frontiers | Research Topics

Human-centred computer audition: Sound, music, and healthcare

Edited by
Kun Qian, Gyorgy Fazekas, Björn Wolfgang Schuller,
Shengchen Li and Zijin Li

Published in
Frontiers in Digital Health
Frontiers in Physics
Frontiers in Psychology
Frontiers in Computer Science

ISSN 1664-8714
ISBN 978-2-8325-4196-8
DOI 10.3389/978-2-8325-4196-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Human-centred computer audition: Sound, music, and healthcare

Topic editors

Kun Qian — Beijing Institute of Technology, China

Gyorgy Fazekas — Queen Mary University of London, United Kingdom

Björn Wolfgang Schuller — Imperial College London, United Kingdom

Shengchen Li — Xi'an Jiaotong-Liverpool University, China

Zijin Li — Central Conservatory of Music, China

Citation

Qian, K., Fazekas, G., Schuller, B. W., Li, S., Li, Z., eds. (2024). *Human-centred computer audition: Sound, music, and healthcare*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4196-8

Table of contents

frontiers | Frontiers in Digital Health

TYPE Editorial
PUBLISHED 12 December 2023
DOI 10.3389/fdgth.2023.1340517

OPEN ACCESS

EDITED AND REVIEWED BY
Uwe Aickelin,
The University of Melbourne, Australia

*CORRESPONDENCE
Kun Qian
✉ qian@bit.edu.cn

RECEIVED 18 November 2023
ACCEPTED 21 November 2023
PUBLISHED 12 December 2023

CITATION
Qian K, Fazekas G, Li S, Li Z and Schuller BW (2023) Editorial: Human-centred computer audition: sound, music, and healthcare. Front. Digit. Health 5:1340517. doi: 10.3389/fdgth.2023.1340517

Editorial: Human-centred computer audition: sound, music, and healthcare

Kun Qian[1,2]*, Gyorgy Fazekas[3], Shengchen Li[4], Zijin Li[5] and Björn W. Schuller[6,7]

[1]Key Laboratory of Brain Health Intelligent Evaluation and Intervention (Beijing Institute of Technology), Ministry of Education, Beijing, China, [2]School of Medical Technology, Beijing Institute of Technology, Beijing, China, [3]Centre for Digital Music (C4DM), School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom, [4]Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China, [5]Department of Music AI and Music Information Technology, Central Conservatory of Music, Beijing, China, [6]GLAM – the Group on Language, Audio, & Music, Imperial College London, London, United Kingdom, [7]Department of Computer Science, Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany

KEYWORDS

computer audition, machine learning, deep learning, artificial intelligence, brain sciences

Editorial on the Research Topic
Human-centred computer audition: sound, music, and healthcare

1. Introduction

At the time of writing this editorial, OpenAI has announced its newest model called chatGPT-4 Turbo.[1] When dreaming for the blue print that we can better the life via this revolution of AI technologies by foundation models, it is a time for almost every person to think how to live with the powerful artificial intelligence (AI) models in the future.

A future that may also challenge our societies and current living in many ways (1) including or even particularly in healthcare (2). Thinking especially of audio, a similar rise of increasingly capable and powerful foundation models appears at highly accelerated pace and with increasingly emergent behaviour. One of the latest at the time of writing is Uniaudio—showing an overly impressive range of zero-shot abilities (3).

For a long time in the field of health, machines have been taught to "see" and/or to "read" rather than to "listen." This is one of the reasons why more progress was achieved in the field of computer vision (CV) and natural language processing (NLP) rather than computer audition (CA) in this domain. Nevertheless, the promising contributions of audio cannot be ignored for its excellent performance in healthcare (4).

Motivated by the concept of human-centred AI (HAI), we organised the research topic on "Human-Centred Computer Audition: Sound, Music, and Healthcare," which lasted from April 2021 to January 2023. Finally, 10 articles were accepted and published after a rigorous peer-review process. There are 57 authors involved in this research topic.

[1] https://openai.com/blog/new-models-and-developer-products-announced-at-devday

In the remainder of this editorial, we will briefly introduce the published research articles in this research topic collection. Then, insights and perspectives will be given towards the future work.

2. Contributions

The published contributions have covered the planned scope, e.g., computational analysis of sound scenes and events, digital music, computer audition for healthcare, computational paralinguistics, and explainable AI in computer audition. In the following, grouped by categories, we provide a brief description of the collected articles.

2.1 Fast screening of COVID-19

Whether audio could serve as a novel digital phenotype for detection of COVID-19 has been increasingly studied in the past three years (5, 6). Coppock et al. summarise the contributions in the organised INTERSPEECH 2021 Computational Paralinguistics Challenges: COVID-19 Cough, (CCS) and COVID-19 Speech, (CSS) (7). They indicated that, a classifier trained by the infected individuals' respiratory sounds can achieve moderate detection rates of COVID-19. However, whether the audio biomarkers in respiratory sounds of infected individuals are unique for COVID-19 or not is still a question to be answered. Chang et al. introduced a “CovNet” which uses a transfer learning framework to improve the models' generalisation. Experimental results show their models' efficiency by considering a parameter transferring strategy and an embedding incorporation strategy. Akman et al. propose an end-to-end deep neural network model (called “CIdeR”) for exploring the methodological adaptation to new datasets with different modalities. From the experiments, their proposed model can serve across multiple audio types. However, they found that it is difficult to train a common COVID-19 classifier due to the limitations of a joint usage of datasets.

2.2 Domestic activity

Audio tagging of domestic activities can provide important information on health and wellbeing. Yang et al. present an explainable tensor network for monitoring domestic activities via audio signals. They indicated that, the combination of the tensor network can reduce the redundancy of the network.

2.3 Music and brain

Music therapy appears promising for its non-drug characteristic, specifically for treatment of mental disorders (8). However, the influences of music on the brain are still an open question to be answered. Wei et al. contribute a review on neurocognition for timbre perception. They conclude that, timbre perception is promising in psychological application. Further. Liu et al. studied timbre fusion of Chinese and Western instruments. This bears interest, given that in a recent study, timbre features are found to be strongly associated with the human affective states (9). Next, Miyamoto et al. introduce a meta-learning strategy in a music generation system. More fundamentally, Corona-González et al. presented a study on personalised theta and beta binaural beats for brain entrainment. The conclusion made is that the neural resynchronisation was met with both personalised theta and beta binaural beats whereas there seemed to be no different mental conditions achieved.

2.4 Artificial hearing

A disyllabic corpus that could be used to examine the performance of pitch recognition of cochlear implant users was introduced. Wang et al. found that, higher scores of tone recognition tend to be achieved by listeners with longer cochlear implant listening experience.

2.5 Speech emotion recognition

Speech emotion recognition is a widely-studied field in affective computing. The combination of task-specific speech enhancement and data augmentation as a strategy has been used for improving the overall multimodal emotion recognition in noisy conditions. This contribution of Kshirsagar et al. can benefit the speech-based affective information retrieval task in real-world applications.

3. Insights and perspectives

When reading over the collection of this research topic, one finds promising potential of computer audition that can benefit manifold health-related aspects of our life. However, one needs to fully consider the current limitations and keep an eye on the future progress of computer audition.

First, *data scarcity* is still a serious challenge (10) that constrains the fast development of audio based large models. The hardware limitations and further factors impede the collection of high-quality audio data at large scale which could provide sufficient training for current state-of-the-art large models in this domain. Besides, the annotation of audio data (specifically for medical applications) is often difficult. Therefore, advanced strategies such as meta-learning (11), and self-supervised learning should be taken into account prior to the event of generalist (medical) AI (12).

Second, fundamental studies on features, models, and strategies are of interest but limited. Among this collection, we can see some contributions focus on extracting novel audio features to improve the performance of models. We hope to see more works in the future towards the interpretation of the models (13).

Third, the mechanism of the brain's perception of audio is worth exploring in considerably more depth. It will not only be beneficial for building brain-inspired deep learning models, but also for our understanding more deeply music/audio therapy.

Last but not the least, how to leverage the power of the coming large models to discover more possibilities of computer audition is an open question to be answered.

Author contributions

KQ: Writing – original draft, Writing – review & editing; GF: Writing – review & editing; SL: Writing – review & editing; ZL: Writing – review & editing; BS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

This work was partially supported by the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects 2021ZD0201900, the National Natural Science Foundation of China (No. 62272044), and the Teli Young Fellow Program from the Beijing Institute of Technology, China.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Peters MA, Jackson L, Papastephanou M, Jandrić P, Lazaroiu G, Evers CW, et al. AI, the future of humanity: ChatGPT-4, philosophy, education—critical responses. *Educ Philos Theory*. (2023) 1–35.

2. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit Med*. (2023) 6:135. doi: 10.1038/s41746-023-00879-8

3. Yang D, Tian J, Tan X, Huang R, Liu S, Chang X, et al. UniAudio: an audio foundation model toward universal audio generation [Preprint] (2023). Available at: https://doi.org/10.48550/arXiv.2310.00704

4. Qian K, Li X, Li H, Li S, Li W, Ning Z, et al. Computer audition for healthcare: opportunities and challenges. *Front Digit Health*. (2020) 2:1–4. doi: 10.3389/fdgth.2020.00005

5. Coppock H, Jones L, Kiskin I, Schuller BW. COVID-19 detection from audio: seven grains of salt. *Lancet Digit Health*. (2021) 3:e537–8. doi: 10.1016/S2589-7500(21)00141-2

6. Deshpande G, Batliner A, Schuller BW. AI-based human audio processing for COVID-19: a comprehensive overview. *Pattern Recognit*. (2022) 122:1–10. doi: 10.1016/j.patcog.2021.108289

7. Schuller B, Batliner A, Bergler C, Mascolo C, Han J, Lefter I. COVID-19 cough, COVID-19 speech, escalation & primates. *Proc. INTERSPEECH*; Brno, Czechia (2021). p. 431–5.

8. Qian K, Schuller BW, Guan X, Hu B. Intelligent music intervention for mental disorders: insights and perspectives. *IEEE Trans Comput Soc Syst*. (2023) 10:2–9. doi: 10.1109/TCSS.2023.3235079

9. Luo G, Sun S, Qian K, Hu B, Schuller BW, Yamamoto Y, et al. How does music affect your brain? A pilot study on EEG, music features for automatic analysis. *Proc. EMBC*; Sydney, Australia. IEEE (2023). p. 1–4.

10. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri A, Al-dabbagh BSN, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J Big Data*. (2023) 10:46. doi: 10.1186/s40537-023-00727-2

11. Vettoruzzo A, Bouguelia MR, Vanschoren J, Rögnvaldsson T, Santosh K. Advances challenges in meta-learning: a technical review [Preprint] (2023). Available at: https://doi.org/10.48550/arXiv.2307.04722

12. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. (2023) 616:259–65. doi: 10.1038/s41586-023-05881-4

13. Frommholz A, Seipel F, Lapuschkin S, Samek W, Vielhaben J. XAI-based comparison of input representations for audio event classification [Preprint] (2023). https://doi.org/10.48550/arXiv.2304.14019

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fpsyg.2021.764068

Personalized Theta and Beta Binaural Beats for Brain Entrainment: An Electroencephalographic Analysis

César E. Corona-González, Luz María Alonso-Valerdi and David I. Ibarra-Zarate*

Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Monterrey, Mexico

Binaural beats (BB) consist of two slightly distinct auditory frequencies (one in each ear), which are differentiated with clinical electroencephalographic (EEG) bandwidths, namely, delta, theta, alpha, beta, or gamma. This auditory stimulation has been widely used to module brain rhythms and thus inducing the mental condition associated with the EEG bandwidth in use. The aim of this research was to investigate whether personalized BB (specifically those within theta and beta EEG bands) improve brain entrainment. Personalized BB consisted of pure tones with a carrier tone of 500 Hz in the left ear together with an adjustable frequency in the right ear that was defined for theta BB (since f_c for theta EEG band was 4.60 Hz ± 0.70 SD) and beta BB (since f_c for beta EEG band was 18.42 Hz ± 2.82 SD). The adjustable frequencies were estimated for each participant in accordance with their heart rate by applying the Brain-Body Coupling Theorem postulated by Klimesch. To achieve this aim, 20 healthy volunteers were stimulated with their personalized theta and beta BB for 20 min and their EEG signals were collected with 22 channels. EEG analysis was based on the comparison of power spectral density among three mental conditions: (1) theta BB stimulation, (2) beta BB stimulation, and (3) resting state. Results showed larger absolute power differences for both BB stimulation sessions than resting state on bilateral temporal and parietal regions. This power change seems to be related to auditory perception and sound location. However, no significant differences were found between theta and beta BB sessions when it was expected to achieve different brain entrainments, since theta and beta BB induce relaxation and readiness, respectively. In addition, relative power analysis (theta BB/resting state) revealed alpha band desynchronization in the parieto-occipital region when volunteers listened to theta BB, suggesting that participants felt uncomfortable. In conclusion, neural resynchronization was met with both personalized theta and beta BB, but no different mental conditions seemed to be achieved.

Keywords: binaural beats, beta, theta, EEG, brain entrainment

Edited by:
Zijin Li,
Central Conservatory of Music, China

Reviewed by:
Serap Aydin,
Hacettepe University, Turkey
Bei Song,
Harbin Conservatory of Music, China

****Correspondence:***
César E. Corona-González
a00833959@itesm.mx

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *24 August 2021*
Accepted: *04 October 2021*
Published: *11 November 2021*

Citation:
Corona-González CE, Alonso-Valerdi LM and Ibarra-Zarate DI (2021) Personalized Theta and Beta Binaural Beats for Brain Entrainment: An Electroencephalographic Analysis. Front. Psychol. 12:764068. doi: 10.3389/fpsyg.2021.764068

Abbreviations: AP, Absolute power; BB, Binaural beats; BBC, Brain Body Coupling; BOLD, Blood Oxygen Level Dependent; dB HL, Decibels hearing level; EEG, Electroencephalography; f_c, Central frequency; $F_{R_{ear}}$, Right ear stimulation frequency; HR, Heart rate; PSD, Power spectral density; RP, Relative power; S1, Session 1 (theta BB); S2, Session 2 (beta BB); SPL, Sound pressure level; $S_{R,}$ Resting state session.

INTRODUCTION

In 1839, Heinrich Wilhelm Dove found that providing two slightly different tone frequencies, one in each ear, were perceived as a third phantom frequency depicted by the difference of these two frequencies, which was called binaural beats (BB; Keeley, 2006). It was until the 1950's when Robert Monroe formally started to research the clinical application of BB, establishing that the dissimilarity of both frequencies must be within the electroencephalographic (EEG) spectrum, that brain entrainment could be elicited (Berger and Turow, 2011). Later, Worden and Marsh (1968) were investigating about electrophysiological effects of sound on the brain. They found that an auditory stimulus provokes a synchronous-neural evoked response which reproduces the frequency and waveform of the incoming stimulus throughout the central auditory pathway. This effect was coined as Frequency Following Response (FFR; Marsh and Worden, 1968; Worden and Marsh, 1968).

Neurons oscillate in several well-known EEG frequency bands. These are delta (δ = 0.1–4 Hz), theta (θ = 4–8 Hz), alpha (α = 8–13 Hz), beta (β = 13–30 Hz), and gamma (γ > 30 Hz). Normally, delta band is present while deep sleep; theta band is in extremely relaxation, drowsiness, or meditation; alpha is best seen at rest with eyes closed; beta is present during problem solving and focusing; and gamma is characterized by cognitive and motor functions (Siuly et al., 2016). In terms of EEG frequency bands, BB has a frequency difference within the band range of interest. For each EEG frequency band, the following corresponding BB can be generated: (1) delta BB, theta BB, alpha BB, beta BB, and gamma BB. In theory, each BB produces neural oscillations at the corresponding EEG frequency band, inducing the associated mental state. Tone frequencies stimuli between 450 and 500 Hz are recommended (García Argibay, 2018).

Theta and beta BBs are of particular interest since they may cause states of relaxation and attentiveness, respectively, which are opposite mental states so that they can be easily compared. In a study performed by Jirakittayakorn and Wongsawat (2017b), brain entrainment in theta wave was achieved when subjects listened to a 6 Hz BB for 10 min, promoting meditative states (Jirakittayakorn and Wongsawat, 2017b). Moreover, overwrought states due to insomnia were diminished by theta BB (Choi et al., 2019). In addition, beta BB has been used to improve (1) short-term (Gálvez et al., 2018) and long-term memory (García-Argibay et al., 2017), (2) working memory (Beauchene et al., 2016), (3) focusing levels and problem solution (Simmons, 2016), and (4) attention (Park et al., 2018). Conversely, López-Caballero and Escera (2017) disagree with brain entrainment due to BB, since no differences in EEG power between baseline and BB exposure were found while using theta, alpha, beta, gamma, and upper gamma BB (López-Caballero and Escera, 2017). Additionally, other studies failed on promoting brainwave entrainment using theta (Goodin et al., 2012; Orozco Perez et al., 2020) or beta BB (Goodin et al., 2012; Vernon et al., 2012). Another example is the work undertaken by Gao et al. (2014), where EEG signals were studied while delta, theta, alpha, and beta BB were applied. They did not find any brain entrainment after 20 min of BB stimulation (5 min per band, followed by a 2-min break between bands). Nevertheless, relative power variations within the four bands were thought to yield neural connectivity changes (Gao et al., 2014).

As shown by past studies, contradictory findings have been found. On one side, BB has shown to be successful in practice. On the other side, no EEG modulation (brainwave entrainment) has been achieved in all the BB studies. It is hypothesized that the BB effect can be always achieved if individual frequency bands of brain oscillations are found and used to generate BB. According to the Brain Body Coupling (BBC) theorem established by Klimesch (2018), brain and body (e.g., gastric waves, motion oscillations, blinks, and heart rate) oscillations are coupled to each other at rest (Klimesch, 2018). Therefore, individual brain-body frequency bands can be found if one of the brain-body oscillations is known, for example, heart rate (HR). On this basis, it is proposed to generate personalized theta and beta BB in accordance with individual theta and beta EEG frequency bands (previously found by BBC theorem), and then analyze the EEG modulation obtained after theta and beta BB exposure. For this purpose, the present investigation was undertaken as follows. First, 20 volunteers were recruited to whom audiometry was applied, and resting HR was taken (see section "Participants"). Second, theta and beta BBs were generated according to the HR of a participant (see section "Binaural Auditory Stimuli"). Third, the experiment was performed in two phases: (1) environment for binaural stimuli, where subjects were instructed about the experimental procedure (see section "Environment for Binaural Stimuli") and (2) presentation of binaural stimuli, where EEG recordings were collected while participants were exposed to BBs (see section "Presentation of Binaural Stimuli"). Finally, EEG data analysis was carried out, which consisted in preprocessing (see section "Preprocessing"), processing (see section "Processing"), and statistical evaluation (see section "Statistical Evaluation").

MATERIALS AND METHODS

Participants

For this study, 20 healthy students of Tecnológico de Monterrey (six women and 14 men) aged between 19 and 24 years old were recruited (i.e., a convenience sampling method was undertaken). All of them reported not having musical experience and voluntarily consented to their participation in the study. This study was previously approved by the Ethical Committee of the Medicine School at Tecnológico de Monterrey (CONBIOETICA-19-CEI-011-20161017).

Data Acquisition and Equipment

To record EEG activity, the mBrainTrain system was used. This is a Bluetooth-interface EEG device of 24 channels (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, M1, M2, AFz, CPz, POz) positioned according to the 10/20 international system, as shown in **Figure 1**. The channels M1 and M2 were set as references and the channel FCz as ground. The sampling frequency was 250 Hz. The mBrainTrain has the Smarting Streamer software, which was used to verify electrode impedances to be below 5 kΩ. To set up the experimental

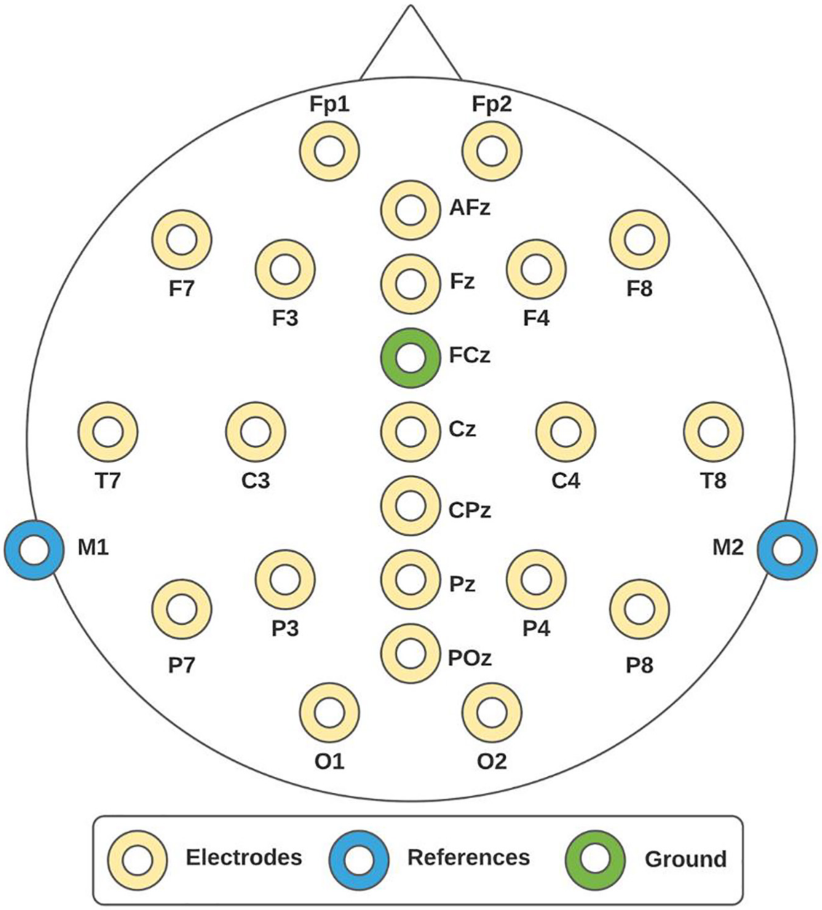


FIGURE 1 | 10/20 system. Electroencephalographic (EEG) montage consisted of 24 channels, whereby 22 of them were recording sites (yellow electrodes), one reference (blue electrodes), and one ground (green electrode).

paradigm, OpenVibe was employed. OpenVibe is a free software commonly used in the neuroscience field to design and test brain-computer interfaces and to develop experimental paradigms for offline records (Renard et al., 2010).

To perform the audiometry, an audiometer Interacoustics-AD226 was utilized. Tonal audiometry within ranges from 250 Hz to 2 kHz was used to determine auditory thresholds. Measurements were taken in dB HL. Finally, HR was taken with pulse oximeter Hergom-MD300 and auditory stimuli was given through open-back headphones SHURE-SRH1840.

Binaural Auditory Stimuli

According to Klimesch (2013), biosignals do not vary randomly or arbitrarily. Namely, brain and body signals oscillations are aligned with each other and form a single frequency architecture. The interaction between brain and body may be described as a complex system that couples and decouples according to a specific harmony frequency described by,

$$f_d\left(i\right) = s * 2^i \tag{1}$$

where s is the scaling factor, i refers to the biosignal of interest, and f is the fundamental frequency of the biosignal oscillation. When $i = 0$, f_d refers to cardiac activity. When $i < 0$, f_d refers to breathing rhythms (including Mayer waves that are the lowest frequency in the respiratory process), blood pressure waves, rhythmic fluctuations in the blood oxygen level-dependent (BOLD) signal at intrinsic mode fluctuations, and gastric waves. When $i > 0$, f_d refers to brain oscillations [delta ($i = 1$), theta ($i = 2$), alpha ($i = 3$), beta ($i = 4$), gamma ($i = 5$)]. In addition, upper, and lower frequencies of each fundamental frequency can be, respectively, estimated by,

$$uf_b\left(i\right) = \frac{1.25 \times 2^{i+1}}{g} \tag{2}$$

$$lf_b\left(i\right) = \left(1.25 \times 2^{i-1}\right) \times g \tag{3}$$

Recently, Klimesch (2018) empirically demonstrated that the resonance of a biosignal is harmonized with other ones at resting state. Some examples are:

- During respiration, HR increases at inhalation and decreases at exhalation.
- Heart rate presents a clear tendency 10:1 frequency ratio relative to breathing rate owing to energy demands and emotional regulation.
- Gastric waves explain 8% of alpha band modulation of EEG signals, and 15% of the BOLD variance is explained by the gastric phase.
- The slow frequency that modulates the envelope of the electromyographic signals is originated from neural mechanisms of motor control and resonance frequency of body parts.

In the light of the above evidence, this research proposes the design of personalized BB based on individual HR. That is, when $i = 0$ in Eq. (1) then,

$$f_d(0) = s * 2^0 = s = HR \ [Hz]$$

Having estimated s, central frequency (f_c) of theta and beta EEG rhythms were calculated in accordance with the individual HR at hand. Note that resulting fundamental frequencies were not subjective, they were rather relative to the human body function, and their values are around the well clinically established frequency bands.

As theta BB consisted of pure tones of 500 Hz for the left ear, f_c of theta EEG band was used to adjust the frequency for the right ear (F_{R_ear}). Similarly, beta BB consisted of pure tones of 500 Hz for the left ear. Then, f_c of beta EEG band was used to adjust the frequency for the right ear. That is,

$$F_{R\_ear} = 500 \ Hz - f_c \tag{4}$$

The resulting EEG frequency bands for each volunteer, and the corresponding BB produced after the individual EEG frequency band identification, are reported in **Table 1**. The computational algorithm to generate personalized BB in accordance with the BBC theorem was programmed in MATLAB programming language and was published in MathWorks File Exchange Forum[1].

Experimental Paradigm and Protocols

Each volunteer participated in two BB sessions on different days. In the first session (S1), theta BB was applied, and in the second session (S2), beta BB was used. Both BB exposures were for 20 min, since listening to BB longer than 20 min may lead to mental fatigue (Jirakittayakorn and Wongsawat, 2017a). All the participants were seated on a comfortable chair and in a quiet room. All of them were asked to keep their eyes closed during BB stimulation. The procedure was conducted in two steps: (1) environment for binaural stimuli and (2) presentation of binaural stimuli.

Environment for Binaural Stimuli

First, the purpose of the study was explained to the participant, and after agreeing to their participation, they signed a consent form. Second, the Neurologic Evaluation Questionnaire from Neuroscience Institute of University of Guadalajara (Balart Sánchez, 2017) was applied to assess their medical history about neurological health. Third, tonal audiometry and the HR at resting state were taken. Audiometry and HR per volunteer are reported in **Table 2**. Fourth, the volunteer was asked to sit down and relax while putting on the EEG cap and electrodes impedances were controlled to be kept below 5 $k\Omega$. Finally, paradigm instructions were given to the participant. **Figure 2** shows the whole preparation sequence. For S2, volunteer preparation started from step 6 in **Figure 2**.

Presentation of Binaural Stimuli

First, volunteers were instructed to keep their eyes closed for 3 min. Hereinafter, this EEG recording is referred to as to baseline or resting state session (S_R). Second, volunteers were stimulated with their personalized theta and beta BB for 20 min in S1 and S2, respectively. The sound level was 60 dB SPL for both sessions (World Health Organization [WHO] and International Telecommunication Union [ITU], 2019).

Signal Analysis

Power spectral density is a method to extract the power content of a signal in the frequency domain. Power spectral density (PSD) utilizes the Discrete Fourier Transform to obtain the periodogram. Welch's method is one of the most applied algorithms to estimate PSD (Zhang, 2019). In this study, PSD was calculated for the three different mental states: (1) S_R, (2) S1, and (3) S2. The analysis was carried out in two steps: (1) preprocessing and (2) processing.

Preprocessing

Electroencephalographic signals were preprocessed in MATLAB using the EEGLab toolbox, developed by the Swartz Center for Computational Neuroscience at the University of California, San Diego (Martínez-Cancino et al., 2020). For preprocessing, the sampling frequency was 250 Hz, the Direct Current component was removed, and a bandpass filter from 0.1 to 100 Hz was utilized along with a band-stop filter for removing 59–61 Hz. Both filters were IIR Butterworth 8th order. FCz was set as the ground electrode and re-referencing was regarding M1 and M2 average. Then, visual inspection was required for cleaning up EEG signals from abrupt changes due to muscular artifacts. Finally, Independent Component Analysis was applied for ocular and cardiac artifacts removal. **Figures 3A,B** exemplify the muscular and ocular artifacts of one of the volunteers. The raw EEG data set is freely accessible and is available in the Mendeley database at https://data.mendeley.com/datasets/ppz3r5j2n2/2.

Processing

To quantify BB effects, PSD was extracted from all volunteers in S_R, S1, and S2 conditions. For Fourier transform algorithms, such as Welch's method, stationarity must be satisfied. However, EEG signals can be segmented into short windows where stationarity is assumed (Nunez et al., 2016), especially when emotional processes are mediated by visual or audio-visual stimuli (Aydın, 2020). For that reason, PSD was applied to each volunteer dataset

[1] https://www.mathworks.com/matlabcentral/fileexchange/99544-personalized-binaural-beats-generator

TABLE 1 | Volunteer data.

Volunteer	EEG fundamental frequency (f_c)		Theta BB (S1)		Beta BB (S2)	
	Theta (Hz)	Beta (Hz)	Left (Hz)	Right $F_{R_{ear}}$ (Hz)	Left (Hz)	Right $F_{R_{ear}}$ (Hz)
1	4.33	17.33	500	495.67	500	482.67
2	4.8	19.2	500	495.2	500	480.8
3	4	16	500	496	500	484
4	3.87	15.47	500	496.13	500	484.53
5	5.4	21.6	500	494.6	500	478.4
6	4.2	16.8	500	495.8	500	483.2
7	5.13	20.53	500	494.87	500	479.47
8	4.67	18.67	500	495.33	500	481.33
9	5.93	23.73	500	494.07	500	476.27
10	4.73	18.93	500	495.27	500	481.07
11	4.67	18.67	500	495.33	500	481.33
12	5.93	23.73	500	494.07	500	476.27
13	3.87	15.47	500	496.13	500	484.53
14	4.53	18.13	500	495.47	500	481.87
15	4.73	18.93	500	495.27	500	481.07
16	3.2	12.8	500	496.8	500	487.2
17	4	16	500	496	500	484
18	5.53	22.13	500	494.47	500	477.87
19	4.87	19.47	500	495.13	500	480.53
20	4.73	18.93	500	495.27	500	481.07
Mean	**4.60**	**18.42**	Mean	**495.34**	Mean	**481.37**
S.D	**0.70**	**2.82**	S.D	**0.70**	S.D	**2.82**

f_c for theta and beta bands and $F_{R_{ear}}$ for BB customizing for S1 and S2.

TABLE 2 | Heart rate (HR) and audiometry values per volunteer.

Volunteer	HR	Right ear (dB)						Left ear (dB)					
		250 Hz	500 Hz	750 Hz	1 kHz	1.5 kHz	2 kHz	250 Hz	500 Hz	750 Hz	1 kHz	1.5 kHz	2 kHz
1	65	25	10	15	0	0	0	10	10	10	5	0	0
2	72	15	10	10	5	5	0	20	10	10	5	0	0
3	60	15	10	10	10	0	5	10	10	10	5	5	5
4	58	25	20	15	15	10	5	20	15	20	15	10	15
5	81	10	10	5	5	10	10	10	5	5	0	5	5
6	63	20	5	10	5	10	5	10	10	5	5	5	0
7	77	10	5	5	10	0	0	5	5	5	0	0	5
8	70	10	15	15	10	5	5	5	10	10	10	5	5
9	89	20	10	15	10	10	5	10	5	10	5	5	0
10	71	5	15	10	0	0	0	15	20	15	5	0	0
11	70	10	15	15	10	5	5	15	15	15	10	5	5
12	89	20	15	10	15	10	5	15	5	10	10	5	0
13	58	10	5	10	10	5	0	0	5	5	5	0	0
14	68	5	0	0	0	5	0	0	5	5	0	0	0
15	71	10	5	5	5	5	0	15	5	5	0	5	0
16	48	15	10	10	5	10	5	20	10	10	5	5	0
17	60	5	5	10	5	10	0	10	5	5	0	0	0
18	83	30	25	15	10	5	5	25	15	20	10	5	0
19	73	5	5	10	5	10	0	10	5	5	0	0	0
20	71	15	10	10	5	10	5	10	10	10	5	10	0

HR was taken at rest in each volunteer, whereas audiometry was performed from 250 Hz to 2 kHz.

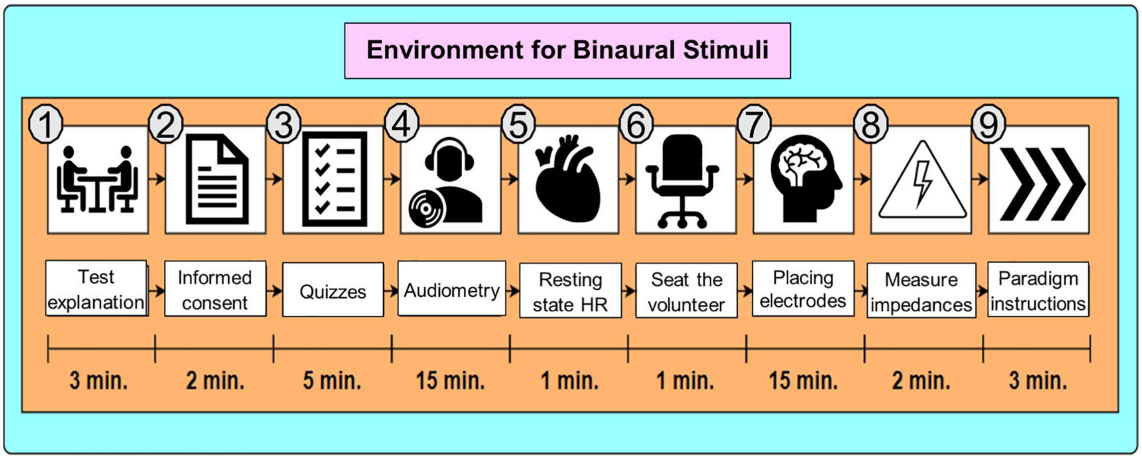


FIGURE 2 | Environment for Binaural Stimuli. The preparation stage in S1 was about 47 min long. For S2, it was about 21 min long since the procedure started at step 6 (sitting down the volunteer).

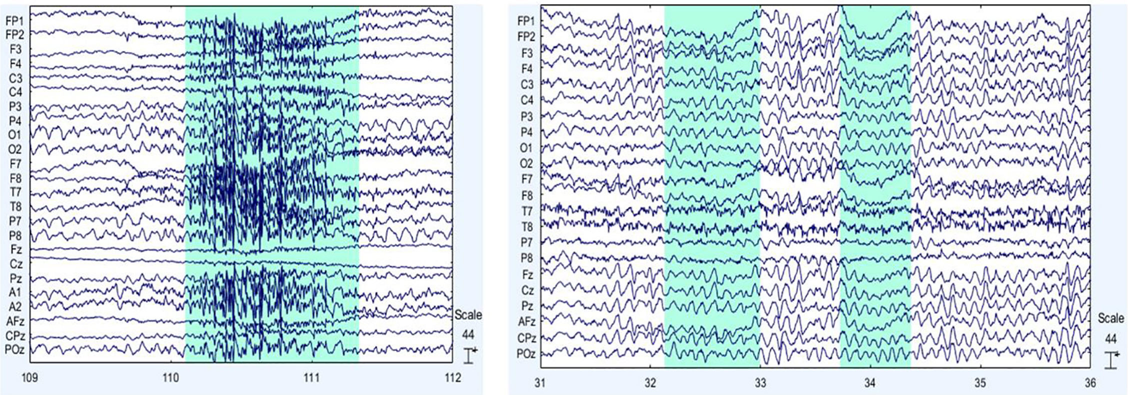


FIGURE 3 | Main EEG artifacts are ocular and muscular electrical activity. On the left **(A)**, muscular artifacts are shown, and on the right **(B)**, ocular artifacts are presented. The sections shaded in blue indicate examples of the muscular (on the left) and eye (on the right) artifacts themselves.

using a windowing of 1 s and an overlapping of 50%. The analysis was carried out taking into consideration two parameters from PSD: (1) absolute power (AP) from S_R, S1, and S2; and (2) relative power (RP) from S1/S_R and S2/S_R. Analyzing AP can provide spectral information regarding neural activity before and after listening to theta or beta BB (Park, 2020). Thus, brain entrainment can be identified in S1 if theta BB triggers the highest AP in theta EEG band. Similarly, brain entrainment can be detected for S2 if beta BB elicits the highest AP in the beta EEG band. However, as comparing AP for the three mental states, it is difficult to differentiate precise changes in EEG frequencies. Therefore, using RP for S1/S_R and S2/S_R allows to directly compare the influence of both theta and beta BB over resting state, so that brain entrainment can be supported. **Figure 4** summarizes these two analyses.

Power Spectral Density – Absolute Power

Power spectral density was calculated for S_R, S1, and S2 to obtain the AP of each mental state. AP values were allocated in three different matrices (one for each state), which dimensions refer to volunteers (20) by channels (22) by AP values (126). Then, these matrices were averaged by volunteers and transformed into a decibel (dB) scale. As the highest frequencies of interest are in the beta range, only frequencies from 0 to 30 Hz were considered for the analysis. Finally, power values within these frequencies were compared in every channel for S_R, S1, and S2.

Power Spectral Density – Relative Power

Once PSD for S_R, S1, and S2 was individually estimated, a data standardization was performed to the power values of S1 and S2, both regarding S_R. The standardized value represents the Relative Power (RP) between BB session and S_R. Eq. (5) summarizes the calculation of RP:

$$RP_s^v = {AP_s^v}/{AP_R^v} \qquad (5)$$

where,

RP_s^v represents the relative power of volunteer "v" and session "s" (theta BB or beta BB).

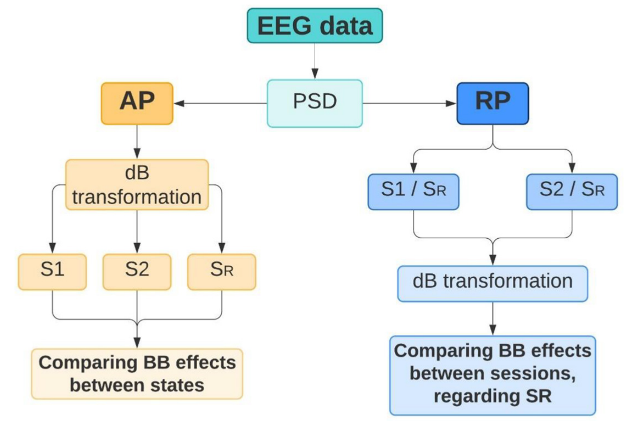

EEG data
AP
PSD
RP
dB transformation
S1
S2
SR
Comparing BB effects between states
S1 / SR
S2 / SR
dB transformation
Comparing BB effects between sessions, regarding SR

FIGURE 4 | Power spectral density (PSD) analysis. Absolute power (AP) and relative power (RP) were extracted from EEG data in order to identify brain changes that can denote brain entrainment.

AP_s^v is AP from PSD of volunteer "v" of session "s" and
AP_R^v is AP from PSD of volunteer "v" at resting state.

It should be noted that Eq. (5) was applied to every channel. Therefore, RP for every volunteer across all channels has been calculated until now. After that, two averaged-by-volunteer matrices were computed, one for S1/S$_R$ and the other for S2/S$_R$. Finally, dB transformation was applied. These ratios may manifest the power variation on theta, alpha, and beta bands and may confirm if brain entrainment has been achieved for the specific band (i.e., increase in theta band activity due to theta BB or increase in beta band activity due to beta BB). Owing to the wide frequency range of the gamma band, it was excluded from this analysis. Considering that the theta band was the minimum frequency range for inducing the FFR effect, the delta band was also rejected for the analysis.

Maximum power in theta and beta EEG bands was expected when volunteers were listening to theta and beta BBs, respectively. Therefore, changes in magnitude power of EEG data could be seen and brain entrainment may be accomplished.

Statistical Evaluation

Statistical analysis was carried out for AP and RP estimates. With respect to AP estimates, the comparison was between sessions (1) S$_R$ with S1, and (2) S$_R$ with S2. First, the Shapiro–Wilk test for normality (Ahad et al., 2011) was applied for these two pairs of data. Once normality was confirmed for both cases, a separate *t*-test was performed to find significant differences between S$_R$-S1 and S$_R$-S2 through all channels. Afterward, Cohen's d effect size was utilized to estimate the magnitude of these differences as "negligible" ($d < 0.2$), "small" ($0.2 \leq d < 0.5$), "medium" ($0.5 \leq d < 0.8$), or "large" ($d \geq 0.8$) (Magnusson, 2021). Regarding RP estimates, normality was also tested using the Shapiro–Wilk test for the power difference between theta, alpha, and beta bands in both S1 and S2 in every channel. Subsequently, two alternately two-way ANOVA tests were performed to RP in S1 and S2. The tested values included averaged RP values, channels, and bands. These two ANOVAs were aimed to verify if statistically significant changes in RP amongst EEG bands and channels were achieved after BB stimulation. Finally, a Tukey test was realized (Daniel and Cross, 2013) to locate these differences. A significance level of 0.05 was used in all statistical tests.

PRODUCTION OF THETA AND BETA BINAURAL BEATS SOUND

Theta and beta BB were designed in MATLAB at *.wav* format, with a sampling frequency of 44,100 Hz (Pejrolo and Metcalfe, 2017). BB had an amplitude-modulated sound composed of two pure frequency tones. According to HR estimations shown in **Table 2**, f_c for each EEG band was calculated based on BBC theorem (Klimesch, 2018). By way of illustration, assuming an HR of 70 bpm, *s* turns out as:

$$^{70}\!/_{60} \;=\; \mathbf{1.16\bar{6}};$$

From Eqs (1) and (4), it is explained the mathematical procedure for individual f_c for theta and beta bands, and F_{R_ear} for the personalized design of BB routines, respectively. These values are calculated as follows:

- Theta band:

$$f_c \;=\; 1.16\bar{6} * \; 2^2 \;=\; 4.67\ Hz$$

$$F_{R_{ear\theta}} \;=\; 500\ Hz - \; 4.67\ Hz \;=\; 495.33\ Hz$$

- Beta band:

$$f_c \;=\; 1.16\bar{6} * \; 2^4 \;=\; 18.67\ Hz$$

$$F_{R_{ear\beta}} \;=\; 500\ Hz - \; 18.67\ Hz \;=\; 481.33\ Hz$$

Hence, stimulation frequencies to create theta BB were 500 and 495.33 Hz and for beta BB were 500 and 481.33 Hz. **Table 1** shows f_c and F_{R_ear} values of all volunteers.

RESULTS

Twenty volunteers aged between 19 and 24 were recruited, who reported normal hearing thresholds, good neurological history, and no musical experience. Before S1 was performed, S$_R$ was taken as the baseline for 3 min. Afterward, S1 and S2 lasted 20 min each. The data analysis was based on AP and RP of EEG signals within 0 and 30 Hz. Findings are mentioned below.

Comparison Between Theta and Beta Binaural Beats Effects: Absolute Power Estimation

Absolute power from S1 and S2 were compared with S$_R$ to identify if BB exposure elicited changes in neural activity. In **Figure 5**, it is exhibited average AP from the three conditions in

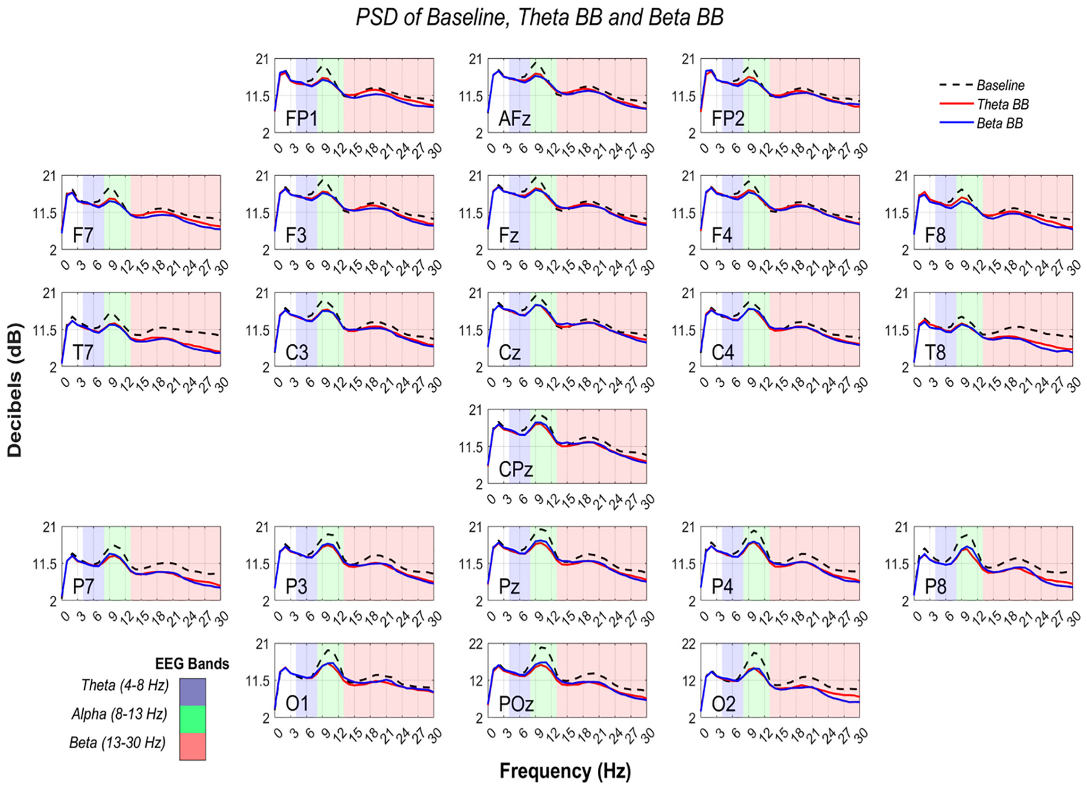


FIGURE 5 | Comparison of PSD from S1, S2, and S_R. AP in dB (*Y*-axis) are shown from 0 to 30 Hz (*X*-axis). Black dashed line represents average AP from S_R, whereas solid lines depict average AP from theta binaural beats (BB) (red) and beta BB (blue). The colored background separates theta (blue), alpha (green), and beta (red) bands.

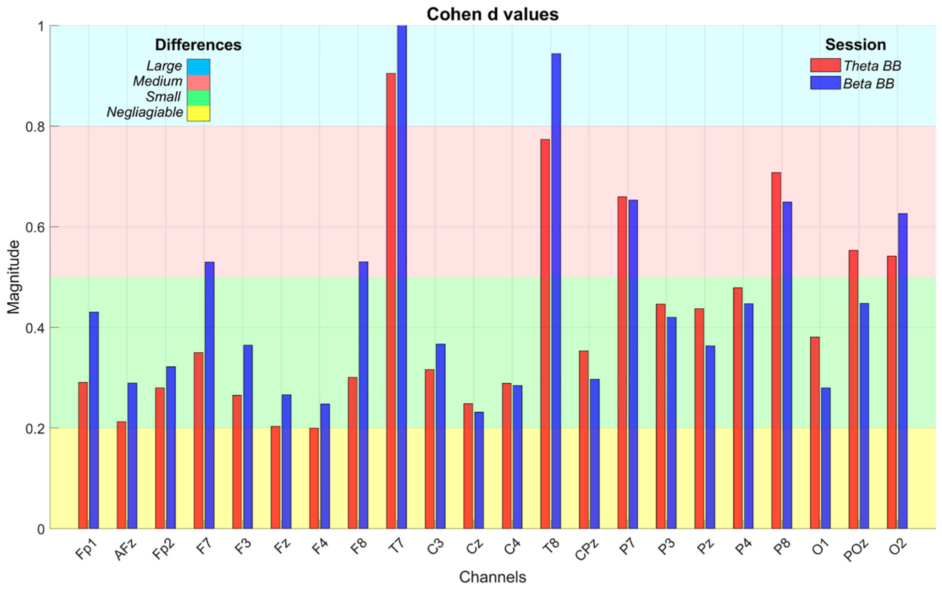


FIGURE 6 | Cohen's d effect size values. The bar plot depicts Cohen's *d* values to estimate the magnitude of statistical differences in absolute power when S1 (red bar) and S2 (blue bar) were compared with S_R. *X*-axis exhibits EEG channels whereas *Y*-axis is the Cohen's *d* value, ranging from 0 to 1. The colored background regions represent the intervals to specify the magnitude of the differences throughout channels and between sessions, such as negligible (yellow), small (green), medium (red), and large (blue).

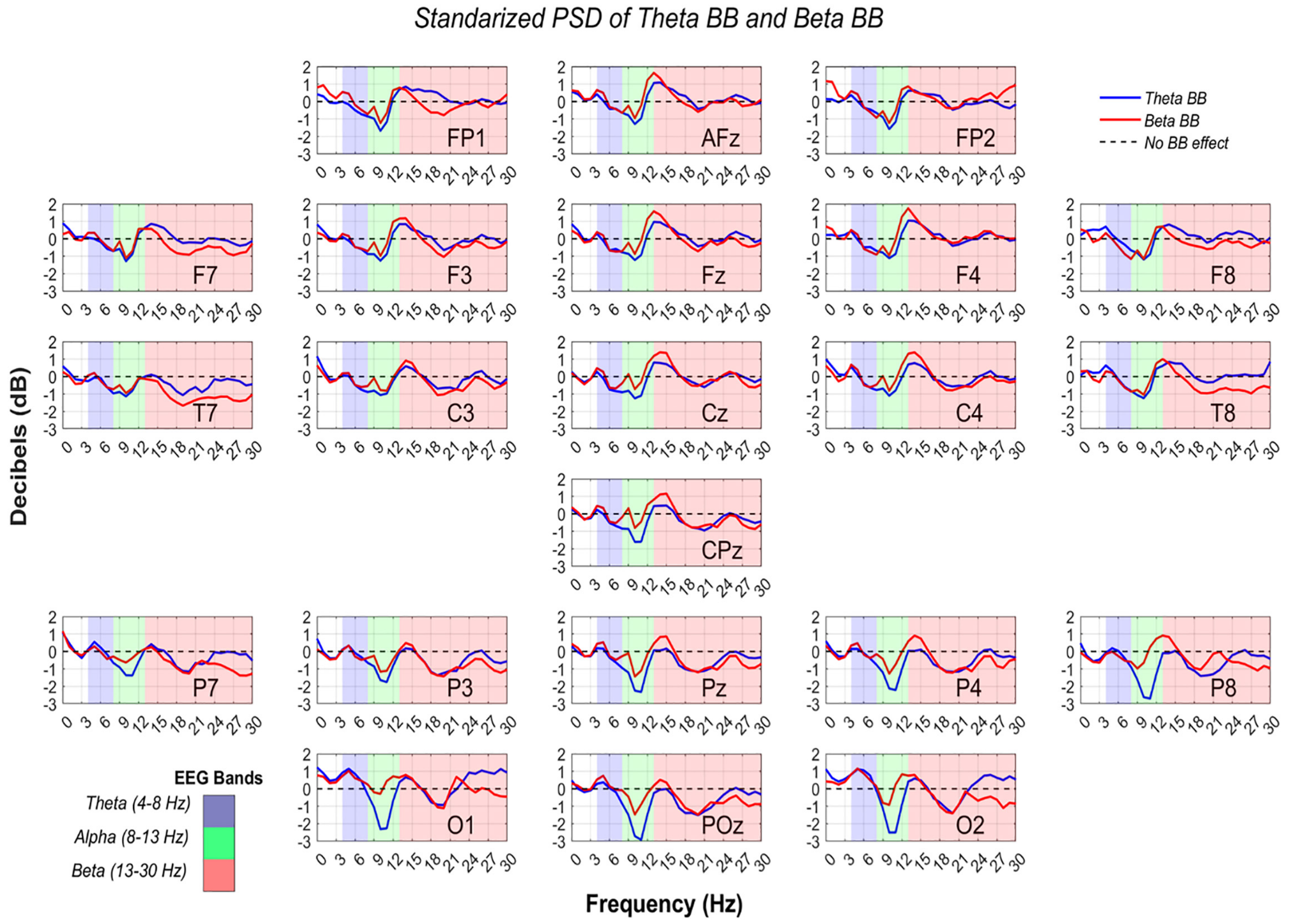


FIGURE 7 | Power spectral density from theta and beta BB, regarding resting state. RP values from theta BB (blue solid line) and beta BB (red solid line) are shown, where *X*-axis is depicted by frequency ranging from 0 to 30 Hz and *Y*-axis is the averaged power ratio between S1/S_R and S2/S_R in dB. Every plot represents a different channel, which was labeled in the lower-right corner. Measurements are highlighted in blue for the theta band, in green for the alpha band, and in red for the beta band.

dB within the frequency range of 0–30 Hz. PSD was estimated across all channels.

At an initial glance, it seemed that average S1, S2, and S_R triggered similar brain activity. To dismiss this issue, a paired *t*-test was applied to AP values of (1) S1 with S_R and (2) S2 with S_R, for all channels. Significant differences were found across all the channels in both comparisons ($p < 0.05$). However, a Cohen's d effect size test was implemented to estimate the magnitude of these differences. Cohen's *d* values are graphically expressed in **Figure 6**.

Comparison Between Electroencephalographic Frequency Bands: Relative Power Estimation

In order to confirm if brain entrainment was achieved, the following conditions must be met: (1) for theta BB, an increase in S1/S_R ratio in the theta band, or (2) for beta BB, an increment in S2/S_R ratio in the beta band. Thus, RP from S1/S_R and S2/S_R were compared across theta, alpha, and beta bands. **Figure 7** shows RP from all channels in dB, throughout EEG frequency ranges where theta, alpha, and beta were colored in blue, green, and red, respectively.

So far, it is known that statistical differences in all channels were identified between BB sessions and resting state, but it does not confirm if brain entrainment was attained. For this reason, a two-way ANOVA was consecutively applied to RP from S1 and S2 to verify if neural activity between bands were dissimilar due to BB. The ANOVA data for each matrix contained RP values from S1 or S2, channels, and bands (theta, alpha, and beta). Significant differences were found for alpha-theta bands ($p = 0$) and alpha-beta bands ($p = 0$). Nevertheless, no statistical difference was obtained for theta-beta ($p = 0.5577$).

DISCUSSION

This study was focused on investigating if brain entrainment could be achieved from the modulation of EEG signals by personalized theta and beta BB stimulation. The EEG analysis followed two approaches: (1) AP, obtained from individual PSD of S1, S2, and $S_{R,}$ and (2) RP, computed by S1/S_R and S2/S_R power ratios. In the following, the results of this research are discussed for each method.

Absolute Power Estimation

As can be seen from **Figure 6**, T7 and T8 showed the highest differences when BB sessions were compared against S_R, followed by medium differences in P7, P8, and O2. After listening to either theta or beta BB, small to medium changes in AP occurred in the remaining channels, suggesting that significant effects due to BB were not entailed or, in other words, brain entrainment was not achieved. As the greatest changes were seen over the temporal lobe, followed by the parietal one, they may be associated with auditory perception and sound location (Benarroch, 2006; Bizley and Cohen, 2013; Goldstein, 2014).

Relative Power Estimation

Since data standardization was carried out, a 0 dB value means that BB session and baseline neural activity were comparable. On the contrary, when dB > 0, brain activity was stronger during BB exposure than baseline. Thus, either theta or beta BB elicited neural synchronization. Similarly, given a dB < 0, it implies that brain activity was higher in the resting state in comparison with the BB session, which means that neural desynchronization was induced (Watkinson and Clarke, 2018).

According to Cohen (2014), an oscillation from −2 dB to +2 dB implies a percentage change from −36.9 to +58.8%. In other words, if RP was equal to −2 dB, it means that baseline activity was higher than in BB session by 36.9%, (i.e., BB induced 36.9% of neural desynchronization). Moreover, a +2 dB change means that the BB session had stronger brain activity than baseline by 58.8% (i.e., BB triggered 58.8% neural synchronization) (Cohen, 2014).

Interestingly, RP in the parieto-occipital region was lesser than −2 dB for alpha EEG band while listening to theta BB, specifically over Pz, P4, P8, O1, POz, and O2 recording sites. This decrease in dB value is explained by greater brain activity at the resting state regarding theta BB. For these channels, as alpha desynchronization occurred only for S1, we suggest that theta BB probably disturbed volunteers instead of inducing them into meditative or relaxed states, even though when they were just seated with closed eyes and doing no task. A theory of this behavior may be related to volunteers feeling uncomfortable when listening to the theta BB (Crespo et al., 2013; Lee et al., 2019).

In conclusion, neural resynchronization was met with both personalized theta and beta BB, but no different mental conditions seemed to be achieved.

Limitations of the Study

The aim of the study was mainly limited due to (1) conditions of participants, and (2) the use of open-back headphones. First, for more specificity in sample selection, hours of sleep in the night before the study and psychological conditions, such as stress or anxiety, should have been considered. These factors can disturb brain oscillations. Second, as BB was delivered through open-back headphones, the environment auditory stimuli could bias neural information. Therefore, the study should have been implemented in an isolated room.

Future Work

In order to develop a full picture of brain activity due to BB exposure, a suggestion would be that the application of other kinds of BB not used in this study such as alpha BB (Park et al., 2018; Shekar et al., 2018) or gamma BB (Colzato et al., 2017; Shekar et al., 2018). Additional studies that give an insight into brain signal modulation are needed for further understanding of which effects BB induces on humans. Empirical studies where BB effects are behaviorally measured are not enough to demonstrate binaural sound influence on the human mental state.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://data.mendeley.com/datasets/ppz3r5j2n2/2.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethical Committee of the Medicine School at Tecnológico de Monterrey (CONBIOETICA-19-CEI-011-20161017). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CC-G contributed to the design of methodology, programming software for signal acquisition, signal processing, statistical analysis, and writing the original draft. LA-V and DI-Z served as advisor of the research by methodology design, feedback on data analysis and statistical analysis, and review and editing the manuscript. All authors contributed to the manuscript revision, read and approved the submitted version.

FUNDING

CC-G was awarded with a scholarship supported by the National Council for Science and Technology (CONACYT) under grant number 894658.

ACKNOWLEDGMENTS

This study was performed at Neuroengineering and Neuroacoustics Research Group from Tecnológico de Monterrey.

REFERENCES

Ahad, N. A., Yin, T. S., Othman, A. R., and Yaacob, C. R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malays.* 40, 637–641.

Aydın, S. (2020). Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips. *IEEE J. Biomed. Health Inform.* 24, 1695–1702. doi: 10.1109/JBHI.2019.2959843

Balart Sánchez, S. A. (2017). *Exploración de una tarea control para una Interface Cerebro-Computadora utilizando un movimiento imaginario intuitivo*. Guadajalajara: Universidad de Guadalajara.

Beauchene, C., Abaid, N., Moran, R., Diana, R. A., and Leonessa, A. (2016). The effect of binaurall beats on visuospatial working memory and cortical connectivity. *PLoS One* 11:e0166630. doi: 10.1371/journal.pone.0166630

Benarroch, E. E. (2006). *Basic Neurosciences with Clinical Applications*. Filadelfia: Butterworth Heinemann Elsevier.

Berger, J., and Turow, G. (2011). *Music, Science, and the Rhythmic Brain. Cultural and Clinical Implications*. New York: Routledge. doi: 10.4324/9780203805299

Bizley, J. K., and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.* 14, 693–707. doi: 10.1038/nrn3565

Choi, H., Bang, Y. R., and Yoon, I. Y. (2019). Entrainment of binaural auditory beats on subjects with insomnia symptoms. *Sleep Med.* 43:A193. doi: 10.1016/j.sleep.2019.11.198

Cohen, M. X. (2014). *Analyzing Neural Time Series Data. Theory and Practice*. Cambridge: The MIT Press. doi: 10.7551/mitpress/9609.001.0001

Colzato, L. S., Steenbergen, L., and Sellaro, R. (2017). The effect of gamma-enhacing binaural beats on the control of feature bindings. *Exp. Brain Res.* 235, 2125–2131. doi: 10.1007/s00221-017-4957-9

Crespo, A., Recuero, M., Galvez, G., and Begoña, A. (2013). Effect of Binaural Stimulation on Attention and EEG. *Arch. Acoust.* 38, 517–528. doi: 10.2478/aoa-2013-0061

Daniel, W. W., and Cross, C. L. (2013). *Biostatistics: a Foundation for Analysis in the Health Sciences*. Hobokenm: Library of Congress Cataloging-in-Publication.

Gálvez, G., Recuero, M., Canuet, L., and Del-Pozo, F. (2018). Short-term effects of binaural beats on EEG power, functional connectivity, cognition, gait and anxiety in Parkinson's disease. *Int. J. Neural Syst.* 28:1750055. doi: 10.1142/S0129065717500551

Gao, X., Cao, H., Ming, D., Qi, H., Wang, X., Wang, X., et al. (2014). Analysis of EEG activity in response to Binaural beats with different frequencies. *Int. J. Psychophysiol.* 94, 399–406. doi: 10.1016/j.ijpsycho.2014.10.010

García Argibay, M. (2018). *Efecto de la sincronización de las oscilaciones neuronales mediante los tonos binaurales en la memoria, atención, ansiedad y percepción del dolor*. Madrid: Universidad Nacional de Educación a Distancia.

García-Argibay, M., Santed, M. A., and Reales, J. M. (2017). Binaural auditory beats affect long-term memory. *Psychol. Res.* 83, 1124–1136. doi: 10.1007/s00426-017-0959-2

Goldstein, E. B. (2014). *Sensation and Perception*. Belmont: Wadsworth Cengage Learning.

Goodin, P., Ciorciari, J., Baker, K., Carrey, A. M., Harper, M., and Kaufman, J. (2012). A High-Density EEG Investigation into Steady State Binaural Beat Stimulation. *PLoS One* 7:e34789. doi: 10.1371/journal.pone.0034789

Jirakittayakorn, N., and Wongsawat, Y. (2017b). Brain responses to a 6-Hz binaural beat: effects on general theta rhythm and frontal midline theta activity. *Front. Neurosci.* 11:365. doi: 10.3389/fnins.2017.00365

Jirakittayakorn, N., and Wongsawat, Y. (2017a). Brain responses to 40-Hz binaural beat and effects on emotion and memory. *Int. J. Psychophysiol.* 120, 96–107. doi: 10.1016/j.ijpsycho.2017.07.010

Keeley, P. S. (2006). *Stress-proof your life. Beyond Stress Management*. West Conshohocken: Infinity Publishing.

Klimesch, W. (2013). An algorithm for the EEG frequency architecture of consciousness and brain body coupling. *Front. Hum. Neurosci.* 7:766. doi: 10.3389/fnhum.2013.00766

Klimesch, W. (2018). The frequency architecture of brain and brain body oscillations: an analysis. *Eur. J. Neurosci.* 48, 2431–2453. doi: 10.1111/ejn.14192

Lee, M., Song, C. B., Shin, G. H., and Lee, S. W. (2019). Possible Effect of Binaural Beat Combined With Autonomous Sensory Meridian Response for Inducing Sleep. *Front. Hum. Neurosci.* 13:425. doi: 10.3389/fnhum.2019.00425

López-Caballero, F., and Escera, C. (2017). Binaural Beat: a failure to enhance EEG power and emotional arousal. *Front. Hum. Neurosci.* 11:557. doi: 10.3389/fnhum.2017.00557

Magnusson, K. (2021). *R Psychologist*. Available Online at: https://rpsychologist.com/cohend/. (accessed February 18, 2021).

Marsh, J. T., and Worden, F. G. (1968). Sound evoked frequency-following responses in the central auditory pathway. *Laryngoscope* 78, 1149–1163. doi: 10.1288/00005537-196807000-00003

Martínez-Cancino, R., Delorme, A., Truong, D., Artoni, F., Kreutz-Delgado, K., Sivagnanam, S., et al. (2020). The Open EEGLAB portal interface: high-performance computing with EEGLAB. *Neuroimage.* 224:116778. doi: 10.1016/j.neuroimage.2020.116778

Nunez, M. D., Nunez, P. L., and Srinivasan, R. (2016). "Electroencephalography (EEG): neurophysics, experimental methods, and signal processing," in *Handbook of Neuroimaging Data Analysis*, eds H. Ombao, M. Lindquist, W. Thompson, and J. Aston (London: Chapman and Hall), 175–197.

Orozco Perez, H. D., Dumas, G., and Lehmann, A. (2020). Binaural beats through the auditory pathway: from brainstem to connectivity patterns. *eNeuro* 7:2020. doi: 10.1523/ENEURO.0232-19.2020

Park, J., Kwon, H., Kang, S., and Lee, Y. (2018). "2018 International conference on Information and Communication Technology Convergence (ICTC)," in *The Effect of Binaural Beat-Based Audiovisual Stimulation on Brain Waves and Concentration*, (Jeju: IEEE), 420–423. doi: 10.1109/ICTC.2018.8539512

Park, T. (2020). *EEG Power Spectrum*. Basel: Scholarly Community Encyclopedia.

Pejrolo, A., and Metcalfe, S. B. (2017). *Creating Sounds from Scratch: A Practical Guide to Music Synthesis for Producerss and Composers*. New York: Oxford University Press.

Renard, Y., Lotte, F., Gilbert, G., Congedo, M., Maby, E., Delannoy, V., et al. (2010). OpenViBE: an open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments. *Presence* 19, 35–53. doi: 10.1162/pres.19.1.35

Shekar, L., Suryavanshi, C. A., and Nayak, K. R. (2018). Effect of alpha and gamma binaural beats on reaction time and short-term memory. *Natl. J. Physiol. Pharm. Pharmacol.* 8, 829–833. doi: 10.5455/njppp.2018.8.1246506022018

Simmons, L. C. (2016). Binaural auditory beats, a promising therapy and cognitive enhancement. *Wheaton J. Neurosci. Sen. Semin. Res. Abril* 27, 1–7.

Siuly, S., Li, Y., and Zhang, Y. (2016). *EEG Signals Analysis and Classfication. Techniques and Applications*. Cham: Springer. doi: 10.1007/978-3-319-47653-7

Vernon, D., Peryer, G., Louch, J., and Shaw, M. (2012). Tracking EEG changes in response to alpha and beta binaural beats. *Int. J. Psychophysiol.* 93, 134–139. doi: 10.1016/j.ijpsycho.2012.10.008

Watkinson, J. C., and Clarke, R. W. (2018). *Scott-Brown¿s Otorhinolaryngology and Head and Neck Surgery*. Boca Raton: CRC Press. doi: 10.1201/9780429443558

Worden, F. G., and Marsh, J. T. (1968). Frequency-following (microphonic-like) neural responses evoked by sound. *Electroencephalogr. Clin. Neurophysiol.* 25, 42–52. doi: 10.1016/0013-4694(68)90085-0

World Health Organization [WHO], and International Telecommunication Union [ITU] (2019). *Safe Listening Devices and Systems: A WHO-ITU Standard*. Geneva: WHO.

Zhang, Z. (2019). "Chapter 6. Spectral and time-frequency analysis," in *EEG Signal Processing and Feature Extraction*, eds L. Hu and Z. Zhang (Singapore: Springer Nature), 89–116. doi: 10.1007/978-981-13-9113-2_6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Digital Health

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fdgth.2021.799067

CovNet: A Transfer Learning Framework for Automatic COVID-19 Detection From Crowd-Sourced Cough Sounds

Yi Chang [1]*, *Xin Jing* [2], *Zhao Ren* [2,3]* *and Björn W. Schuller* [1,2]

[1] Group on Language, Audio, and Music, Imperial College London, London, United Kingdom, [2] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [3] L3S Research Center, Hannover, Germany

Since the COronaVIrus Disease 2019 (COVID-19) outbreak, developing a digital diagnostic tool to detect COVID-19 from respiratory sounds with computer audition has become an essential topic due to its advantages of being swift, low-cost, and eco-friendly. However, prior studies mainly focused on small-scale COVID-19 datasets. To build a robust model, the large-scale multi-sound FluSense dataset is utilised to help detect COVID-19 from cough sounds in this study. Due to the gap between FluSense and the COVID-19-related datasets consisting of cough only, the transfer learning framework (namely CovNet) is proposed and applied rather than simply augmenting the training data with FluSense. The CovNet contains (i) a parameter transferring strategy and (ii) an embedding incorporation strategy. Specifically, to validate the CovNet's effectiveness, it is used to transfer knowledge from FluSense to COUGHVID, a large-scale cough sound database of COVID-19 negative and COVID-19 positive individuals. The trained model on FluSense and COUGHVID is further applied under the CovNet to another two small-scale cough datasets for COVID-19 detection, the COVID-19 cough sub-challenge (CCS) database in the INTERSPEECH Computational Paralinguistics challengE (ComParE) challenge and the DiCOVA Track-1 database. By training four simple convolutional neural networks (CNNs) in the transfer learning framework, our approach achieves an absolute improvement of 3.57% over the baseline of DiCOVA Track-1 validation of the area under the receiver operating characteristic curve (ROC AUC) and an absolute improvement of 1.73% over the baseline of ComParE CCS test unweighted average recall (UAR).

Keywords: transfer learning, COVID-19, cough, FluSense, COUGHVID

Edited by:
Harry Hochheiser,
University of Pittsburgh, United States

Reviewed by:
Samrat Kumar Dey,
Bangladesh Open University,
Bangladesh
K. C. Santosh,
University of South Dakota,
United States

***Correspondence:**
Yi Chang
y.chang20@imperial.ac.uk
Zhao Ren
zren@l3s.de

Specialty section:
This article was submitted to
Health Informatics,
a section of the journal
Frontiers in Digital Health

Received: *21 October 2021*
Accepted: *29 November 2021*
Published: *03 January 2022*

Citation:
Chang Y, Jing X, Ren Z and Schuller BW (2022) CovNet: A Transfer Learning Framework for Automatic COVID-19 Detection From Crowd-Sourced Cough Sounds. Front. Digit. Health 3:799067. doi: 10.3389/fdgth.2021.799067

1. INTRODUCTION

Since the year 2019, the coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become a global pandemic[1]. As of August 2021, there have been more than 202,000,000 confirmed cases of COVID-19 worldwide, including more than 4,000,000 deaths, reported by the World Health Organization (WHO)[2]. The daily

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it; retrieved 10 August 2021.

[2] https://covid19.who.int/; retrieved 10 August 2021.

increasing COVID-19 cases and deaths have resulted in global lockdown, quarantine, and many restrictions (1). Along with the above measures, a set of following problems have appeared, including the economic downturn (2) and mental health problems (e.g., depression and stress) (1).

Swift and accurate diagnosis of COVID-19 is essential to give patients appropriate treatments and effectively control its transmission (3). The reverse transcription PCR (RT-PCR) from oral-nasopharyngeal swabs identifies viral RNA and is a commonly used instrument for the diagnosis of COVID-19. Nevertheless, high false negative rate and stability issues have been reported (4). In contrast to RT-PCR, chest CT was proven to have high sensitivity and be expedited for diagnosing COVID-19(4). Serological instruments are utilised to diagnose/confirm late COVID-19 cases by measuring antibody responses to the corresponding infection (5). Compared to the above laboratory instruments, which require professionals and special medical equipment, rapid antigen and molecular tests using nasopharyngeal swabs are commercially available due to their swift and simple test procedures, reduced mortality of COVID-19 patients, internal hospital costs, and in-hospital transmission (6). However, rapid tests are still hard-to-follow for non-specialists and are not environment-friendly.

Artificial intelligence has been widely applied to respiratory sounds in the healthcare area (7–9). In a study by (8), a multilayer perceptron based classifier was developed on features extracted from respiratory sounds to screen lung health. Random forests are applied on the filter bank energy-based features to pre-screen the lung health abnormalities (9). COVID-19 patients were reported to have seven common symptoms, including fever, cough, sore throat, headache, myalgia, nausea/vomiting, and diarrhea (10). Among these symptoms, the first two symptoms of COVID-19 are fever and cough (10). As a fast and non-invasive way to detect potential infections in public areas, body temperature measurement has been commonly employed (11). Traditional body temperature measurement with a thermometer usually requires relatively close contact with potential COVID-19 positive individuals (12). Although infrared (IR) thermal cameras provide a non-contact way for mass fever detection, they may not be valid because of the absence of calibration, non-homogeneous devices/protocols, and poor correlation between skin temperature and core body temperature (11). The reading of IR thermal cameras could also be affected by the environmental temperature (11). On the other hand, cough, as a common symptom in many respiratory diseases, is a worthwhile consideration when diagnosing a disease (13). Cough sounds have been used to diagnose asthma, bronchitis, pertussis, pneumonia, etc. (13). Recent studies have also investigated the feasibility of detecting COVID-19 infections from cough sounds. For instance, cough sounds were shown to contain latent features distinguishable between COVID-19 positive individuals and COVID-19 negative individuals (i.e., normal, bronchitis, and pertussis) (14). In Brown et al.'s study (15), cough sounds from COVID-19 positive individuals were reported to have a longer duration, more onsets, higher periods, lower RMS, and MFCC features with fewer outliers. Due to the development of the internet-of-things (IoT), the algorithms for detecting potential COVID-19 positive individuals from cough sounds can be integrated into mobile phones, wearable devices, and robots. Such a rapid, easy-to-use, and environment-friendly instrument will be helpful for real-time and remote pre-screening of COVID-19 infections, thereby supplementing clinical diagnosis and reducing the medical burden.

Since the outbreak of COVID-19, several studies have collected cough samples from COVID-19 positive patients (and COVID-19 negative individuals) to detect COVID-19 infections. Coswara (16) is a crowd-sourced database consisting of various kinds of sounds, including breathing (shallow and deep), coughing (shallow and deep), sustained vowel phonation (/ey/ as in made, /i/ as in beet, /u:/ as in cool), and number counting from one to twenty (normal and fast-paced). Another crowd-sourced database, COUGHVID with cough sounds only (17), was collected via a web interface. To date, the latest version of COUGHVID is publically released with $27,550$ cough recordings.[3] The crowd-sourced University of Cambridge COVID database was reported to have more than 400 cough and breathing recordings (15). The Virufy datasets consist of a Latin American crowd-sourced dataset (31 individuals) and two South Asian clinical datasets (362 and 63 individuals, respectively). Due to the difficulty of collecting cough sounds of confirmed COVID-19 patients and multi-sound (non-cough)/noise in crowd-sourced datasets, most of the above databases are small-scale, leading to a challenge for training robust machine learning models.

With this in mind, we propose a hybrid transfer learning framework for robust COVID-19 detection, where several convolutional neural networks (CNNs) are trained on large-scale databases and fine-tuned on several small-scale cough sound databases for verification. Note that the focus of this paper is not to outperform the state-of-the-art neural networks models for COVID-19 detection from cough sounds; rather, the aim of this study is to provide a framework for mitigating the effect of noise or irrelevant sounds in the crowd-sourcing datasets applied to COVID-19 by training robust CNN models with the transferred knowledge from Flusense and/or COUGHVID. The workflow of this study is indicated in **Figure 1**. The code of this paper is publicly available on GitHub[4].

- The FluSense database (18) was collected in a platform to track influenza-related indicators, such as cough, sneeze, sniffle, and speech. Since it contains various types of sounds existing in crowd-sourced cough datasets, the FluSense dataset is applied in this study.
- Due to the gap in sound type between FluSense and databases with cough sounds only, the COUGHVID database is considered as the target data when CNNs are trained on FluSense as the source data. The trained models on COUGHVID are further adapted to the other two smaller test databases, i.e., Computational Paralinguistics challengE (ComParE) 2021 COVID-19 cough sub-challenge (CCS) (19) and DiCOVA 2021 Track-1 (20).

[3] https://zenodo.org/record/4498364#.YRKa3IgzbD4

[4] https://github.com/ychang74/CovNet

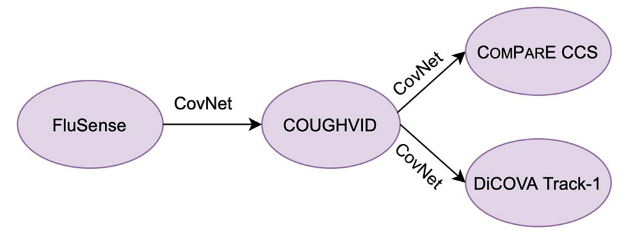


FIGURE 1 | The workflow of this study. CovNet is the proposed transfer learning framework, which includes transferring parameters and incorporating embeddings. CovNet is first applied on the Flusense as the source data, COUGHVID as the target data. Afterwards, to further validate the effectiveness of CovNet, the CovNet based pre-trained COUGHVID models are applied on two smaller Computational Paralinguistics challengE (ComParE) 2021 COVID-19 cough sub-challenge (CCS) dataset and DiCOVA 2021 Track-1 dataset.

- We propose two transfer learning pipelines, i.e., transferring parameters from the source database to the target database for fine-tuning models and incorporating embeddings for expanding models' capability of extracting useful features.

In the following sections, the transfer learning framework is first introduced in section 2, followed by the architecture of the models for COVID-19 detection in section 3. Next, the experimental details are described, and the results are presented and discussed in section 4. Finally, our study is summarised, and the outlook is given in section 5.

2. TRANSFER LEARNING FRAMEWORKS

Transfer learning aims at applying the knowledge learnt from source data to different but related target data and achieving better performance in a cost-effective way (21–23). The source data and target data should be similar, otherwise negative transfer may happen (22, 24). Transfer learning has been successfully applied to COVID-19 detection based on acoustic data (14, 15). In Imran et al.'s study (14), the knowledge was transferred from the cough detection model to the COVID-19 diagnosis model. Brown et al. (15) discovered that VGGish pre-trained on a large-scale YouTube dataset was utilised to extract audio features from raw audio samples for COVID-19 diagnosis.

In this study, two ways of transfer learning are applied. One is to fine-tune the parameters of the networks with the target data. The other is extracting the embeddings from the pre-trained network and applying the embeddings when training the new network for the target dataset. Since the crowd-sourced cough recordings usually contain non-cough audio signals other than cough sounds, such as speech and breathing, the FluSense dataset and the COUGHVID dataset contain similar sound types. Therefore, the knowledge learnt from FluSense data can be employed to improve the performance of models trained on the COUGHDVID dataset. In **Figure 2**, $D_{FluSense}$ is the FluSense dataset, and $D_{COUGHVID}$ means the COUGHVID dataset; convs0 and convs1 represent the convolutional layers/blocks in the neural networks on the FluSense dataset and the COUGHVID dataset, respectively; $FC_{FluSense}$ and $FC_{COUGHVID}$ denotes the fully-connected (FC) layer of corresponding models. When separating the left part with the right part in **Figures 2A,B**, with the training data (x_0, y_0) and (x_1, y_1), we separately train the CNNs on the FluSense and COUGHVID datasets to produce the predicted values $\hat{y}_0$ and $\hat{y}_1$, respectively.

With the parameters and embeddings from the pre-trained FluSense models, as highlighted in blue in **Figure 2**, the COUGHVID models are given the potential to discriminate between the various audio signals, which further helps its COVID-19 detection from crowd-sourced cough signals. Notably, the predicted value $\hat{y}_1$ is the final output of the proposed transfer learning framework.

To further investigate the generalisation ability of CovNet, we apply it to some other small-scale crowd-sourced datasets for COVID-19 detection. In the following, we introduce the two transfer learning methods in greater detail.

2.1. Transferring Parameters

Fine-tuning pre-trained models is an effective transfer learning method by sharing some parameters across tasks (21, 22). In the computer vision area, parameters of pre-trained models on ImageNet (25) are often applied for transfer learning on a wide range of image-related tasks (26–29). Similarly, parameters of pre-trained models on the Audio Set are transferred to many audio-related tasks (30–32). Parameters of pre-trained CNN models on the Audio Set are transferred to the adapting networks for acoustic event recognition (30, 31). Several pre-trained audio neural networks trained on the Audio Set dataset were proposed for other audio pattern recognition tasks (32).

In this study, as indicated in **Figure 2A**, the parameters of the first n convolutional layers/blocks, $convs1_{1,2,...,n}$, of models trained on the COUGHVID dataset, are initialised by the corresponding layers/blocks $convs0_{1,2,...,n}$ of models pre-trained on FluSense dataset. The parameters of $convs1_{1,2,...,n}$ are frozen and not trained, and only the remaining randomly initialised parameters of $convs1_{n+1,n+2,...,N}$ and $FC_{COUGHVID}$ are updated during the training procedure.

2.2. Incorporating Embeddings

The embeddings generated by the convolutional layers carry either low-level edge information or high-level discrimination-related features (22, 23). Moreover, the performance of embeddings appears to be highly scalable with the amount of training data (33). In this study, the pre-trained FluSense models produce embeddings representing high-level or low-level characteristics of various audio types, which can be applied as an additional input to help develop the target model.

Specifically, we feed the crowd-sourced cough recordings from the COUGHVID into the pre-trained Flusnese model and extract the embeddings after certain convolutional layers/blocks. **Figure 2B** exhibits this strategy. Data-point (x_1, y_1) enters the pre-trained FluSense model, and the output embeddings of the n-th convolutional layer/block $convs0_n$ are extracted to be concatenated (on the channel dimension) or added with the embeddings generated by the corresponding $convs1_n$. The

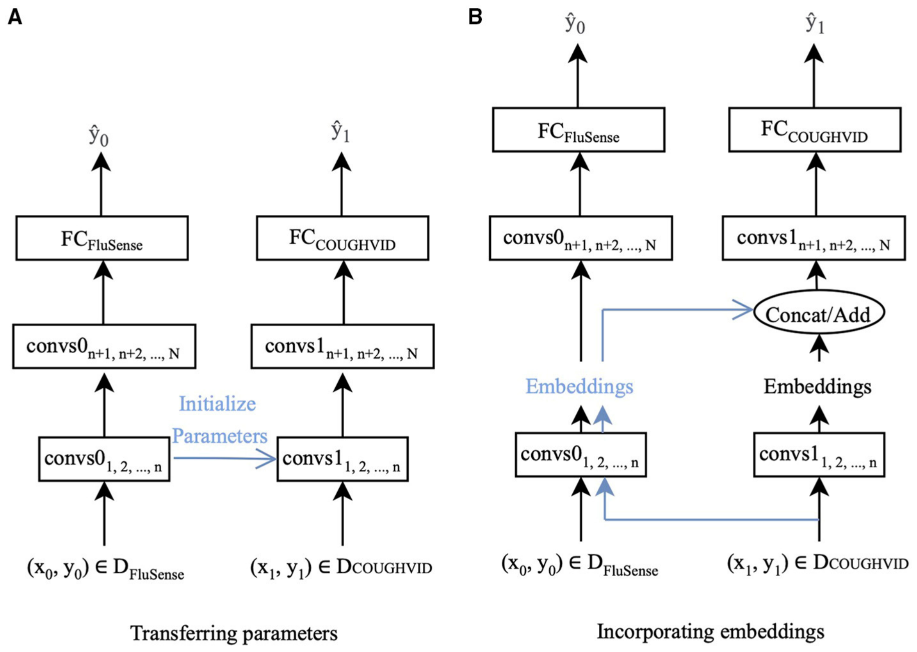


FIGURE 2 | The proposed transfer learning framework :CovNet. **(A)** Parameters of the first *n* convolutional layers/blocks (convs1) of the current COUGHVID model are frozen and initialised by the corresponding first *n* convolutional layers/blocks (convs0) of the pre-trained FluSense model. **(B)** Embeddings are extracted after the *n*-th convs0 of the pre-trained FluSense model. The extracted embeddings are concatenated or added to the current embeddings generated after the *n*-th convs1 of the COUGHVID model.

concatenated or added embeddings enter the next convolutional layer/block convs1_{n+1} for the task of COVID-19 detection.

3. AUTOMATIC COVID-19 DETECTION

Convolutional neural networks have been successfully applied in image-related areas, such as image classification (34–37). When processing audio signals, CNNs have demonstrated their capabilities in extracting effective representations from the log Mel spectrograms (38, 39). In this study, we choose four typical CNN models: base CNN (34), VGG (40), residual network (ResNet) (41), and MobileNet (42). We focus on the proposed transfer learning framework, CovNet, instead of competing with the state-of-the-art models on COVID-19 detection. Therefore, in order to highlight the effectiveness of CovNet, we construct four simple CNN models (i.e., CNN-4, VGG-7, ResNet-6, and MobileNet-6), each of which only has three convolutional layers/blocks. A detailed description of each model is given and analysed in the following subsections.

The log Mel spectrograms are calculated by Mel filter banks and logarithmic operation worked on the spectrograms, which are produced by the Short-Time Fourier Transforms (STFTs) on the original waveforms. In this section, to better evaluate the effectiveness of the proposed transfer learning framework and compare the performance differences among different CNN architectures, four CNNs are employed to deal with the extracted log Mel spectrograms: CNN-4, VGG-7, ResNet-6, and MobileNet-6. Log Mel spectrograms (T,F) are extracted from the audio signals as the input to the CNNs, where T represents the sequence length, and F denotes the log Mel frequency. Before entering the final FC layer, the matrix has the dimension (C_N, N), where C_N is the output channel number of the last convolutional layer, and N is the class number. Specifically, for the FluSense database, N is set to be 9; for the other datasets used in this study, N equals 2. For comparison convenience, we regard the convolutional layers and blocks equally when ordering them in a specific model. In this notation, ResNet-6 and MobileNet-6 have "block2" following the first convolutional layer.

3.1. CNN-4

As shown **Figure 3A**, we propose a simple 4-layer CNN, CNN-4, constructed by three 5 × 5 convolutional layers. To speed up and stabilise the training procedure, each convolutional layer is followed by batch normalisation (43) and the Rectified Linear Unit (ReLU) activation function (44). Afterwards, we apply max pooling for downsampling. The first three local max pooling operations are conducted over a 2 × 2 kernel, and the last max pooling is a global one to summarise the features along the dimension of the sequence length and frequency. Before the final FC layer for the final predicted result, a dropout (45) layer is utilised to address the overfitting issue.

3.2. VGG-7

Very deep CNN, known as VGG, were originally designed with up to 19 weight layers and achieved great performance on the large-scale image classification task (40, 46). VGG or VGG-like architectures were applied to extract audio features from respiratory sound data for COVID-19 detection and obtained good performances (15, 47).

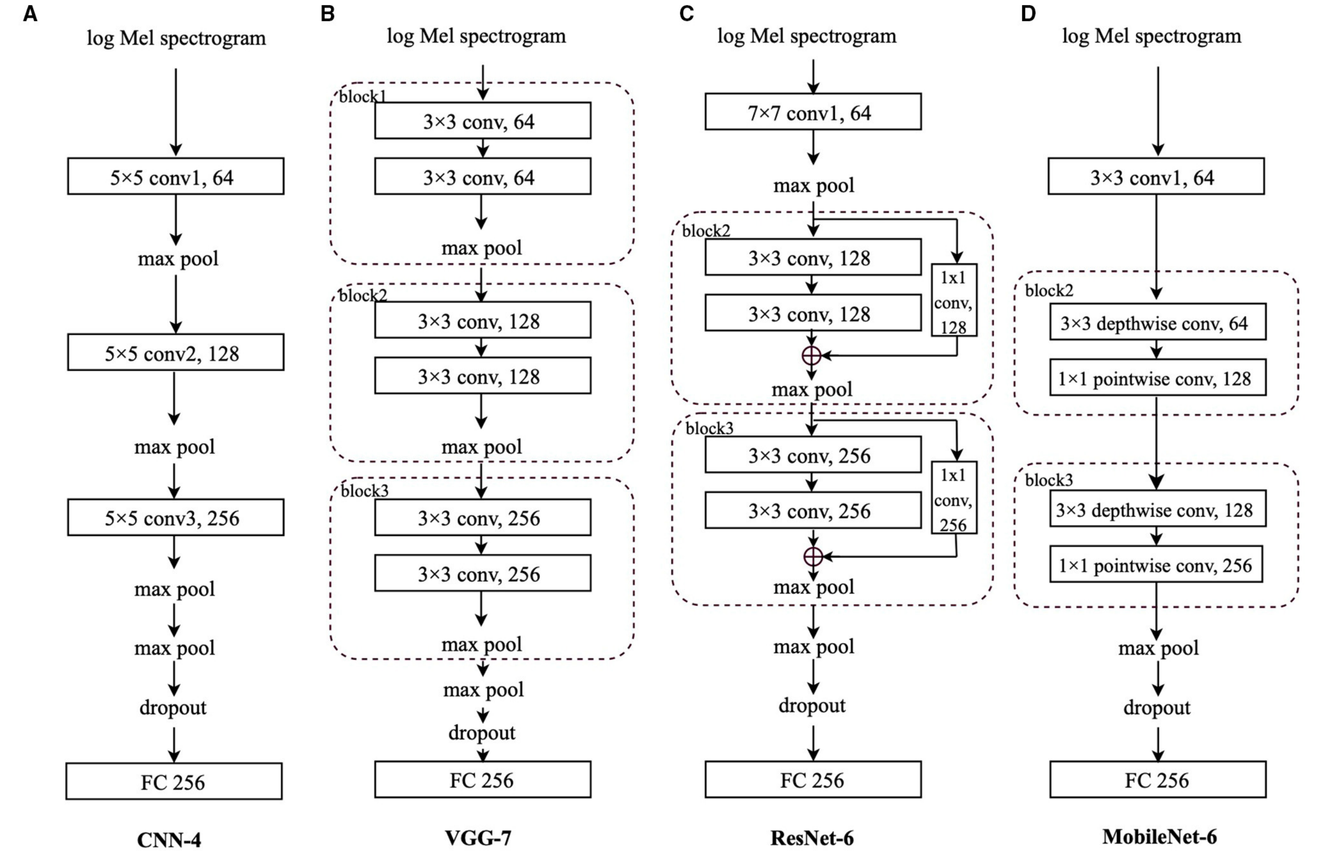


FIGURE 3 | Models' architecture: **(A)** Convolutional neural network-4 (CNN-4), **(B)** VGG-7, **(C)** residual network-6 (ResNet-6), **(D)** MobileNet-6. "conv" stands for the convolutional layer, and "block" indicates the convolutional block. The number before the "conv" is the kernel size; the number after the "conv" is the output channel number. The number after "FC" is the input neurons' size.

As indicated in **Figure 3B**, we adapt the VGG (40) with 7 layers, VGG-7, which is composed of three convolutional blocks and a final FC layer. Although the VGG-7 is simple, different from its original "deep" design, it is still worthwhile to include it for fair comparison with other CNNs in this study. Each block contains two 3 × 3 convolutional layers, each of which is followed by batch normalisation (43) and the ReLU function (44) to stabilise and speed up the training process. Afterwards, a local max pooling layer with a kernel size of 2 × 2 is applied. Following the three blocks, there is also a global max pooling layer working on the sequence length and log Mel frequency dimensions. Before the FC layer, a dropout (45) layer is applied.

3.3. ResNet-6

The Deep ResNet is proposed to address the degradation problem existing in training deeper networks (41) by incorporating shortcut connections between convolutional layers. In Hershey et al.'s (48) study,ResNet has outperformed other CNNs for audio classification on the Audio Set (49). A ResNet based model is constructed for COVID-19 detection from breath and cough audio signals (50).

In this study, we mainly adopt the above mentioned shortcut connections to construct a 6-layer ResNet, ResNet-6. In **Figure 3C**, after the first convolutional layer with a kernel size of 7 × 7 followed by batch normalisation (43) and the ReLU function (44), we apply two convolutional blocks, each of which contains the "shortcut connections" to add the identity mapping with the outputs of two stacked 3 × 3 convolutional layers.

Inside "block2" and "block3," after the first 3 × 3 convolutional layer, the batch normalisation (43) and ReLU function (44) are applied, whereas only the batch normalisation is utilised after the second 3 × 3 convolutional layer. For the channel number consistency, the identity is processed by a 1 × 1 convolutional layer followed by batch normalisation (43); after the addition of the identity and the output of two stacked convolutional layers, we apply the ReLU function (44). The max pooling after the 7 × 7 convolutional layer is a local one with a kernel size of 3 × 3 and the max pooling layers in "block2" and "block3" are also local with a kernel size of 2 × 2; similarly, the last max pooling is a global one, followed by a dropout (45) layer and the FC layer.

3.4. MobileNet-6

Based on depthwise separable convolutions, light-weight MobileNets have been widely applied in mobile and embedded image related applications (42, 51). MobileNets are cost-effective and are explored herein for potential solutions embedded in mobile devices for COVID-19 detection.

We adapt the MobileNet with 6 layers only. As shown in **Figure 3D**, after the first 3 × 3 convolutional layer followed by batch normalisation (43) and the ReLU function (44), each of "block2" and "block3" contains a 3 × 3 depthwise convolutional layer and a 1 × 1 pointwise convolutional layer, respectively. Similarly, batch normalisation (43) and ReLU function (44) are applied after each convolutional layer. Similar to the original MobileNet architecture, we only set one global max pooling layer before the dropout (45) layer and the final FC layer.

4. EXPERIMENTAL RESULTS

With the aforementioned transfer learning framework, the experiments will be presented in this section, including the databases, experimental setup, results, and discussions.

4.1. Databases

To verify the proposed transfer learning framework in this study, the following four datasets are employed.

4.1.1. FluSense

The FluSense (18) project applied a part of the original Audio Set dataset (49), which includes weakly labelled 10-s audio clips from YouTube. After the re-annotation by two human raters for more precise labels in the FluSense (18) project, there are a total of 45, 550 seconds samples in Audio Set that are considered in this study, and they are labelled with the classes of *breathe*, *burp*, *cough*, *gasp*, *hiccup*, *other*, *silence*, *sneeze*, *sniffle*, *snore*, *speech*, *throat-clearing*, *vomit*, and *wheeze*. To mitigate the effect of data imbalance on the classification performance, those classes with a number of samples less than 100 are not considered in our experiments. Therefore, the audio samples labelled with the following nine classes are employed: *breathe*, *cough*, *gasp*, *other*, *silence*, *sneeze*, *sniffle*, *speech*, and *throat-clearing*. For all audio recordings in the above nine classes, we first re-sampled them into 16 kHz. Second, as the audio samples have various time lengths, we split the original samples with a length of greater than or equal to 0.5 s into one or more 1 s segment(s). In particular, for audio samples with a length between 0.5 and 1 s, the audio repeats itself until a full 1 s segment is reached. For those samples with a length greater than 1 s, after a certain number of 1 s segments are split, the remaining signals repeat themselves until a full segment is reached if the remaining one has a length of greater than or equal to 0.5 s; otherwise, the remaining signals are simply abandoned. Furthermore, we split the segments into train/val subsets with a ratio of 0.8/0.2 in a stratified manner. The data distribution of FluSense before and after the pre-processing is shown in **Table 1**.

TABLE 1 | Data distribution of the FluSense data.

	Original	Pre-Processing		
	#	Train	Val	$\sum$
Breathe	167	238	58	297
Cough	2,486	6,148	1,537	7,685
Gasp	337	315	79	394
Other	3,863	15,059	3,765	18,824
Silence	832	1,116	279	1,395
Sneeze	611	540	135	675
Sniffle	589	604	151	755
Speech	2,615	16,614	4,154	20,768
Throat clearing	102	118	29	147
$\sum$	11,602	40,752	10,188	50,940

The "original" column indicates the number of audio samples; whereas the "pre-processing" columns show the number of segments with unified length of 1 s.

TABLE 2 | Data distribution of the COUGHVID data.

#	Train	Test	$\sum$
Negative	5,660	1,415	7,075
Positive	559	140	699
$\sum$	6,219	1,555	7,774

4.1.2. COUGHVID

The on-going crowd-sourced COUGHVID dataset (17) is collected via a web interface[5]. All participants voluntarily record and upload their cough sounds lasting for up to 10 s. In the meantime, the COVID-19 status of each cough sample is self-reported by each participant: *healthy*, *symptomatic without COVID-19 diagnosis*, and *COVID-19*. The information of each participant is optionally self-reported, including the geographic location (latitude, longitude), age, gender, and whether she/he has other pre-existing respiratory conditions, and muscle pain/fever symptoms. As there might be some low-quality audio samples (e.g., noise, speech, etc.), the data collectors trained an extreme gradient boosting (XBG) classifier on 215 audio samples (121 cough and 94 non-cough) to predict the probability of a recording containing cough sounds. For all audio recordings, the sampling frequency is 48 kHz.

In this study, only the classes of *healthy* (i.e., COVID-19 negative) and *COVID-19* (i.e., COVID-19 positive) are considered, as the audio samples with symptomatic status were not explicitly reported by the participants as to whether they were diagnosed with COVID-19 or not. Furthermore, only audio samples with cough sound probabilities greater than 0.9 are included to ensure each audio sample contains cough sounds. Finally, 7, 774 audio samples (COVID-19 negative: 7, 075, COVID-19 positive: 699) are selected for our experiments. Similarly, we split the selected samples into train/test subsets with a ratio of 0.8/0.2, respectively in a stratified manner. **Table 2** shows the data distribution of COUGHVID.

4.1.3. ComParE 2021 CCS

In the INTERPSEECH 2021 ComParE (19), the CCS provides a dataset from the crowd-sourced Cambridge COVID-19 Sound database (15). The participants are asked to provide one to three forced coughs in each recording via one of the following multiple

[5] https://COUGHVID.epfl.ch/; retrieved 09 July 2021.

TABLE 3 | Data distribution of the Computational Paralinguistics challengE (ComParE) COVID-19 cough sub-challenge (CCS) data.

#	Train	Val	Test	∑
Negative	215	183	169	567
Positive	71	48	39	158
∑	286	231	208	725

TABLE 4 | DiCOVA Track-1 data distribution of each fold of cross-validation.

#	Train	Val	∑
Negative	772	193	965
Positive	50	25	75
∑	822	218	1040

platforms: A web interface, an Android app, and an iOS app.[6] The CCS dataset consists of 929 cough recordings (1.63 h) from 397 participants. The data distribution of CCS is shown in **Table 3**. All recordings from the CCS dataset were resampled and converted into 16 kHz. The official training, validation, and test sets in the ComParE challenge are used in this study.

4.1.4. DiCOVA 2021 Track-1

The Track-1 of the DiCOVA challenge 2021 (20) provides cough recordings from 1, 040 participants (COVID-19 negative: 965, COVID-19 positive 75). In the challenge, the dataset was split into five train-validation folds. Each training set consists of 822 cough samples (COVID-19 negative: 772, COVID-19 positive: 50), and each validation set contains 218 cough samples (COVID-19 negative: 193, COVID-19 positive: 25). The additional test set is not used in this study, as it is blind. All cough recordings are sampled at 44.1 kHz. The data distribution of DiCOVA 2021 Track-1 is indicated in **Table 4**.

4.2. Experimental Setup

For faster progress (38), all audio files in the four datasets are re-sampled into 16 kHz. The log Mel spectrograms are extracted with a sliding window size of 512, an overlap of 256 units, and 64 Mel bins.

As for the evaluation metrics, we mainly use unweighted average recall (UAR), since it is more adequate for evaluating the classification performance on imbalanced datasets than accuracy,—the weighted average recall (52, 53). Apart from the UAR, we also calculate the area under the receiver operating characteristic curve (ROC AUC) score.

The proposed CNNs consist of three convolutional layers/blocks. The number of output channels for the three convolutional layers/blocks is 64, 128, and 256, respectively. During the training procedure of the neural networks, the cross-entropy loss is utilised as the loss function. To overcome the class imbalance issue, we re-scale the weight parameter for each class in the loss function. Since this study focuses on the transfer learning framework, we do not further mitigate the class imbalance issue through down-/up-sampling.

For single learning (i.e., training from scratch) on the FluSense and the COUGHVID datasets, the optimiser is set to "Adam" with an initial learning rate of 0.001, which is scheduled to be reduced by a factor of 0.4 when there is less than 0.01 improvement of the UAR after every 4 of 30 epochs in total. When transferring parameters, we set the initial learning rate as 0.0001; for incorporating embeddings, the initial learning rate is set to be 0.001.

When applying the strategy of transferring parameters introduced in section 2.1 to training the COUGHVID model, we experiment with only setting the following layer(s) trainable: the FC layer, the convolutional layer/block (conv/block) 3 & FC layer, conv/block 2 − 3 & FC layer, and conv/block 1 − 3 & FC layer, respectively. The remaining layer(s)/block(s) are initialised based on the pre-trained FluSense models' corresponding parameters and are frozen during the whole training procedure. As for the incorporating embeddings strategy described in section 2.2, we investigate the concatenation and addition of two embeddings generated from the conv/block 3, conv/block 2, and conv/block 1, respectively. One embedding is from the pre-trained FluSense model, and the other one is the COUGHVID model trained from scratch.

To further validate the effectiveness of the CovNet, we apply the pre-trained COUGHVID models on the ComParE CCS dataset and the DiCOVA Track-1 dataset. Specifically, we train the four CNNs introduced in section 3 from scratch. Afterwards, we choose up to two COUGHVID models with the best performance (best AUC or best UAR) as the pre-trained models. With the chosen pre-trained COUGHVID models and their strategies (layer(s)/block(s) number and transfer learning strategies), we transfer the parameters or embeddings of the above chosen COUGHVID models to the current train-from-scratch models on the ComParE and DiCOVA datasets during the training. Finally, we choose the best results to compete with official baselines: the average validation AUC 68.81% (20) for the DiCOVA Track-1 dataset, and test UAR without fusion 64.7% (19) for ComParE CCS. Similarly, when training models from scratch or applying the incorporating embeddings method, we set the initial learning rate as 0.001, whereas if the transferring parameters are utilised, the initial learning rate is set as 0.0001.

4.3. Results

In **Table 5**, we focus on performance differences on the COUGHVID test dataset between single learning (training from scratch) models and the models produced by the proposed transfer learning strategies in section 2. For convenience, the best test AUC and test UAR of every model under three transfer learning strategies are shown in bold face. We can see that there are some improvements in test AUC/UAR, especially for the VGG-7 and MobileNet-6. In the following analysis, we compare the absolute difference between performances. On the COUGHVID test dataset, with the transfer learning, the VGG-7 obtains an improvement of 2.62% AUC ($p < 0.1$ in a one-tailed z-test) and an improvement of 3.75% UAR ($p < 0.05$ in a one-tailed z-test); the MobileNet-6 achieves 3.77% improvement in

[6] https://www.covid-19-sounds.org/; retrieved 15 July 2021

TABLE 5 | Models' performances [AUC/UAR %] on FluSense and COUGHVID test datasets.

		Layers	CNN-4	ResNet-6	VGG-7	MobileNet-6
Single Learning	FluSense	—	93.55/65.27	93.91/64.76	93.23/63.86	91.26/58.24
	COUGHVID	—	66.14/59.43	68.86/60.43	65.15/56.42	64.17/54.83
Transfer Learning	Parameters	FC	58.59/53.68	61.35/57.50	54.68/54.14	56.91/53.93
		conv/block 3 & FC	68.04/57.04	67.01/57.97	64.97/57.15	67.88/**59.71**
		conv/block 2-3 & FC	69.05/**60.98**	**67.89**/**59.25**	64.92/**59.79**	**67.94**/58.93
		conv/block 1-3 & FC	**69.43**/55.54	66.23/56.31	**67.31**/56.17	65.21/55.64
	Embeddings Cat	conv/block 3	**67.73**/**60.65**	**67.21**/59.45	**65.85**/**58.27**	64.32/**56.46**
		conv/block 2	67.30/57.81	66.17/55.59	65.58/52.30	**67.36**/52.31
		conv/block 1	65.15/59.30	65.35/**59.77**	58.67/51.92	66.37/53.77
	Embeddings Add	conv/block 3	**66.76**/**59.30**	64.27/**58.88**	66.08/**60.17**	**65.94**/**58.24**
		conv/block 2	66.39/58.82	64.55/57.27	**67.77**/58.55	64.37/57.19
		conv/block 1	65.91/57.17	**64.63**/58.21	63.85/58.97	64.17/56.60

Single learning indicates training from scratch and transfer learning includes "Parameters" (transferring parameters), "Embeddings Cat," and "Embeddings Add" (incorporating emebdddings). The Models' performances with transfer learning are based on the COUGHVID dataset. For "Parameters," the "Layers" column indicates the layers that are randomly initialised and trainable during the training procedure, and the remaining layers are frozen and initialised by the pre-trained FluSense models; for "Embeddings Cat," "Embeddings Add," and "Layers," the column lists the convolutional layer/block (conv/block), after which embeddings incorporation happens. For convenience, the best test AUC and test UAR of every model under three transfer learning strategies are shown in bold face.

AUC ($p < 0.05$ in a one-tailed z-test) and 4.88% improvement in UAR ($p < 0.005$ in a one-tailed z-test). Moreover, for all constructed CNN models, only setting the FC layer trainable and freezing other layers with parameters transferred from pre-trained FluSense models achieves almost the lowest AUC/UAR among all transfer learning settings.

For the transferring parameters strategy, we can see that most best test AUC/UAR cases are obtained by only setting the convolutional layer/block (conv/block) $2-3$ & FC layer trainable or the conv/block $1-3$ & FC layer trainable. With the embeddings cat method, models' performances are mostly better than single learning models' and the most best results are achieved by concatenating the embeddings output by the conv/block 3. With the embeddings addition method, models also mostly outperform the single learning ones, and similarly, most best results are obtained by adding embeddings after the conv/block 3.

In **Table 6**, first, we can see that with the proposed transfer learning strategies on the pre-trained COUGHVID models generated by the CovNet, most of the models' performances improve a lot compared with the single learning models' performance. Specifically, transferring parameters improves the test UAR on ComParE by 9.05% for the VGG-7 ($p < 0.05$ in a one-tailed z-test); the transferring parameters improves the validation AUC on DiCOVA by 1.12, 3.86, and 5.22 % for the CNN-4, ResNet-6, and VGG-7, respectively (in a one-tailed z-test, not significant, $p < 0.05$, and $p < 0.005$, respectively). The incorporating embeddings improves the test UAR on ComParE data by 1.47, and 1.11% for the CNN-4, and VGG-7, respectively; the incorporating embeddings improves the validation AUC of DiCOVA by 3.62%, 8.85, 7.46, and 2.20% for the CNN-4, ResNet-6, VGG-7, and MobileNet-6, respectively (in a one-tailed z-test, $p < 0.05$, $p < 0.001$, $p < 0.001$ and not significant, respectively).

Second, as the numbers in bold indicate better performance than the baseline, we can see that most models learnt through the transfer learning framework outperform the official baselines, even though the models here are quite simple. Notably, the best test UAR 66.43% on ComParE CCS data is achieved by the VGG-7 with transferring parameters, which is 1.73% above the official baseline; the CNN-4 with incorporating embeddings the achieves the best validation AUC 72.38% on the DiCOVA Track-1, which is 3.57% higher than the baseline ($p < 0.05$ in a one-tailed z-test). **Figure 4** displays the confusion matrices for above-mentioned best UAR on the ComParE CCS dataset and best validation AUC on the DiCOVA Track-1 dataset. We can see that the models recognise negative samples very well, but the positive ones are frequently confused with the negative ones.

4.4. Discussion

In **Table 5**, if comparing the performance of single learning CNNs and transfer learning CNNs, we find that there is no improvement or even slightly worse performance of transfer learning methods on the ResNet-6 model. ResNet gains accuracy from increased neural network depth (41), which may explain the performance of the simple ResNet-6 in this study. Apart from fine-tuning the parameters of FC layers only, almost all other CNN models obtain better performance after the transfer learning, proving the usefulness of the knowledge transferred from the FluSense dataset for recognising COVID-19 on the COUGHVID dataset. Setting FC layers trainable only limits the generalisation of the pre-trained FluSense models.

For fine-tuning parameters of different layers, fine-tuning the weights of the convolutional layers/blocks $2-3$ & FC layer obtains better performance. Since the target dataset COUGHVID is not large-scale enough compared with the FluSense one, fine-tuning the entire network (convolutional layers/blocks $1-3$ &

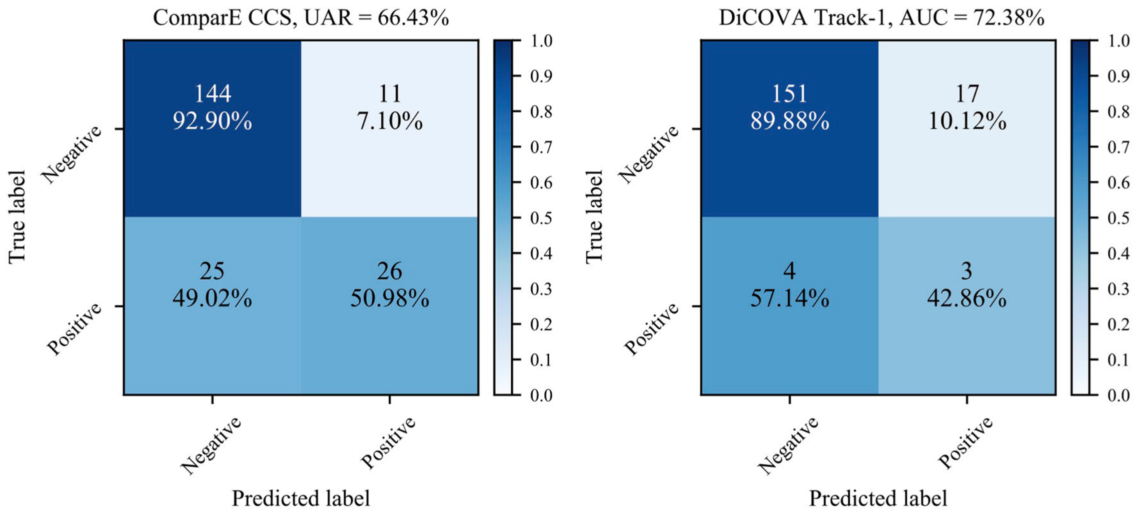


FIGURE 4 | Confusion matrices for the best performance on the COMPARE CCS test set and the DiCOVA validation set. For the DiCOVA dataset, since its test dataset is not accessible, the numbers are averaged over the five cross-validation folds.

TABLE 6 | Models' performances [%], validation AUC on the DiCOVA Track-1 dataset, and test UAR on the ComParE dataset, with single learning (train from scratch), and the proposed transfer learning strategies.

		Dataset	Baseline	CNN-4	ResNet-6	VGG-7	MobileNet-6
Single Learning	–	ComParE	64.70	63.35	61.78	57.38	63.80
		DiCOVA	68.81	68.76	62.53	64.88	64.27
Transfer Learning	Parameters	ComParE	–	61.24	60.01	**66.43**	57.22
		DiCOVA	–	**69.88**	66.39	**70.10**	63.29
	Embeddings	ComParE	–	**64.82**	60.67	58.49	63.37
		DiCOVA	–	**72.38**	**71.38**	**72.34**	66.47

Pre-trained COUGHVID models and their corresponding transfer learning settings are chosen based on the best performance in ***Table 5****. "Embeddings" here include addition/concatenation. The numbers in bold are higher than the baseline.*

FC layer) might encounter an overfitting issue (23). Specifically, earlier layers/blocks generate low-level, generic features, which do not change significantly during the training procedure (23). Conversely, the convolutional layer/block 3 herein generates more high-level, domain-dependent representations. As for the embeddings incorporation, concatenation and addition of the embeddings achieve similar results, which indicates that both operations equally transfer the knowledge learnt from the FluSense dataset. Furthermore, we find that incorporating the embeddings after the convolutional layer/block 3 mostly outperforms the operations on other layers/blocks. This can be caused by more discrimination power obtained by applying the pre-trained FluSense models.

From **Table 6**, we further validate the generalisation ability of the proposed CovNet with the DiCOVA Track-1 and COMPARE CCS datasets. By competing with the official baselines, even simple CNNs can also achieve better performance with the proposed transfer learning methods. Therefore, the considered CovNet appears robust and can provide useful knowledge when detecting COVID-19 from crowd-sourced cough recordings. However, the performance improvement over the COMPARE CCS baseline by incorporating the embeddings method is not obvious, which might be caused by the inherent data difference between the FluSense and COUGHVID datasets and the COMPARE CCS dataset. Moreover, the CovNet works very well on the DiCOVA track-1 dataset, especially the incorporating embeddings. Perhaps, the embeddings from the pre-trained COUGHVID models carry more beneficial knowledge compared with parameters of convolutional layers on the DiCOVA dataset.

The main purpose of this study is to introduce and prove the usefulness of the transfer learning framework CovNet, instead of competing with the state-of-the-art performance on the DiCOVA Track-1 dataset (54–56) and COMPARE CCS dataset (19). The constructed four CNN models are so simple that each of them only contains three convolutional layers/blocks; we do not apply any data augmentation techniques and the only input to the networks are the original log Mel spectrograms.

5. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a transfer learning framework, CovNet, containing transferring parameters and incorporating embeddings. Transferring parameters indicate fine-tuning the models by initialising and freezing some parameters with the pre-trained model; incorporating embeddings describe concatenating or adding the embeddings generated by a pre-trained model with the embeddings produced by the current model.

The effectiveness and generalisation ability of the proposed transfer learning framework was demonstrated when developing simple CNNs for COVID-19 detection from crowd-sourced cough sounds. In the future, one should consider deeper neural networks to further improve performance through transfer learning. Moreover, other knowledge transfer architectures, such as multi-task learning (57) and domain adaption (58) can be explored.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YC contributed to the study design, experimenting, manuscript drafting, and editing. XJ contributed to the experimenting and manuscript editing. ZR contributed to the study design, manuscript drafting, and editing. BS supervised the whole process, from study design, overall implementation, to manuscript drafting, and editing. All authors approved the submitted version.

FUNDING

This study was partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS), the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), and the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor (Grant No. 01DD20003).

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the holders of the FluSense, COUGHVID, ComParE 2021 CCS, and DiCOVA challenge 2021 Track-1 datasets for providing collected data for research purposes. Moreover, thanks to all the participants for their data points.

REFERENCES

1. Atalan A. Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. *Ann Med Surg*. (2020) 56:38–42. doi: 10.1016/j.amsu.2020.06.010
2. Inoue K, Hashioka S, Kawano N. Risk of an increase in suicide rates associated with economic downturn due to COVID-19 pandemic. *Asia Pac J Public Health*. (2020) 32:367. doi: 10.1177/1010539520940893
3. Schuller BW, Schuller DM, Qian K, Liu J, Zheng H, Li X. COVID-19 and computer audition: an overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis. *Front Digit Health*. (2021) 3:14. doi: 10.3389/fdgth.2021.564906
4. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, et al. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. *J Med Virol*. (2020) 92:903–8. doi: 10.1002/jmv.25786
5. Tang YW, Schmitz JE, Persing DH, Stratton CW. Laboratory diagnosis of COVID-19: Current issues and challenges. *J Clin Microbiol*. (2020) 58:e00512-20. doi: 10.1128/JCM.00512-20
6. Dinnes J, Deeks JJ, Berhane S, Taylor M, Adriano A, Davenport C, et al. Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database Syst Rev*. (2021) 3:1–15. doi: 10.1002/14651858.CD013705.pub2
7. Santosh KC. Chapter 1: Speech processing in healthcare: can we integrate? In: Dey N, editor. *Intelligent Speech Signal Processing*. New York, NY: Academic Press (2019). p. 1–4. doi: 10.1016/B978-0-12-818130-0.00001-5
8. Mukherjee H, Sreerama P, Dhar A, Obaidullah SM, Roy K, Santosh KC, et al. Automatic lung health screening using respiratory sounds. *J Med Syst*. (2021) 45:19. doi: 10.1007/s10916-020-01681-9
9. Mukherjee H, Salam H, Santosh K. Lung health analysis: adventitious respiratory sound classification using filterbank energies. *Int J Pattern Recogn Artif Intell*. (2021) 2021:2157008. doi: 10.1142/S0218001421570081
10. Larsen JR, Martin MR, Martin JD, Kuhn P, Hicks JB. Modeling the onset of symptoms of COVID-19. *Front Public Health*. (2020) 8:473. doi: 10.3389/fpubh.2020.00473
11. Buoite Stella A, Manganotti P, Furlanis G, Accardo A, Ajčević M. Return to school in the COVID-19 era: considerations for temperature measurement. *J Med Eng Technol*. (2020) 44:468–71. doi: 10.1080/03091902.2020.1822941
12. Wei W, Wang J, Ma J, Cheng N, Xiao J. A real-time robot-based auxiliary system for risk evaluation of COVID-19 infection. In: *Proc. Interspeech*. Shanghai (2020). p. 701–5. doi: 10.21437/Interspeech.2020-2105
13. Alqudaihi KS, Aslam N, Khan IU, Almuhaideb AM, Alsunaidi SJ, Ibrahim NM, et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. *IEEE Access*. (2021). doi: 10.1109/ACCESS.2021.3097559
14. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform Med Unlocked*. (2020) 20:100378. doi: 10.1016/j.imu.2020.100378
15. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: *Proc. ACM SIGKDD*. New York, NY: ACM (2020). p. 3474–84. doi: 10.1145/3394486.3412865
16. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, R N, et al. Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: *Proc. Interspeech*. Shanghai: Interspeech (2020). p. 4811–5. doi: 10.21437/Interspeech.2020-2768
17. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci Data*. (2021) 8:1–10. doi: 10.1038/s41597-021-00937-4
18. Al Hossain F, Lover AA, Corey GA, Reich NG, Rahman T. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. New York, NY: ACM (2020). p. 1–28. doi: 10.1145/3381014
19. Schuller B, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, et al. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. In: *Proc. Interspeech*. Brno (2021). p. 431–5. doi: 10.21437/Interspeech.2021-19
20. Muguli A, Pinto L, R N, Sharma N, Krishnan P, Ghosh PK, et al. DiCOVA challenge: dataset, task, and baseline system for COVID-19 diagnosis using acoustics. In: *Proc. Interspeech*. Brno (2021). p. 901–5. doi: 10.21437/Interspeech.2021-74
21. Torrey L, Shavlik J. Transfer learning. In: Ganchev T, Sokolova M, Rada R, Garcia-Laencina PJ, Ravi V, editors. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods,*

and Techniques. Hershey, PA: IGI Publishing (2010) p. 242–64. doi: 10.4018/978-1-60566-766-9.ch011

22. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowledge Data Eng*. (2010) 22:1345–59. doi: 10.1109/TKDE.2009.191
23. Mehdipour Ghazi M, Yanikoglu B, Aptoula E. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*. (2017) 235:228–35. doi: 10.1016/j.neucom.2017.01.018
24. Cao B, Pan SJ, Zhang Y, Yeung DY, Yang Q. Adaptive transfer learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Atlanta, AL: AAAI (2010). p. 407–12.
25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. (2017) 60:84–90. doi: 10.1145/3065386
26. Kornblith S, Shlens J, Le QV. Do better imagenet models transfer better? In: *Proc. CVPR*. Long Beach, CA (2019). doi: 10.1109/CVPR.2019.00277
27. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput Biol Med*. (2021) 128:104115. doi: 10.1016/j.compbiomed.2020.104115
28. Raghu M, Zhang C, Kleinberg J, Bengio S. *Transfusion: Understanding Transfer Learning for Medical Imaging*. Red Hook, NY: Curran Associates Inc. (2019).
29. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162
30. Pons J, Serra J, Serra X. Training neural audio classifiers with few data. In: *Proc. ICASSP*. Brighton (2019). p. 16–20. doi: 10.1109/ICASSP.2019.8682591
31. Kumar A, Khadkevich M, Fagen C. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In: *Proc. ICASSP*. Calgary, AB (2018). p. 326–30. doi: 10.1109/ICASSP.2018.8462200
32. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans Audio Speech Lang Process*. (2020) 28:2880–94. doi: 10.1109/TASLP.2020.3030497
33. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: robust DNN embeddings for speaker recognition. In: *Proc. ICASSP*. Calgary, AB (2018). p. 5329–33. doi: 10.1109/ICASSP.2018.8461375
34. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision-ECCV 2014*. Cham: Springer International Publishing (2014). p. 818–33. doi: 10.1007/978-3-319-10590-1_53
35. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W. CNN-RNN: a unified framework for multi-label image classification. In: *Proc. CVPR*. Las Vegas, NV (2016). doi: 10.1109/CVPR.2016.251
36. Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, et al. HCP: a flexible CNN framework for multi-label image classification. *IEEE Trans Pattern Anal Mach Intell*. (2016) 38:1901–7. doi: 10.1109/TPAMI.2015.2491929
37. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. In: *Proc. ICARCV*. Marina Bay Sands (2014). p. 844–8. doi: 10.1109/ICARCV.2014.7064414
38. Ren Z, Baird A, Han J, Zhang Z, Schuller B. Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. In: *Proc. ICASSP*. Barcelona (2020). p. 7184–88. doi: 10.1109/ICASSP40776.2020.9054087
39. Kong Q, Yu C, Xu Y, Iqbal T, Wang W, Plumbley MD. Weakly labelled AudioSet tagging with attention neural networks. *IEEE/ACM Trans Audio Speech Lang Process*. (2019) 27:1791–802. doi: 10.1109/TASLP.2019.2930913
40. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *Proc. ICLR*. San Diego, CA (2015).
41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proc. CVPR*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
42. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications (2017). *arXiv [Preprint]. arXiv*: 1704.04861. Available Online at: https://dblp.uni-trier.de/rec/journals/corr/HowardZCKWWAA17.html?view=bibtex
43. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proc. ICML*. Lille: ICML (2015). p. 448–56.
44. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: *Proc. ICML*. Madison, WI (2010). p. 807–14.
45. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. (2014) 15:1929–58. Available online at: https://jmlr.org/papers/v15/srivastava14a.html
46. Sitaula C, Belayet Hossain M. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. *Appl Intell*. (2021) 51:2850–63. doi: 10.1007/s10489-020-02055-x
47. Lella KK, Pja A. Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath. *Alexandria Eng J*. (2021) 61:1319–34. doi: 10.1016/j.aej.2021.06.024
48. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In: *Proc. ICASSP*. New Orleans, LA (2017). p. 131–5. doi: 10.1109/ICASSP.2017.7952132
49. Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio Set: An ontology and human-labeled dataset for audio events. In: *Proc. ICASSP*. New Orleans, LA (2017). p. 776–80. doi: 10.1109/ICASSP.2017.7952261
50. Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller B. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. *BMJ Innovations*. (2021) 7:356–62. doi: 10.1136/bmjinnov-2021-000668
51. Nayak SR, Nayak DR, Sinha U, Arora V, Pachori RB. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study. *Biomed Signal Process Control*. (2021) 64:102365. doi: 10.1016/j.bspc.2020.102365
52. Schuller B, Batliner A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. 1st ed. New York, NY: Wiley Publishing (2013). doi: 10.1002/9781118706664
53. Rosenberg A. Classifying skewed data: importance weighting to optimize average recall. In: *Proc. Interspeech*. Portland, OR (2012). p. 2242–5. doi: 10.21437/Interspeech.2012-131
54. Sodergren I, Nodeh MP, Chhipa PC, Nikolaidou K, Kovacs G. Detecting COVID-19 from audio recording of coughs using random forests and support vector machines. In: *Proc. Interspeech*. Brno: Interspeech (2021) p. 916–20. doi: 10.21437/Interspeech.2021-2191
55. Das RK, Madhavi M, Li H. Diagnosis of COVID-19 using auditory acoustic cues. In: *Proc. Interspeech*. Brno: Interspeech (2021) p. 921–5. doi: 10.21437/Interspeech.2021-497
56. Harvill J, Wani YR, Hasegawa-Johnson M, Ahuja N, Beiser D, Chestek D. Classification of COVID-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation. In: *Proc. Interspeech*. (2021) p. 926–30. doi: 10.21437/Interspeech.2021-799
57. Caruana R. Multitask learning. *Mach Learn*. (1997) 28:41–75. doi: 10.1023/A:1007379606734
58. Wang M, Deng W. Deep visual domain adaptation: a survey. *Neurocomputing*. (2018) 312:135–53. doi: 10.1016/j.neucom.2018.05.083

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

REVIEW
published: 29 March 2022
doi: 10.3389/fpsyg.2022.869475

A Review of Research on the Neurocognition for Timbre Perception

Yuyan Wei [1], *Lin Gan* [2]* *and Xiangdong Huang* [1]

[1] *Department of Electrical and Information Engineering, Tianjin University, Tianjin, China,* [2] *Department of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin, China*

Edited by:
Shengchen Li,
Xi'an Jiaotong-Liverpool University,
China

Reviewed by:
Zhongzhe Xiao,
Soochow University, China
Haifeng Li,
Harbin Institute of Technology, China

****Correspondence:***
Lin Gan
ganlin@tju.edu.cn

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Psychology

Received: *04 February 2022*
Accepted: *07 March 2022*
Published: *29 March 2022*

Citation:
Wei Y, Gan L and Huang X (2022) A Review of Research on the Neurocognition for Timbre Perception. Front. Psychol. 13:869475. doi: 10.3389/fpsyg.2022.869475

As one of the basic elements in acoustic events, timbre influences the brain collectively with other factors such as pitch and loudness. Research on timbre perception involve interdisciplinary fields, including physical acoustics, auditory psychology, neurocognitive science and music theory, etc. From the perspectives of psychology and physiology, this article summarizes the features and functions of timbre perception as well as their correlation, among which the multi-dimensional scaling modeling methods to define timbre are the focus; the neurocognition and perception of timbre (including sensitivity, adaptability, memory capability, etc.) are outlined; related experiment findings (by using EEG/ERP, fMRI, etc.) on the deeper level of timbre perception in terms of neural cognition are summarized. In the meantime, potential problems in the process of experiments on timbre perception and future possibilities are also discussed. Thought sorting out the existing research contents, methods and findings of timbre perception, this article aims to provide heuristic guidance for researchers in related fields of timbre perception psychology, physiology and neural mechanism. It is believed that the study of timbre perception will be essential in various fields in the future, including neuroaesthetics, psychological intervention, artistic creation, rehabilitation, etc.

Keywords: timbre perception, neurocognitive, psychology, EEG, ERP, fMRI

1. INTRODUCTION

Timbre is a complex and abstract concept. Compared to other acoustic characteristics such as pitch and loudness, academic research on timbre started late and drew less attention, since timbre has been considered one of the most difficult acoustic features to comprehend. The Acoustical Society of America defined timbre in the 1960s as follows: the attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar. However, this definition only describes timbre from the dimension of loudness and pitch, rather than from the nature of timbre itself. In fact, timbre is not a single property, since it arises from an event produced by a single or several sound sources that are perceptually fused or blended into a single auditory image (Siedenburg et al., 2019). It contains not only auditory superficial features, but also rich auditory cognitive characteristics. Therefore, to systematically study timbre, it is necessary to integrate timbre perception with neurocognition, which requires a high-level interdisciplinary combination of psychology, physiology, neurology, physics, etc. To some extent, the difficulty of interdisciplinarity also leads to the fact that most of the timbre-relevant research works still stay in the exploratory stage.

Meanwhile, with the increasing progress of brain science and cognitive science, the last decade has witnessed an upsurge of interest in timbre. As far as research methods are concerned, besides the traditional behavioral science and psychology, the neural mechanism research based on brain imaging technologies such as EEG and fMRI has been increasingly applied. Regarding the spatial response of the brain stimulated by the timbre, the exploration range has extended from the initial auditory cortex to the overall analysis of multiple brain regions. Moreover, in the exploration of neural representation of timbre, besides peripheral auditory system, neurons researches at the mesoscopic level have also made further breakthroughs. For example, by imitating over 1,000 neurons in the mammalian primary auditory cortex as well as from simulated cortical neurons, Patil et al. (2012) constructed a neuro-computational framework to explore timbre classification. Meanwhile, the timbre stimuli also become increasingly complex, which have developed from simple mode of auditory stimuli (such as monophonic and synthetic sounds) to more complex stimuli with cognitive sensations (such as melody and natural sounds). These studies impose great significance on both the neural mechanism of the brain reaction to timbre and the aesthetic perception of timbre.

This article aims to summarize the existing timbre-related research works in the field of neurocognition, which are sorted out into four parts. The first part addresses how researchers link the perceptual dimension of timbre to the quantitative dimension of acoustics from the perspective of psychophysics. The second part gives a comprehensive discussion on the brain's perception of timbre,which includes memory capability, adaptability etc. The third part focuses on the research of event-related potentials that are related to timbre. Finally, in the fourth part, the spatial distribution characteristics of brain perception on timbre are summarized. Besides, the overall diagram addressing the structural relationship of these four parts is illustrated in **Figure 1**.

2. DIMENSIONS OF TIMBRE

The earliest research of timbre can be traced back to the work of Helmholtz and Ellis (1855) and Stumpf (1926), and their research mostly studied timbre just from the perspective of physics which ignored the perceptual qualities of timbre. However, with the deepening of the timbre research, it has been found that timbre, as a complex perceptual property of a specific fused auditory event, is also involved in psychology and other disciplines. In the 1970s, a pioneering work was started by Plomp (1970) and Wessel (1973), who studied timbre perception from the perspective of the psychophysics. Following this, multiple dimensions were developed to study the psychological perception of timbre (Grey, 1977; McAdams, 1993; Handel, 1995; Hajda et al., 1997; Toiviainen et al., 1998). The following will describe in detail how these dimensions arise and what they refer to, from which we will also reveal some important but unsettled problems for discussion.

2.1. Timbre Space

Through the exploration of the internal properties of timbre perception, the concept of timbre space was established through the well-known Multidimensional Scaling (MDS) research method (Grey, 1977; McAdams, 1993; Handel, 1995; Hajda et al., 1997; Toiviainen et al., 1998). This method links the people perception (psychology) to the timbre's physical properties (physics) *via* inference from the rating data of pairwise timbres, which can implement categorization of timbres without relying on any prior processing of the physical or perceptual structure of the timbre. Firstly, pairwise coupling is conducted among the given timbre set (i.e., any timbre should be transversely paired with other timbres). Secondly, all listeners are asked to rate the differences of all timbre pairs subjectively. Thirdly, these rating results are compared with each other so that a geometry space named "the timbre space" (an example of a three-dimensional timbre space diagram was illustrated in **Figure 2**) can be generated, from which the Multidimensional Scaling (MDS) model can be built up. In such timbre space, those timbres with similar properties tend to be closer to each other, and *vice versa*.

The basic MDS modeling method is based on the underlying assumption that all timbres are equally treated and all listeners are of the same level of perceptual ability (Kruskal, 1964a,b; Plomp, 1976). In other words, this modeling neither imposes any weight on certain special timbre nor makes any distinction among listeners. To further improve the modeling performance, a series of variants of MDS modeling approaches emerged by means of moderately relaxing this assumption. Among them, the EXSCAL algorithm (Krumhansl, 1989; Winsberg and Carroll, 1989) incorporates the specificity of every timbre. For the INDSCAL algorithm (Carroll and Chang, 1970; Wessel, 1973; Miller and Carterette, 1975; Plomp, 1976; Grey, 1977) and CLASCAL algorithm (Winsberg and Soete, 1993; McAdams et al., 1995), the listeners need to be categorized into several groups in terms of their abilities (or specificities) of timbre perception, which are treated with different weights accordingly. The CONSCAL algorithm (Winsberg and Soete, 1997; Caclin et al., 2005) can yield accurate models customized for individual listeners through continuous mapping operations on the timbre positions along perceptual dimensions by using spline functions. Generally speaking, since the above modified MDS algorithms can provide more accurate multidimensional timbre space, these variants tend to perform better in describing features, structures, and qualities of different timbres compared to the basic MDS method.

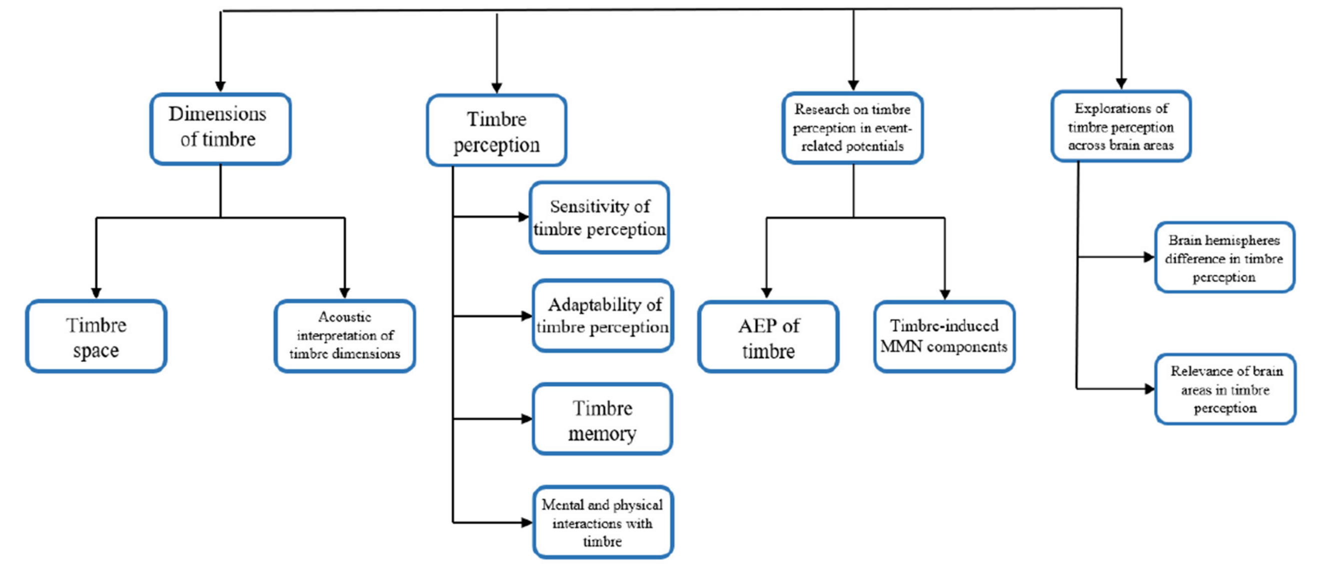

Dimensions of timbre
Timbre perception
Research on timbre perception in event-related potentials
Explorations of timbre perception across brain areas
Sensitivity of timbre perception
Adaptability of timbre perception
Timbre memory
Mental and physical interactions with timbre
Timbre space
Acoustic interpretation of timbre dimensions
AEP of timbre
Timbre-induced MMN components
Brain hemispheres difference in timbre perception
Relevance of brain areas in timbre perception

FIGURE 1 | Overall structural diagram of this review.

2.2. Acoustic Interpretation of Timbre Dimensions

The timbre space generated by the MDS modeling is about perception dissimilarity for sounds with similar pitch, duration, and loudness, and it represents the common perception dimension of timbre. A basic assumption is that these perceptual dimensions are orthogonal and should be represented by independent physical properties. These physical properties are used as an acoustic interpretation of timbre, which are called the audio descriptors.

These audio descriptors can be acquired by combining different perceptual dimensions and acoustic-related physical parameters, which can be categorized into descriptors of temporal, spectral, and spectrotemporal (Peeters et al., 2011). Temporal descriptions refer to the time aspect of sound. Some of them are directly extracted from waveforms, but most are usually extracted from time energy envelopes. Spectral descriptions usually refer to the local features of frequency contents. Spectrotemporal descriptions usually refer to the spectral changes across multiple time frames.

Generally speaking, most studies (Grey and Gordon, 1978; Iverson and Krumhansl, 1993; Krimphoff et al., 1994; McAdams et al., 1995; Kendall et al., 1999) agree that the following descriptors can represent the characteristics of different timbres: (1) Spectral centroid: It represents the relationship between low and high harmonics. Specifically, the greater the amplitudes of the high- frequency components relative to the low-frequency components are, the higher the spectral centroid is and thus the clearer and brighter the sound is. For example, the oboe has a higher spectral centroid than the French horn. (2) Attack time: It indicates a transition period, during which the amplitude of a particular harmonic increases from the perceptible threshold level to the maximum value. The shorter the attack time is, the more acute the timbre feels. For example, string instruments have a longer attack time than percussion instruments. (3) Spectral flux: The evolving degree of the spectral shape within a duration. (4) Spectral irregularity: It is relevant to the intensity of even harmonics relative to odd harmonics. If the amplitudes of even harmonics are relatively lower than odd harmonics, the sound tends feel hollow.

Audio descriptors play important roles in characterizing the psychoacoustics of timbre, which help explain the timbre perception in acoustic fields. However, the current research on timbre descriptors is still in confusion: how many descriptors can comprehensively describe a timbre? How does people perceive timbre?—with a linear or nonlinear combination of descriptors? How to evaluate the interpretability of an individual descriptor? These questions are still worth further exploration.

3. TIMBRE PERCEPTION

Limited by the multidimensional and complex characteristics of timbre, either the traditional research based on acoustic characteristics or the spectrum analysis and the psychological subjective evaluation meets challenges in exploring timbre perceptions. Since timbre perception is ultimately fulfilled by the brain, some studies attempted to combine physical cognition and neural perception of timbre for the purpose of uncovering the interactions between different timbre dimensions. For example, Caclin et al. (2006) showed that different dimensions of timbre are processed in parallel when the brain perceives timbre. In conclusion, it is an indispensable work to explore the mechanism on how the brain processes timbre perception.

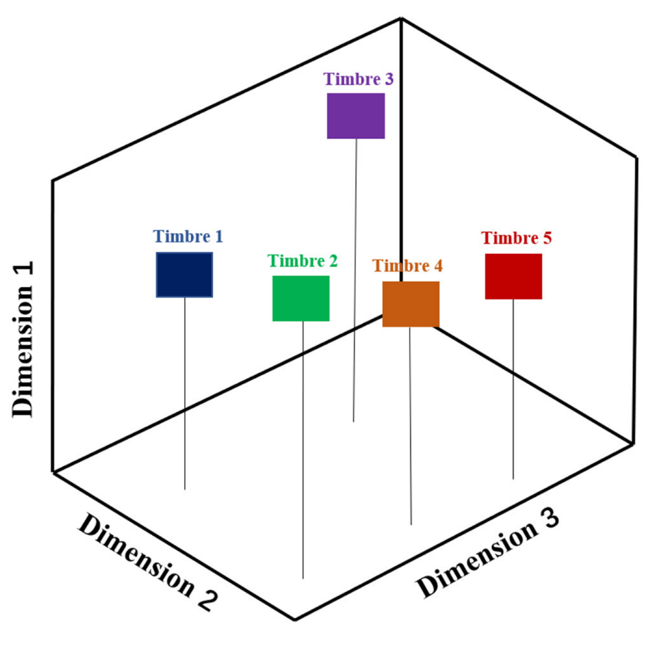


FIGURE 2 | The three-dimensional timbre space diagram of five different timbres. The geometric distance between two timbres corresponds to the perceived differences between them, and the spatial dimensions are correlated with acoustical physical properties.

3.1. Sensitivity of Timbre Perception

Studies have found that the human brain can perceive the difference of timbre. Peynirciou et al. (2016) attempted to acquire timbre perception difference through the artificial mixture of real musical instruments. They conducted two experiments. In one experiment, the subjects were asked to hear several fragments of the mixture of different instruments, from which they judged the mixing degree of these musical instruments. In the other experiment, the participants were asked to identify the different timbres that contained specific proportion of mixed instruments. These two experimental results showed that participants could accurately perceive the timbre differences of these instruments. Moreover, it was also found that the subjects who received music training and those who did not receive music training showed similar response patterns. Meanwhile, Samson et al. (1997) conducted experiments using synthesized timbres, which only subtly differed in frequency spectra and time information. They proved that the human brain was very sensitive to the perception of timbre differences, which was consistent to the conclusion drawn by Peynircioğlu.

3.2. Adaptability of Timbre Perception

The human brain has a certain degree of adaptability to the perception of timbre (i.e., the perceptual after-effect). To verify this effect, Piazza et al. (2018) asked the participants to be repeatedly exposed to two sounds (e.g., clarinet and oboe, male and female voice) and then these subjects were asked to hear one of them. The results showed that: when the subjects solely listened to sound A (or B), they naturally incorporated the auditory perception effects of A with those of B. Moreover, the experiment also proved that such after-effects were robust for moderate pitch changes. This adaptation contributes to the stability of timbre perception and the extensibility of familiar timbre. It actually enhances the sensitivity to novel or rare auditory objects, such as the timbre of an unfamiliar human voice.

3.3. Timbre Memory

It is often taken for granted that timbre can be easily memorized in the brain. However, the memory of timbre actually requires a complex mnemonic architecture, which delicately keeps track of sound identities and concurrently manages timbre operations (such as sensory processing, information storage, and matching of representations). Poulin-Charronnat et al. (2004) found that changing the instrumental timbre will affect the memory of tonal excerpts in human brain during the study of tonal and atonal music memory. Trainor et al. (2004) also found that the timbre change perturbs infants' melody memory. In the study on synthesized timbres, Golubock and Janata (2013) found that all the differences along the dimensions of spectral centroid, attack time, and spectral flux would influence the capacity of working memory. Meanwhile, Schellenberg and Habashi (2015) studied the memory capability influenced by the lags of timbre stimuli. Specifically, in the melody recognition test, by altering the lags between exposure and test spanning, which were set as 10 min, 1 day, and 1 week, they surprisingly discovered that the lag alternations did not significantly affect the timbre memory capability.

3.4. Mental and Physical Interactions With Timbre

In some circumstances, changing in timbre may awaken people's overall physiological and psychological responses, which is especially obvious in sensorimotor system. Many studies (Behrens and Green, 1993; Gabrielsson and Juslin, 1996; Leman et al., 2009) have found that musicians can use gesture language to convey emotional intentions under the influence of timbre. Therefore, people may not just passively listen to different timbres, which means, timbre changes in turn can also promote the extent of musicians' involvement in the overall state of the body (Halpern et al., 2004). Overy and Molnar-Szakacs (2009) proposed the "shared affective motion experience" (SAME) hypothesis based on the basic level of noisy and normal timbres. Combined with behavioral research, Blumstein et al. (2012) proved that noisy timbre could cause more vigorous physical activity than non-noisy timbre in terms of evoking the limbic nervous system response. Following this, Wallmark et al. (2018) supplemented the consensus of the predecessors through the embodied cognition research paradigm. They conducted experiments on different monophonic timbres and composite music timbres, which were then converted into noisy timbres by means of pitch shifting techniques. Their experimental results show that, such noisy timbres are able to arouse greater physical exertion and produce a lower emotional response than a non-noisy timbre, and that the noisy timbres can evoke responses in the motor system of the brain. It can be seen that, in addition to the significant impact on the human brain's auditory dimensions, timbre can also inspire a listener to produce emotional actions, which in turn reflects the perception of the listener.

Although efforts have been extensively made to study the brain responses of timbre, there are still many unsolved problems including the brain perception process of timbre identification, the exploration of brain regions for different functionalities of timbre perceptions etc. Therefore, the field of timbre perception study is still very young, which is expected to bring about breakthroughs through integrating varieties of neurocognitive experimental methods and new techniques of data processing such as EEG/ERP and fMRI.

4. RESEARCH ON TIMBRE PERCEPTION IN EVENT-RELATED POTENTIALS

As a popular physiological means to effectively reflect human brain activities, EEG (electroencephalogram) has been increasingly used in the research of brain cognitive mechanism. The EEG arises from the potential oscillations in the brain (i.e., excitatory postsynaptic potentials), and the current is afferent from the cortex of the thalamus to activate the parietal dendrites (Schaefer, 2017). Early EEG experiments mainly focused on the oscillation of the brain wave (which refers to spontaneous EEG), whose voltage can be collected in the experimental process. In contrast, later studies paid more attention to a special EEG component that had a time-locked relationship with psychological events, namely ERP (event-related potential). ERP has three notable features: one is that the waveform change is either positive or negative; the second is that the waveform change should behave as a sufficiently high intensity (amplitude); and the third is that the wave change occurs at a specific moment after the stimulus (latency period) are triggered (Wang, 2011). Because ERP can reflect mental activity in millisecond accuracy and thus has a high time resolution without causing any brain damage, it is applicable to explore the brain cognitive mechanism of short-term sound stimulus, such as the brain response to transient timbre stimulation.

4.1. AEP (Auditory Evoked Potential) of Timbre

Early studies have found that for auditory stimuli, N1 (Näätänen and Picton, 1987) and P2 (Celesia, 1976) are typical ERP components that reflect human auditory perception and auditory classification. Therefore, in numerous electrophysiological studies (Auzou et al., 1995; Liu et al., 2018; Hamlin et al., 2019; Banerjee et al., 2021) of timbre perception, researchers often treat N1 and P2 as indicators to reveal the neural mechanism of timbre processing.

Many studies (Helmholtz and Ellis, 1855; Fletcher, 1934; Seashore, 2008) have found that the timbre is closely related to the harmonic structure. Therefore, some researchers have carried out investigations on pure tone (lacking harmonic structures) and complex tone with the same baseband but different harmonics. Meyer et al. (2006) obtained ERP from 16 healthy subjects, who were required to distinguish between complex instrumental monophonic sounds (piano, trumpet, and violin) and simple pure sounds that lacked timbre characteristics. Analyses showed that, compared to pure tones, N1 and P2 responded more strongly to the tones of the instrument. At the same time, Tardón et al. (2021) attempted to discover the variations of the electrophysiological responses of the brain by simultaneously changing the acoustic characteristics of music, demonstrating that the amplitudes of the N1 and P2 components increased when the spectral flux, one of the dimensions of timbre, was mutated. Moreover, Pantev et al. (2001) observed that, for professional trumpet players and violinists, the timbre arising from playing their own instruments tended to evoke stronger N1 event-related potential components than other timbre did.

Besides the timbre stimuli, AEP can also be evoked *via* auditory imagination of timbre. Studies (Tuznik et al., 2018) have proved that the difference in imaginary timbre can be reflected in event-related potentials. Specifically, they found that timbre imagination is able to evoke other ERPs such as LPC in addition to N1 and P2. Moreover, LPC was found more sensitive to timbre changes of imagined sounds than other AEPs. In addition, it was also discovered that, once the ERP was successfully evoked, whether a subject had experienced music training or not, the magnitude of N1, P2, and LPC potentials were not affected. Nevertheless, when performing the same auditory imagination tasks related to timbre, the success rate of musicians was higher than that of non-musicians.

4.2. Timbre-Induced MMN Components

MMN (Mismatch Negativity) is an important component of event-related potentials (Luck, 2009), which is obtained by the Oddball paradigm. This paradigm involves two types of sound stimuli: standard stimuli and deviation stimuli. The standard stimuli appear with high probability, whereas deviation stimuli appear with low probability. To acquire MMN, ERPs evoked by both the standard stimuli and the deviation stimuli is need to be respectively superimposed and averaged. Then, the ERP evoked by the deviation stimuli is subtracted from the ERP evoked by the standard stimuli, from which a difference wave can be generated and treated as the desired MMN. The waveform of a MMN appears as a negative deflection, which occurs during 100–250 ms after the stimulus (i.e., the latency period is 100–250 ms; Näätänen et al., 2004). Particularly, the MMN can also be evoked even if the listener is in the coma state, which can be applied as an automatic indicator of the hearing mechanism recovery in the early treatment on the hearing-impaired.

Christmann et al. (2014) explored how timbre variation affected the MMN on the condition that other variables remained unchanged by means of the spectrally rotated technique. The experiment proved that, MMN evoked by instrumental sounds with timbre characteristics occurred earlier than those evoked by pure tones without timbre characteristics. This result indicates that the brain tends to be more sensitive to tones with rich harmonic structures, but is not influenced by pitch changes. Caclin et al. (2008) measured the variations of MMN by altering single-dimensional timbre characteristics (attack time, spectral centroid, even harmonic attenuation) and their combinations, from which he concluded that there existed some neural cells dedicatedly processing acoustics in the brain. Torppa et al.

(2018) found that when the children with cochlear implants (CIs) perceived timbre differences (from piano to Cymbal) in noise, the amplitude of MMN would change significantly. This conclusion suggests the importance of MMN in studying the timbre perception of children with CIs in noisy circumstances.

Since MMN is sensitive to varieties of timbre characteristics, it is usually employed as an indicator for timbre classification. Specifically, by altering the types of standard timbre stimuli or deviation timbre stimuli, MMN of different features (such as amplitudes, latency periods) can be evoked, thus reflecting listeners' abilities in identifying varieties of timbres, such as distinguishing between pure tones and overtones (Tervaniemi et al., 1997), distinguishing timbres with different spectral complexities (Tervaniemi et al., 2000), and distinguishing musical timbres that convey different emotions (Goydke et al., 2004). In addition, concerning applications of MMN, it was found that people with cochlear implants still had the ability to distinguish timbres although they were weaker than normal people (Koelsch et al., 2004).

In general, various ERP components induced by timbre actually provide a window to observe brain responses to timbre changes, which also helps researchers explore the neural mechanism of the brain's timbre perception. Nevertheless, to further discover this mechanism, it is necessary to analyze the timbre response distinctions across different brain areas.

5. EXPLORATIONS OF TIMBRE PERCEPTION ACROSS BRAIN AREAS

With the development of magnetic resonance imaging technology, fMRI (functional Magnetic Resonance Image) and PET (Positron Emission Computed Tomography) have gradually been adopted, through which the response mechanism of the brain to timbre perception in different spatial locations is being discovered. Related studies cover the spatial characteristics of brain's perception on spectral and temporal information of sounds (Zatorre and Belin, 2001; Hall et al., 2002), the effect on the spatial location distribution of brain from variations of timbre harmonics (Menon et al., 2002), the response difference of brain locations to the sound spectrum envelope (Warren et al., 2005), etc. These studies related to spatial features will facilitate the extension of explorable brain regions for timbre perception, such as the areas spreading from the auditory cortex to the whole brain.

5.1. Brain Hemispheres Difference in Timbre Perception

Some studies based on auditory perception have found that the left brain has an advantage in the processing of sound properties in the time domain (Robin et al., 1990), while the right brain has an advantage in dealing frequency-domain information (Zatorre and Belin, 2001; Menon et al., 2002). And based on the disclosure that timbre is closely related to both harmonic structure and time structure, researchers have found that timbre discrimination relies on the whole auditory cortex in the brain, while at the same time timbre perception also has the right-side advantage (Platel et al., 1997).

Studies have proved that the temporal lobe is involved in the brain's processing of timbre. On this basis, Samson and Zatorre (1994) carried out the timbre discrimination study on patients with unilateral temporal lobe resection, and found that only those subjects with right temporal lobe resection were affected in timbre discrimination. These findings supported the functional role of the right temporal lobe in timbre discrimination, and the responses of those subjects with left temporal lobe resection were not obvious. Leaver and Rauschecker (2010) also proved that right superior temporal regions were active in the processing of different timbres on instruments. However, their conclusion that the basic attributes of music perception were mainly biased toward the right hemisphere was then challenged by Johnsrude et al. (1997). Specifically, their study examined the processing of attack time in non-percussion sounds by using PET technology, which showed that the subjects had obvious activation foci in the left orbitofrontal cortex and left fusiform gyrus. Later, Samson et al. (2002) also showed that both the left and right hemispheres were involved in timbre processing. They found that the patients with left temporal lobe lesions were not influenced in distinguishing single sounds, but when the single sounds appeared in the background of a melody, the patients were unable to judge the degree of dissimilarity. At the same time, Menon et al. (2002) also revealed that the left brain and right brain exhibited some asymmetry in reaction to timbre stimuli. Such asymmetry expressed in the fact that the activation of the left temporal lobe was significantly posterior compared with the right hemisphere.

5.2. Relevance of Brain Areas in Timbre Perception

At present, studies have shown that the response of cerebral cortex to sound stimulus is mainly distributed in areas of the primary auditory cortex, superior temporal gyrus, superior temporal sulcus and Heschl's gyrus, prefrontal ventrolateral area, etc. (Platel et al., 1997; Caclin et al., 2007; Samson et al., 2011; Wallmark et al., 2018). Beside the above areas, some researchers believe that timbre, as a complex multi-dimensional perceptual property, may be related to the activities of brain areas that do not correspond to the processing of some low-level auditory stimuli. Thus, Meyer et al. (2006) used the EEG imaging method of low-resolution electromagnetic tomography (LORETA) and proved that timbre perception involved not only the two sides of the auditory cortex, but also the middle region of the brain that was related to emotion and auditory imagination. The research of Alluri et al. (2012) proved that when participants listened to timbres of "bright" qualities, the putamen (basal ganglia) would be activated. Wallmark et al. (2018) explored the neural dynamics of single-tone timbres at different noise levels to determine which areas of the brain were involved in the processing of noisy brain stimuli. It turned out that timbre processing was related to the sensorimotor area. At the same time, Blumstein et al. (2012) confirmed that the motion and edge responses caused by different timbres had certain differences.

The brainstem, an important part of the central auditory system, has also been studied to explore timbre perception. Strait et al. (2012) revealed that, a musician's auditory brainstem behaved as unique responses to his own frequently-exercising instrument, whereas it also showed insensitivity to other instruments with distinct timbres. By reviewing the work on the auditory brainstem's ability to respond to complex sounds, Anderson and Kraus (2010) found that timbre can be applied as an objective neural index for hearing-in-noise abilities.

6. DISCUSSION

On basis of the above retrospect, the findings of timbre perception can be obtained by either psychological approaches or physiological approaches. The psychological approaches typically refer to the basic multi-dimensional scaling modeling method and its variants, from which a series of audio descriptors concerning the related dimensions can be derived. The physiological approaches usually rely on signal acquisition means that are related to cerebral neural activities including EEG/ERP, fMRI, and PET, which provide various perspectives to explore the neural mechanism of the brain's timbre perception.

In general, it can be concluded that timbre perception is promising in psychology- and neurocognition-related fields. Specifically, timbre perception can play crucial roles in future applications such as music creation, auditory neuroaesthetics, and human-computer interaction experiences, if efforts are made in the following directions.

(1) Interdisciplinary fusion should be strengthened. Up to now, in both the timbre space modeling and timbre abstract encoding from low to high levels in the brain, the inter-discipline permeation is still not sufficient. For example, the aforementioned mental and physical interaction with timbre stimuli can hardly be interpreted well due to the lack of inter-discipline permeation. Therefore, it is urgent to integrate multiple academic fields including psychology, neurocognition, physical acoustics etc. Only in this way can the mechanism of the timbre-relevant perception be explored deeply.

(2) More attention should be paid to the relevance with other acoustic characteristics. Most of the existing timbre research only focuses on the characteristics of timbre itself. In fact, because timbre rarely appears in a single form, timbre perception inevitably is affected by other characteristics such as pitch, loudness, melody, and rhythm. Therefore, it is necessary to incorporate the relevance between timbre and other characteristics. Essentially,if the timbre research is placed in a comprehensive environment which organically links all these elements, the study of timbre perception will be more productive.

(3) More emphasis should be placed on individual differences. At present, only few studies address individual differences in timbre perception, and most of them only focus on the difference between musicians and non-musicians. Nevertheless, higher diversity should be considered concerning individual difference. For example, individuals with hearing impairments or with poor auditory perception should also be considered. The efforts on exploring the characteristics of special individuals' timbre perception will further promote the advancement of auditory aesthetics and neuromedicine.

AUTHOR CONTRIBUTIONS

YW: data collection, data analyses, and writing the article. LG: study idea, study design, and manuscript revision. XH: study design and manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a grant from National Natural Science Foundation of China (No. 6210-7029).

ACKNOWLEDGMENTS

We thank all members of our group for their support.

REFERENCES

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59, 3677–3689. doi: 10.1016/j.neuroimage.2011.11.019

Anderson, S., and Kraus, N. (2010). Sensory-cognitive interaction in the neural encoding of speech in noise: a review. *J. Am. Acad. Audiol.* 21, 575. doi: 10.3766/jaaa.21.9.3

Auzou, P., Eustache, F., Etevenon, P., Platel, H., Rioux, P., Lambert, J., et al. (1995). Topographic EEG activations during timbre and pitch discrimination tasks using musical sounds. *Neuropsychologia* 33, 25–37. doi: 10.1016/0028-3932(94)00100-4

Banerjee, A., Sanyal, S., Roy, S., Nag, S., Sengupta, R., and Ghosh, D. (2021). A novel study on perception-cognition scenario in music using deterministic and non-deterministic approach. *Phys. A Stat. Mech. Appl*. 567, 125682. doi: 10.1016/j.physa.2020.125682

Behrens, G. A., and Green, S. B. (1993). The ability to identify emotional content of solo improvisations performed vocally and on three different instruments. *Psychol. Mus*. 21, 20–33. doi: 10.1177/030573569302100102

Blumstein, D. T., Bryant, G. A., and Kaye, P. D. (2012). The sound of arousal in music is context-dependent. *Biol. Lett*. 8, 744–747. doi: 10.1098/rsbl.2012.0374

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., et al. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *J. Cogn. Neurosci*. 18, 1959–1972. doi: 10.1162/jocn.2006.18.12.1959

Caclin, A., Giard, M.-H., Smith, B. K., and McAdams, S. (2007). Interactive processing of timbre dimensions: a garner interference study. *Brain Res*. 1138, 159–170. doi: 10.1016/j.brainres.2006.12.065

Caclin, A., McAdams, S., Smith, B. K., and Giard, M.-H. (2008). Interactive processing of timbre dimensions: an exploration with event-related potentials. *J. Cogn. Neurosci*. 20, 49–64. doi: 10.1162/jocn.2008.20001

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. *J. Acoust. Soc. Am*. 118, 471–482. doi: 10.1121/1.1929229

Carroll, J. D., and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young"? decomposition. *Psychometrika* 35, 283–319. doi: 10.1007/BF02310791

Celesia, G. G. (1976). Organization of auditory cortical areas in man. *Brain* 99, 403–414. doi: 10.1093/brain/99.3.403

Christmann, C. A., Lachmann, T., and Berti, S. (2014). Earlier timbre processing of instrumental tones compared to equally complex spectrally rotated sounds as revealed by the mismatch negativity. *Neurosci. Lett*. 581, 115–119. doi: 10.1016/j.neulet.2014.08.035

Fletcher, H. (1934). Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *J. Acoust. Soc. Am*. 6, 59–69. doi: 10.1121/1.1915704

Gabrielsson, A., and Juslin, P. N. (1996). Emotional expression in music performance: between the performer's intention and the listener's experience. *Psychol. Mus*. 24, 68–91. doi: 10.1177/0305735696241007

Golubock, J. L., and Janata, P. (2013). Keeping timbre in mind: working memory for complex sounds that can't be verbalized. *J. Exp. Psychol. Hum. Percept. Perform*. 39, 399–412. doi: 10.1037/a0029720

Goydke, K. N., Altenmüller, E., Möller, J., and Münte, T. (2004). Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. *Brain Res. Cogn. Brain Res*. 21, 351–359. doi: 10.1016/j.cogbrainres.2004.06.009

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am*. 61, 1270–1277. doi: 10.1121/1.381428

Grey, J. M., and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am*. 63, 1493–1500. doi: 10.1121/1.381843

Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (1997). Methodological Issues in Timbre Research. Erlbaum: Psychology Press; Taylor and Francis, p. 253–306.

Hall, D. A., Johnsrude, I. S., Haggard, M. P., Palmer, A. R., Akeroyd, M. A., and Summerfield, A. Q. (2002). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 12, 140–149. doi: 10.1093/cercor/12.2.140

Halpern, A. R., Zatorre, R. J., Bouffard, M., and Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia* 42, 1281–1292. doi: 10.1016/j.neuropsychologia.2003.12.017

Hamlin, M. P., Mofle, T., Whitten, H., Roberts, K., Scheuber, S. H., and Scheuber, S. H. (2019). Emotional and neurological responses to timbre. University of Central Oklahoma.

Handel, S. (1995). Timbre perception and auditory object identification. Hearing 2, 425–461. doi: 10.1016/B978-012505626-7/50014-5

Helmholtz, H., and Ellis, A. (1855). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press.

Iverson, P., and Krumhansl, C. (1993). Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am*. 94, 2595–2603. doi: 10.1121/1.407371

Johnsrude, I. S., Zatorre, R. J., Milner, B., and Evans, A. C. (1997). Left?hemisphere specialization for the processing of acoustic transients. *NeuroReport* 8, 1761–1765. doi: 10.1097/00001756-199705060-00038

Kendall, R. A., Carterette, E. C., and Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Mus. Percept*. 16, 327–363. doi: 10.2307/40285796

Koelsch, S., Wittfoth, M., Wolf, A., Müller, J., and Hahne, A. (2004). Music perception in cochlear implant users: an event-related potential study. *Clin. Neurophysiol*. 115, 966–972. doi: 10.1016/j.clinph.2003.11.032

Krimphoff, J., Mcadams, S., and Winsberg, S. (1994). Caractérisation du timbre des sons complexes.ii. analyses acoustiques et quantification psychophysique. *J. Phys. IV France*. 4, C5-625–C5-628. doi: 10.1051/jp4:19945134

Krumhansl, C. L. (1989). "Why is musical timbre so hard to understand?," in *Structure and Perception of Electroacoustic Sound and Music, Proceedings of the Marcus Wallenberg Symposium 1998*. (Stockholm).

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1–27. doi: 10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115–129. doi: 10.1007/BF02289694

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci*. 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Leman, M., Desmet, F., Styns, F., van Noorden, L., and Moelants, D. (2009). Sharing musical expression through embodied listening: a case study based on Chinese Guqin music. *Mus. Percept*. 26, 263–278. doi: 10.1525/mp.2009.26.3.263

Liu, X., Xu, Y., Alter, K., and Tuomainen, J. (2018). Emotional connotations of musical instrument timbre in comparison with emotional speech prosody: evidence from acoustics and event-related potentials. *Front. Psychol*. 9, 737. doi: 10.3389/fpsyg.2018.00737

Luck, S. J. (2009). "Event-related potentials," in *APA Handbook of Research Methods in Psychology*, ed D. L. Long (Washington, DC: American Psychological Association), 35–79.

McAdams, S. (1993). Recognition of sound sources and events. *Thinking in sound: The cognitive psychology of human audition*. 146. p. 198. doi: 10.1093/acprof:oso/9780198522577.003.0006

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol. Res*. 58, 177–192. doi: 10.1007/BF00419633

Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., et al. (2002). Neural correlates of timbre change in harmonic sounds. *NeuroImage* 17, 1742–1754. doi: 10.1006/nimg.2002.1295

Meyer, M., Baumann, S., and Jäncke, L. (2006). Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *NeuroImage* 32, 1510–1523. doi: 10.1016/j.neuroimage.2006.04.193

Miller, J. R., and Carterette, E. C. (1975). Perceptual space for musical structures. *J. Acoust. Soc. Am*. 58, 711–720. doi: 10.1121/1.380719

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. *Clin. Neurophysiol*. 115, 140–144. doi: 10.1016/j.clinph.2003.04.001

Näätänen, R., and Picton, T. W. (1987). The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Overy, K., and Molnar-Szakacs, I. (2009). Being together in time: musical experience and the mirror neuron system. *Mus. Percept*. 26, 489–504. doi: 10.1525/mp.2009.26.5.489

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., and Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport* 12, 169–174. doi: 10.1097/00001756-200101220-00041

Patil, K., Pressnitzer, D., Shamma, S. A., and Elhilali, M. (2012). Music in our ears: the biological bases of musical timbre perception. *PLoS Comput. Biol*. 8:e1002759. doi: 10.1371/journal.pcbi.1002759

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: extracting audio descriptors from musical signals. *J. Acoust. Soc. Am*. 130, 2902–2916. doi: 10.1121/1.3642604

Peynirciou, Z. F., Brent, W., and Falco, D. E. (2016). Perception of blended timbres in music. *Psychol. Mus*. 44, 625–639. doi: 10.1177/0305735615578313

Piazza, E. A., Theunissen, F. E., Wessel, D., and Whitney, D. (2018). Rapid adaptation to the timbre of natural sounds. *Sci. Rep*. 8:13826. doi: 10.1038/s41598-018-32018-9

Platel, H., Price, C. J., Baron, J. C., Wise, R. J. S., Lambert, J., Frackowiak, R. S., et al. (1997). The structural components of music perception. A functional anatomical study. *Brain* 120(Pt 2), 229–243. doi: 10.1093/brain/120.2.229

Plomp, R. (1970). Timbre as a multi-dimensional attribute of complex tones. *Frequency Analysis and Periodicity Detection in Hearing*. 397–414.

Plomp, R. (1976). Aspects of tone sensation: a psychophysical study. Academic Press.

Poulin-Charronnat, B., Bigand, E., Lalitte, P., Madurell, F., Vieillard, S., and McAdams, S. (2004). Effects of a change in instrumentation on the recognition of musical materials. *Mus. Percept*. 22, 239–263. doi: 10.1525/mp.2004.22.2.239

Robin, D. A., Tranel, D., and Damasio, H. (1990). Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions. *Brain Lang*. 39, 539–555. doi: 10.1016/0093-934X(90)90161-9

Samson, F., Zeffiro, T. A., Toussaint, A., and Belin, P. (2011). Stimulus complexity and categorical effects in human auditory cortex: an activation likelihood estimation meta-analysis. *Front. Psychol*. 1, 241. doi: 10.3389/fpsyg.2010.00241

Samson, S., and Zatorre, R. J. (1994). Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologia* 32, 231–240. doi: 10.1016/0028-3932(94)90008-6

Samson, S., Zatorre, R. J., and Ramsay, J. O. (1997). Multidimensional scaling of synthetic musical timbre: perception of spectral and temporal characteristics. *Can. J. Exp. Psychol*. 51, 307–315. doi: 10.1037/1196-1961.51.4.307

Samson, S., Zatorre, R. J., and Ramsay, J. O. (2002). Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. *Brain* 125(Pt 3), 511–523. doi: 10.1093/brain/awf051

Schaefer, H.-E. (2017). Music-evoked emotions-current studies. *Front. Neurosci*. 11:600. doi: 10.3389/fnins.2017.00600

Schellenberg, E. G., and Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Mem. Cogn*. 43, 1021–1031. doi: 10.3758/s13421-015-0519-1

Seashore, C. E. (2008). Psychology of music. *Music Educ. J*. 23, 30–33.

Siedenburg, K., Saitis, C., and McAdams, S. (2019). "The present, past, and future of timbre research," in *Timbre: Acoustics, Perception, and Cognition*. (Springer International Publishing), 1–19. doi: 10.1007/978-3-030-14832-4_1

Strait, D. L., Chan, K., Ashley, R., and Kraus, N. (2012). Specialization among the specialized: auditory brainstem function is tuned in to timbre. *Cortex* 48, 360–362. doi: 10.1016/j.cortex.2011.03.015

Stumpf, C. (1926). *Die sprachlaute; experimentell-phonetische untersuchungen (nebst einem anhang über instrumentalklänge)*. Springer.

Tardón, L. J., Rodríguez-Rodríguez, I., Haumann, N. T., Brattico, E., and Barbancho, I. (2021). Music with concurrent saliences of musical features elicits stronger brain responses. *Appl. Sci*. 11:9158. doi: 10.3390/app11199158

Tervaniemi, M., Schröger, E., Saher, M., and Näätänen, R. (2000). Effects of spectral complexity and sound duration on automatic complex-sound pitch processing in humans-a mismatch negativity study. *Neurosci. Lett*. 290, 66–70. doi: 10.1016/S0304-3940(00)01290-8

Tervaniemi, M., Winkler, I., and Näätänen, R. (1997). Pre-attentive categorization of sounds by timbre as revealed by event-related potentials. *NeuroReport* 8, 2571–2574. doi: 10.1097/00001756-199707280-00030

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., and Näätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Mus. Percept*. 16, 223–241. doi: 10.2307/40285788

Torppa, R., Faulkner, A., Kujala, T., Huotilainen, M., and Lipsanen, J. (2018). Developmental links between speech perception in noise, singing, and cortical processing of music in children with cochlear implants. *Mus. Percept*. 36, 156–174. doi: 10.1525/mp.2018.36.2.156

Trainor, L. J., Wu, L., and Tsang, C. (2004). Long-term memory for music: infants remember tempo and timbre. *Dev. Sci*. 7, 289–296. doi: 10.1111/j.1467-7687.2004.00348.x

Tuznik, P., Augustynowicz, P., and Francuz, P. (2018). Electrophysiological correlates of timbre imagery and perception. *Int. J. Psychophysiol*. 129, 9–17. doi: 10.1016/j.ijpsycho.2018.05.004

Wallmark, Z., Iacoboni, M., Deblieck, C., and Kendall, R. A. (2018). Embodied listening and timbre: Perceptual, acoustical, and neural correlates. *Mus. Percept*. 35, 332–363. doi: 10.1525/mp.2018.35.3.332

Wang, L. (2011). *Research on the behavior and EEG of musical tone timbre perception* (Master's thesis). University of Electronic Science and Technology, Chengdu, China.

Warren, J. D., Jennings, A. R., and Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *NeuroImage* 24, 1052–1057. doi: 10.1016/j.neuroimage.2004.10.031

Wessel, D. L. (1973). "Psychoacoustics and music: A report from Michigan State University," in *PACE: Bulletin of the Computer Arts Society*. 30, 1–2.

Winsberg, S., and Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika* 54, 217–229. doi: 10.1007/BF02294516

Winsberg, S., and Soete, G. D. (1993). A latent class approach to fitting the weighted Euclidean model, clascal. *Psychometrika* 58, 315–330. doi: 10.1007/BF02294578

Winsberg, S., and Soete, G. D. (1997). Multidimensional scaling with constrained dimensions: conscal. *Br. J. Math. Stat. Psychol*. 50, 55–72. doi: 10.1111/j.2044-8317.1997.tb01102.x

Zatorre, R. J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 11, 946–953. doi: 10.1093/cercor/11.10.946

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/fphy.2022.863291

Neural Network Model Based on the Tensor Network for Audio Tagging of Domestic Activities

LiDong Yang[1], RenBo Yue[1], Jing Wang[2] and Min Liu[3]*

[1]School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou, China, [2]School of Information and Electronics, Beijing Institute of Technology, Beijing, China, [3]China Mobile Research Institute, Beijing, China

Due to the serious problem of population aging, monitoring of domestic activities is increasingly important. Audio tagging of domestic activities is very suitable when the visual data are unavailable due to the interference from light and the environment. Aiming at solving this problem, a neural network model based on the tensor network is proposed for audio tagging of domestic activities that is more interpretable than traditional neural networks. The introduction of the tensor network can compress the network parameters and reduce the redundancy of the training model while maintaining a good performance. First, the important features of a Mel spectrogram of the input audio are extracted through the convolutional neural networks (CNNs). Then, they are converted into the high-order space corresponding with the tensor network. The spatial structure information and important features can be further extracted and retained through the matrix product state (MPS). Large patches of the featured data are divided into small local orderless patches when using the tensor network. The final tagging results are obtained through the MPS layers which is just a tensor network structure based on the tensor train decomposition. In order to evaluate the proposed method, the DCASE 2018 challenge task 5 dataset for monitoring domestic activities is selected. The results showed that the average F1-score of the proposed model in the test set of the development dataset and validation dataset reached 87.7 and 85.9%, which are 3.2 and 2.8% higher than the baseline system, respectively. It is verified that the proposed model can perform better and more efficiently for audio tagging of domestic activities.

Keywords: tensor network, matrix product state (MPS), tensor train decomposition, audio tagging, neural network

Edited by:
Shengchen Li,
Xi'an Jiaotong-Liverpool University,
China

Reviewed by:
Qiuqiang Kong,
University of Surrey, United Kingdom
Xi Shao,
Nanjing University of Posts and
Telecommunications, China

****Correspondence:***
Jing Wang
wangjing@bit.edu.cn

Specialty section:
This article was submitted to
Interdisciplinary Physics,
a section of the journal
Frontiers in Physics

Received: *27 January 2022*
Accepted: *15 March 2022*
Published: *12 April 2022*

Citation:
Yang LD, Yue RB, Wang J and Liu M (2022) Neural Network Model Based on the Tensor Network for Audio Tagging of Domestic Activities. Front. Phys. 10:863291. doi: 10.3389/fphy.2022.863291

1 INTRODUCTION

The world is facing the problem of population aging. It is estimated that by 2050, the number of people over 64 years will exceed 20% of the world's population. According to the survey, 40% of the elderly will live alone at home [1]. This will lead to many social problems, such as the increase in diseases and healthcare costs, the shortage of nursing staff, and the increase in the number of people unable to live independently. Therefore, it is imperative to develop ambient intelligence-assisted living tools to help the elderly live independently at home [2]. The first task is to detect what is happening at home. Audio tagging is very suitable when the visual data are unavailable due to the interference from light and the environment. Audio tagging associate tags with the audio and identifies the events that generate the audio. Audio tagging of domestic activities has important applications in smart home robots, monitoring of domestic activities, and the lives of the elderly [3].

For the problem of audio tagging, Gong [4] proposed PSLA, a collection of model-agnostic training techniques. It includes ImageNet pre-training, balanced sampling, data augmentation, label augmentation, and model aggregation. The results we obtained outstripped the best previous systems. Puy [5] proposed a model based on separable convolutions, which uses separable convolutions in channel, time, and frequency dimensions to control the complexity of the network and achieved good results in terms of effect and complexity. The widely used dataset for audio signals is DCASE (Detection and Classification of Acoustic Scenes and Events). DCASE 2018 Challenge Task 5 [6] is specifically used for audio tagging for domestic activities. This tagging task provides the development and validation datasets and baseline system and requires identifying nine classes of events in domestic activities within 10-s clips. The audio data are collected by four linearly arranged microphones. There are many ways to process microphone array audio, among which Wang [7] proposed a modeling method that uses the channel mode, time mode, and frequency mode as the three dimensions to construct a three-dimensional tensor space, which has achieved good results. In the tensor completion method proposed by Yang [8], tensor modeling of multi-channel audio signals with the missing data has achieved good results.

Among the submitted systems in DCASE 2018 Challenge Task 5, the baseline system of this task trains a single classifier model that takes a single channel as the input. The learner in the baseline system is based on a neural network architecture using convolutional and dense layers. As input, log Mel-band energies are provided to the network for each microphone channel separately [9]. Inoue [10] put forward a combination method of a data-enhanced front-end module and a back-end module based on the CNN classification method. First, it enhances the input data by shuffling and mixing the sound clips. Its data enhancement method helped increase the variation of training samples and reduce the impact of unbalanced datasets. Then, the input of the CNN, as a classifier, is the log-Mel spectrogram of the enhanced data. The system proposed by Tanabe [11] is a combination of the front-end modules based on blind signal processing and the back-end modules based on machine learning. The front-end modules employ blind dereverberation and blind source separation. They use spatial cues without machine learning to avoid overfitting. The back-end modules employ one-dimensional convolutional neural network (1DCNN)-based architecture and VGG16-based architecture for the individual front-end modules. All of the probability outputs are ensembled. In addition, through mix-up-based data augmentation, overfitting is avoided in the back-end modules. TC2DCNN [12] is extended by operating the convolutions along the two dimensions of time and channel, not along the frequency axis, since similar patterns in different frequency bands do not necessarily belong to the similar audio event. INRC_2D [13] combines a deep neural network with a scattering transform. Each audio segment is first represented by two layers of scattering transform. The four denoised transforms of each of the two layers are combined together. Each of the fused layers is processed in parallel by two neural network (NN) architectures, RESNET, and a long short-term memory (LSTM) network, with a joint fully connected layer. The VGGish model proposed by Kong [14], which has an AlexNetish 8-layer CNN with global max pooling, has achieved good results.

The tensor network is a sparse data structure designed for the efficient representation and manipulation of the ultra-high dimensional data to achieve better interpretability of the data. It is similar to the kernel method in machine learning [15]. Through feature mapping, the original linearly inseparable data are converted to a high-dimensional space. In this space, a hyperplane can be linearly separable. But the number of parameters will be very large. Tensor train decomposition (also called the matrix product state) is a kind of tensor decomposition specifically for high-dimensional data. Wang [16] uses tensor train decomposition in a compressed HRTF, which is closer to the original HRTF than other methods. Therefore, tensor train decomposition is used to approximate the tensor networks. Matrix product state is the first tensor network to be discovered and used, which can be efficiently used in the simulation of the ground state of an infinite one-dimensional system. In recent years, tensor networks based on matrix product states have shown good performance in classification. For example, Stoudenmire [17] encoded the MNIST data into a tensor network, and the tensor network was trained to obtain the probability of each class to complete the classification. Efthymiou [18] proposed a new contraction method for Fashion-MNIST, which realizes the parallel compression of the horizontal edges, and then the vertical compression, which further accelerates the training speed. Selvan [19] proposed a lonet tensor network, which overcomes the shortcomings of the MPS tensor network, that is, the loss of spatial correlation when used for large resolutions. It is used for the two-dimensional classification of medical images and has achieved good results. While achieving good results, compared with other models, the GPU usage is significantly lower than that of the other models. PEPS [20] is a two-dimensional extension of the matrix product state. Although it has achieved great success, its algorithmic complexity is much higher than that of the matrix product state. MERA [20] is an experimental state of the ground state of a one-dimensional quantum system, which is inherently scale-invariant. In the MERA, tensors are connected to reproduce the holographic geometry. There are also other kinds of tensor network structures which have higher complexity than the MPS and can be used in other applications such as applied mathematics, chemistry, physics, machine learning, and many other fields.

In the article, a neural network model based on the tensor network is proposed for audio tagging of domestic activities. This article draws on the research results of the simplest and most mature matrix product state in the tensor network, hoping to achieve a balance between the complexity and effectiveness of the network model. An end-to-end tensor network-based neural network model is constructed and trained with the Mel spectrograms. After going through the convolutional layers, important features are extracted. Then, the MPS tensor network further extracts the features and gives the tagging

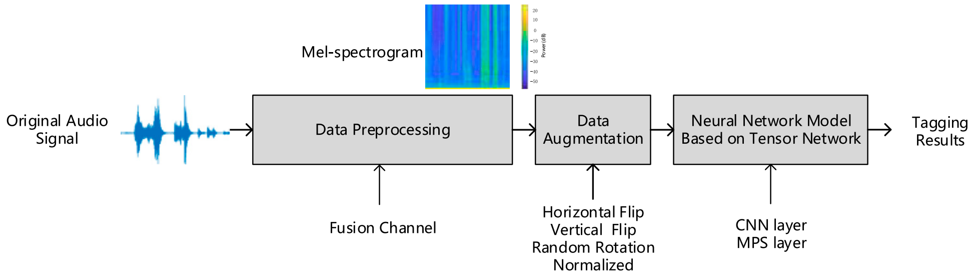


FIGURE 1 | Flow chart of audio tagging.

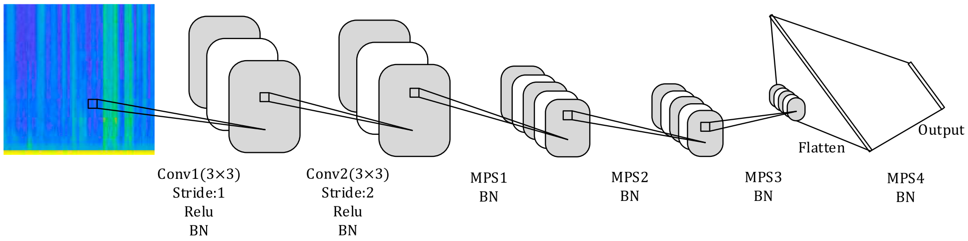


FIGURE 2 | Structure of the neural network model based on the tensor network.

results. This can not only achieve good tagging results but also compresses the network through tensor train decomposition, which has a smaller number of parameters than the traditional CNN. The F1-score is used to evaluate the performance of the proposed method. In terms of tagging performance, the performance of the proposed model is compared with other models. Compared with the results of the development dataset and the validation dataset of DCASE 2018 challenge task 5, the proposed method achieved better results. This article is a beneficial attempt to combine the tensor networks and neural networks and can also be extended to other deep learning sound signal processing fields.

The rest of this article is organized as follows: **Section 2** introduces the neural network model based on the tensor network proposed in this article in detail. **Section 3** introduces the parameter settings and experimental results of the proposed method, which are analyzed in terms of precision, recall, and F1-score, respectively. This article is concluded in **Section 4**.

2 NEURAL NETWORK MODEL BASED ON TENSOR NETWORK

As the experimental flowchart shows in **Figure 1**, the proposed audio tagging method consists of three main stages, namely, data preprocessing, data augmentation, and neural network model based on the tensor network. Data preprocessing first performs channel fusion [21] on the audio, then takes the log after FFT, and then the Mel spectrogram is obtained by mapping the Mel frequency.

The structure of the neural network model based on the tensor network is shown in **Figure 2**. Convolutional layers are used for extracting deeper feature representations. Important spatial structure and time information will be retained in the middle MPS layers. Finally, the retained information enters the MPS decision layer after being flattened to obtain the audio tagging results.

2.1 Data Preprocessing and Augmentation

The Mel spectrogram as the audio feature of the original signal is used in the proposed method. The Mel spectrogram converts the ordinary frequency scale of the spectrogram into the Mel frequency scale. After framing, the fbank feature is extracted through the Mel filter bank [22]. The energy value distribution range is summarized and is then linearly corresponded to blue-yellow [23]. In this article, 128 triangular filters are used to form a Mel filter bank, which corresponds to the objective law that the higher the frequency, the duller the human ear is.

Data augmentation uses horizontal flip, vertical flip, and random rotation to enlarge the training data, avoid overfitting, and enhance the robustness of the model.

2.2 Neural Network Model Based on the Tensor Network

2.2.1 CNN Feature Extraction

CNN [24] is used to process the multi-dimensional data, such as the two-dimensional images with many channels. CNN uses shared weights, local connections, pooling, and other layers to organize the attributes of natural signals. The convolutional layer, ReLU layer, and pooling layer are the most commonly used CNN layers.

The basic purpose of the convolutional layer is to determine the local connections between the features and map their information to a specific feature map. The convolution of the input I with filter $F \in \mathbf{R}^{2a_1+2a_2}$ is given as follows:

$$(I*F)_{n,m} = \sum_{k=-a_1}^{a_1} \sum_{l=-a_2}^{a_2} F_{k,l} I_{n-k,m-l}, \tag{1}$$

where a_1 and a_2 determine the size of the convolution kernel along the x and y directions. ReLU $(g(z) = \max(0, z))$ [25] is a non-linear function which is applied to feature mapping created by the convolutional layer. The BN [26] layer normalizes each mini batch throughout the entire network, reducing the internal covariate shift caused by the progressive transforms. The BN layer is used to reduce the training time of the CNN and the sensitivity of network initialization. Therefore, this layer is used for normalization in the proposed network model.

2.2.2MPS Tensor Network

The tensor network notation is a brief graphical representation of the high-dimensional tensors. It not only makes it easier and more intuitive to process the high-dimensional tensors but also provides an insight into how to achieve more efficient operations. For a more comprehensive introduction to the tensor networks, references in [27]can be referred.

The MPS (matrix product state) [17, 18] is a one-dimensional tensor network structure, which is based on tensor train decomposition [28]. It uses chain-connected small tensors to represent the high-dimensional tensors.

For a neural network model based on the tensor network, the generated Mel spectrograms must first be mapped to the high-dimensional space corresponding to the tensor network. According to **Eq. 2**, each pixel of the Mel spectrogram is mapped to a two-dimensional space.

$$\left|x_n^{[l]}\right\rangle = \cos\frac{x_n^{[l]}\pi}{2}|0\rangle + \sin\frac{x_n^{[l]}\pi}{2}|1\rangle, \tag{2}$$

where $|\rangle$ is the Dirac symbol in physics, representing the state vector. $|0\rangle$ means blue with low energy, and $|1\rangle$ means yellow with high energy, where l represents the order of the Mel spectrogram, and n represents the pixel order in the Mel spectrogram. The function with cos(πx/2) and sin(πx/2) is one of the mapping methods. After inputting the spectrogram, the data of each pixel are normalized to be between 0 and 1; using cos(πx/2) and sin(πx/2) can accurately represent the information in the pixel. After mapping, $|x_n^{[l]}\rangle$ can represent all the magnitudes of energy in the Mel spectrogram. After all the pixels are mapped, Mel spectrograms can be expressed as **Eq. 3** and also be expressed as **Eq. 4** using the tensor network notation.

$$\left|X^{[l]}\right\rangle = \left|x_1^{[l]}\right\rangle \otimes \left|x_2^{[l]}\right\rangle \otimes \ldots \otimes \left|x_N^{[l]}\right\rangle, \tag{3}$$

$$\mathbf{\Phi}(x) = \boldsymbol{\phi}(x_1) \otimes \boldsymbol{\phi}(x_2) \otimes \ldots \otimes \boldsymbol{\phi}(x_N), \tag{4}$$

where ⊗ represents the tensor product. x represents the Mel spectrogram of each input, and N is the total number of pixels in the Mel spectrogram. $\phi(x_1)$ is the representation of the first pixel in the Mel spectrogram mapped to a two-dimensional space, and $\Phi(x)$ is the high-dimensional mapping form of the Mel spectrogram. Given the high-dimensional features, for the input Mel spectrogram, the decision function of the event tagging can be expressed as

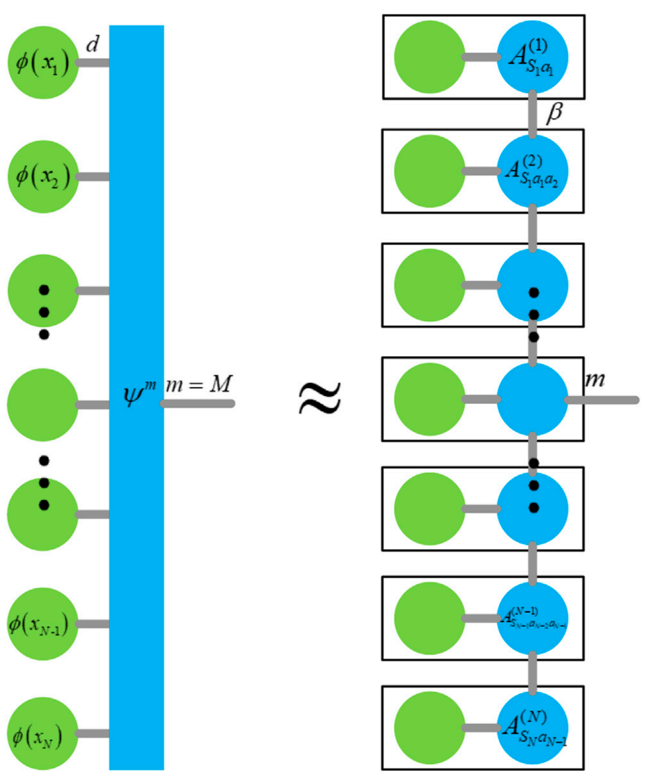


FIGURE 3 | Linear model of the decision module in audio tagging.

$$f^m(x) = \psi^m \cdot \mathbf{\Phi}(x), \tag{5}$$

$$m = \arg\max f^m(x). \tag{6}$$

Here, m represents M categories, $m = [0, 1, \ldots, M-1]$, where ψ^m is the trainable weight tensor. The model of the decision module in audio tagging is shown on the left of **Figure 3** and in **Eq. 5**. ψ^m is a weight tensor, and its dimension is as high as $M \cdot 2^N$, which is difficult to be calculated. After decomposing ψ^m into the chained small tensors through the MPS, the two-dimensional space that can be mapped with each pixel can be contracted with the weight tensor ψ^m. In this way, the calculation can only be carried out between the small tensors, without directly calculating the weight tensors with high dimensionality. **Figure 3** is a linear model of the decision module in audio tagging represented by the tensor network notation. For details on the tensor network notation, reference in [27]can be referred. As shown by the small green tensor in **Figure 3**, $\Phi(x)$ is the form in which the two-dimensional space mapped by each pixel is connected to the weight tensor ψ^m. The nodes in the first column are the pixels of each Mel spectrogram after being mapped to the two-dimensional space. They are connected to the weight tensor obtained after the training. There is an index m on the right side of ψ^m, whose dimension is the number of the final tagging classes.

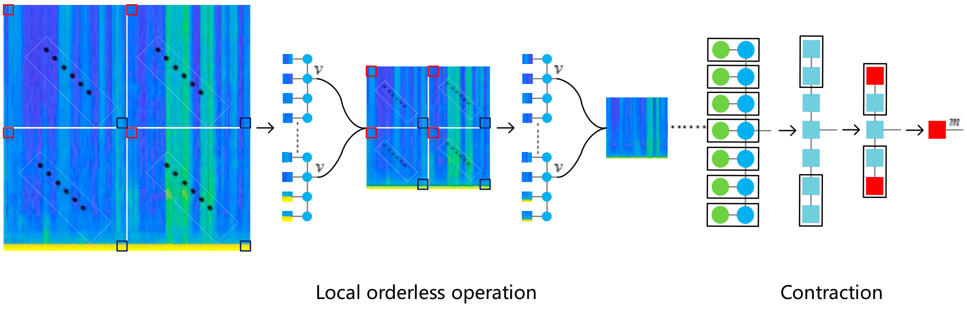


FIGURE 4 | Local orderless operation and contraction.

This mapping method will result in a huge number of parameters in the weight tensor. The matrix product state is the name for tensor train decomposition in physics. It approximates a large tensor to the product form of several second-order and third-order tensors. In this way, the contraction can be performed in the way on the right side of **Figure 3**, to avoid the direct calculation of the ultra-high dimensional tensor, and the calculation amount will be greatly reduced. A high-dimensional tensor T is decomposed into an approximate tensor $\tilde{T}$ by the tensor train [28], as shown in **Eq. 7**.

$$\tilde{T} = \sum_{a_1 a_2 \ldots a_{N-1}} A^{(1)}_{S_1 a_1} A^{(2)}_{S_1 a_1 a_2} \ldots A^{(N-1)}_{S_{N-1} a_{N-2} a_{N-1}} A^{(N)}_{S_N a_{N-1}}. \tag{7}$$

The weight tensor ψ^m is approximated by the product form of some two-dimensional and three-dimensional tensors according to **Eq. 7**. The approximated weight tensor is shown in **Eq. 8** and on the right of **Figure 3**.

$$\psi^{m,i_1,i_2,\ldots,i_N} = \sum_{\alpha_1,\alpha_2,\ldots\alpha_N} A^{i_1}_{\alpha_1} A^{i_2}_{\alpha_1\alpha_2} A^{i_3}_{\alpha_2\alpha_3} \ldots A^{m,i_j}_{\alpha_j\alpha_{j+1}} \ldots A^{i_N}_{\alpha_N}, \tag{8}$$

where$A$ is the decomposed second-order and third-order tensors. The subscript i_j is called the free index, and the free index m corresponds to the right side of **Figure 3**, and its dimension is the number of tagging classes. The subscript a_j is an auxiliary indicator, and its dimension is called the bond dimension, which controls the quality of the approximation. The size of the bond dimension determines the size of the tensor. The components of the tensor A are the variational parameters determined through the training.

2.2.3 Local Orderless Operation

Since MPS is a one-dimensional tensor network, the neighboring pixels in the spectrogram are usually highly correlated. Therefore, directly flattening and inputting the Mel spectral feature into the MPS layer will cause the loss of spatial information. Spatial information includes the information of a single frame in the vertical direction, as well as the information between the frames in the horizontal direction, which is very important for audio tagging. In order to solve this problem, the local orderless operation according to the local orderless theory is used in the tensor network [29, 30]. The local orderless operation divides a large patch into many small patches. After the small patches are contracted, the dimension of the output vector is $\boldsymbol{v}$, and $\boldsymbol{v}$ is set to the same size as the bond dimension. This step can be interpreted as using a vector of dimension $\boldsymbol{v}$ to represent small patches of information, similar to feature extraction. Each small patch contains global features, which can better preserve the spatial information.

First, the Mel spectrogram is divided into four parts, as shown in **Figure 4**. The first pixel of each part is taken out and combined into a 2 × 2 local orderless small patch, as shown in the red box in **Figure 4**. Then, the pixels of each part are combined according to this step, until the last pixel in the black box, as shown in **Figure 4**. The pixel order in the patch is shown in **Eq. 9**.

$$\begin{aligned} \mathrm{P}^K &= \begin{pmatrix} K & , & K + \dfrac{\mathrm{W}}{2} \\ K + \dfrac{\mathrm{H} \times \mathrm{W}}{2} & , & K + \dfrac{(\mathrm{H}+1) \times \mathrm{W}}{2} \end{pmatrix} \quad \forall K \\ &= 1, \ldots, (\mathrm{H} \times \mathrm{W})/4, \end{aligned} \tag{9}$$

where P^K represents the local orderless small patch, the superscript K represents the sequential number of small patches, and H and W represent the height and width of the Mel spectrogram, respectively.

Then the small patches are flattened and input into the MPS layer to contract. Then all the output vectors $\boldsymbol{v}$ are reshaped into images. The converted graph has a smaller resolution than the previous Mel spectrogram, but the important information will be preserved. This operation is repeated on the converted image. After the three MPS layers, the resolution of the generated image is already very small, but the features and spatial information of the original Mel spectrogram are well-preserved.

2.2.4 Contraction and Optimization

After the three MPS layers of contraction, a small size image has been generated. It has spatial structure information and important features of the Mel spectrogram. It is flattened into the last MPS layer, as shown in **Figure 4**. In line with the implementation method from the MPS in Miller [31], the horizontal edges are first contracted in parallel to get the contracted tensors, and then, these tensors are contracted vertically. The output is generated by connecting the free indicators of the tensor. A recent work has proposed a more effective calculation method [32, 33], which is expected to further accelerate the calculation speed.

3 EXPERIMENTS

3.1 Datasets

In the experiment, development and validation datasets of DCASE 2018 challenge task 5 are used to evaluate the audio tagging for domestic activities. DCASE 2018 challenge task 5 is a derivative of the SINS dataset. It contains a continuous recording of one person living in a holiday home over a period of 1 week. It was collected using a network of 13 microphone arrays distributed over the entire home. The microphone array consisted of four linearly arranged microphones. For this task, seven microphone arrays are used in the living room and kitchen area combined. The continuous recordings are split into audio segments of 10 s. These audio segments are provided as individual files along with the ground truth. The dataset contains 72,984 audio files. Each audio segment contains four channels. It is organized with nine class labels consisting of absence, cooking, dishwashing, eating, social activity, vacuum cleaning, watching TV, and working. The audio files are recorded with 16 kHz sampled frequency, and the number of files in each class are not the same.

3.2 Evaluation Method

In this experiment, the development dataset and validation dataset are divided into the training set, validation set, and test set with a ratio of 8:1:1, respectively. The evaluation criteria include the precision rate, recall rate, and F1-score. Precision is the ratio of real positive samples to samples that are predicted to be positive, which is specific to the predicted samples. Recall is the ratio of the correct predictions to the positive cases in the sample, which is specific to the actual samples. The F1 score is calculated based on recall and precision. The experimental results in this article are the results of the development and the validation datasets in the divided test set, respectively. These criteria are obtained by calculating the confusion matrix given by **Eqs 10–12**.

$$Precision = \frac{TP}{TP + FP}, \tag{10}$$

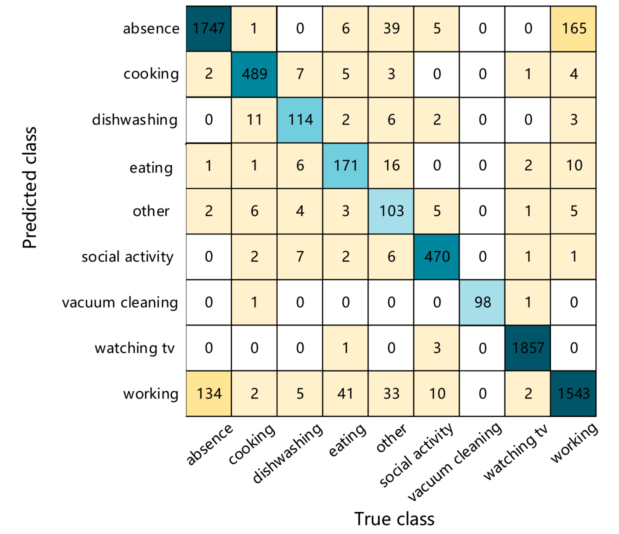
$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}, \tag{12}$$

where TP is the number of true positive results, TN is the number of true negative results, FP is the number of false positive results, and FN is the number of false negative results.

3.3 Experimental Setup and Result

There are four channels (C_1, C_2, C_3, C_4) within one audio signal. The four channels are manually averaged [21] to yield C_5, where $C_5 = (C_1 + C_2 + C_3 + C_4)/4$ so as to better fuse the four channels and augment the dataset. The audio signal is converted to a Mel spectrogram, as described in **Section 2**. The window type, window size, overlap, and FFT size parameters are set to Hamming, 480, 240, and 480, respectively. The Hamming window is adopted for signal framing as it can effectively overcome the leakage phenomenon [34]. The dimension of the Mel spectrogram is 336 × 336 × 3 as the input of the neural network model based on the tensor network, which is composed of the two convolutional layers and four MPS layers. It is then the horizontal flip, vertical flip, and random rotation that enlarge the training data, avoid overfitting, and enhance the robustness of the model. The batch size is set to 256, bond dimension is set to 5, and the initial learning rate is 0.001. The optimizer and loss function used in the training are Adam and cross-entropy loss function. The structure of the neural network model based on the tensor network is shown in **Figure 2**, and the states of TP, TN, FP, and FN in the test set of the development dataset are shown for each class on the confusion matrix in **Figure 5**.



	absence	cooking	dishwashing	eating	other	social activity	vacuum cleaning	watching tv	working
absence	1747	1	0	6	39	5	0	0	165
cooking	2	489	7	5	3	0	0	1	4
dishwashing	0	11	114	2	6	2	0	0	3
eating	1	1	6	171	16	0	0	2	10
other	2	6	4	3	103	5	0	1	5
social activity	0	2	7	2	6	470	0	1	1
vacuum cleaning	0	1	0	0	0	0	98	1	0
watching tv	0	0	0	1	0	3	0	1857	0
working	134	2	5	41	33	10	0	2	1543

FIGURE 5 | Confusion matrix for the test set of the development dataset.

As can be seen from the confusion matrix in **Figure 5**, the abscissa is the true class, and the ordinate is the predicted class. The blue square indicates that the predicted class is the same as the true class. The color intensity corresponds to the number of audio tagging. It can be found from **Figure 5** that the proposed model judges 165 working audios as absence and 134 absence audios as working. Because people may make very small noises at work, it is easy to confuse it with the absence class. In addition, the model judges 39 and 33 other class audios as absence and working class, and many labels for other class audios cannot be distinguished since the other class is not a class of specific activities. There are many types of features extracted from the other class audio, and the common features of other class are difficult to learn. As a result, a lot of audio signals are near the decision boundary, and it is easy to be misjudged as absence, working, and eating. But for the prediction results, it can be found from **Figure 5** that the prediction results for the other class are more accurate, proving the better learning ability of the tensor network model.

In order to compare with other models more intuitively, other performance criteria including precision, recall, and F1-score are separately given in **Table 1** for each class of DCASE 2018 challenge task 5. It can be seen from the **Table 1** that the

TABLE 1 | Neural network model based on the tensor network performance criteria in the test set of the development dataset.

Class	Precision/%	Recall/%	F1-score/%
Absence	89.00	92.63	90.78
Cooking	95.69	95.30	95.49
Dishwashing	82.61	79.72	81.14
Eating	82.61	74.03	78.09
Other	79.84	50.00	61.49
Social activity	96.11	94.95	95.53
Vacuum cleaning	98.00	100.00	98.99
Watching TV	99.79	99.57	99.68
Working	87.18	89.14	88.15
Average value	90.09	86.15	87.70

precision rate of the other class is much higher than the recall rate. This shows that the prediction of the other class is more accurate, but many other class audios are prone to judgment errors. Since the other class is not a class of specific activities, the tensor network can better learn the common features of the class. But for the other class audio with less commonality, it is less possible to identify the deeper rules.

Table 2 is a comparison between the proposed model and other models, which represent several typical and commonly used networks, including the CNN, RESNET, and LSTM. This experiment selected the three models to compare with the neural network model based on tensor network (NNMBTN) model, namely, the baseline system [9], TC2DCNN [12], and INRC_2D [13]. The baseline system uses a neural network architecture based on the convolutional layers and dense layers. TC2DCNN is extended by operating the convolutions along the two dimensions of time and channel. INRC_2D is processed in parallel by RESNET and long short-term memory (LSTM) network, with a fully joint connected layer.

It can be seen from **Table 2** that the F1-score of the proposed method on the test set of the DCASE 2018 challenge task 5 development set is 87.70%, which is 3.2% higher than the baseline system, 1.95% higher than the TC2DCNN system, and 0.86% higher than INRC_2D system. The proposed method has five classes that are higher than the baseline, TC2DCNN, and INRC_2D. This shows that the tensor network model can identify the important features well after obtaining the features extracted by the convolutional layer. At the same time, the spatial information of the audio is well-preserved. Compared with the other models, the tensor network has powerful representation ability in the high-dimensional space and can separate the different classes of audio with hyperplane. There is little difference in the F1-score performance on cooking, vacuum cleaning, and working. The score advantage of other classes is more obvious, 5.61% higher than the INRC_2D system, which shows that for classes of not specific activities, the tensor network can also learn the features better.

The data provided in the evaluation set are based on the sensor nodes that do not exist in the development set and can provide data from the same nodes in the development set. The F1-scores of each model in the test set of the validation set are shown in **Table 3**.

Compared with the results on the development set, it can be seen from **Table 3** that the proposed model in the test set of the validation set is lower than the other models in the two categories of eating and social activities and higher than the other models in both categories of dishwashing and vacuuming. The advantages of the other categories are still obvious. The average F1-score reaches 85.9%, which is 9.0% higher than TC2DCNN, 4.2% higher than INRC2D, and 2.8% higher than the baseline. The F1-scores of the neural network model based on the tensor network are relatively stable, which proves that the proposed network has a good generalization ability. On the whole, the proposed model has better ability to extract and learn the important features of the data.

In order to better demonstrate the compression ability of the MPS to the network, the MPS layer in the proposed model is replaced by the convolutional layer, max pooling layer, and fully connected layer. We compared the proposed model (NNMBTN: 2CNN+4MPS) with the traditional CNN-based model which is composed of four CNNs, Maxpool, and a fully connected layer. The model comparison results are shown in **Table 4**.

It can be seen from **Table 4** that the parameters of the proposed model are one quarter smaller than that of the traditional neural network after replacement, and the effect is also better than that of the traditional neural network, which

TABLE 2 | Comparison of the neural network model based on the tensor network with other models in the test set of the development dataset.

Class	Detecting F1-score (%) for the used methods			
	Baseline system [9]	TC2DCNN [12]	INRC_2D [13]	NNMBTN
Absence	85.41	86.62	83.95	90.78
Cooking	95.14	93.34	95.47	95.49
Dishwashing	76.73	72.68	78.00	81.14
Eating	83.64	87.03	89.68	78.09
Other	44.76	53.81	55.88	61.49
Social activity	93.92	93.94	93.97	95.53
Vacuum cleaning	99.31	99.79	100.00	98.99
Watching TV	99.59	99.38	99.40	99.68
Working	82.03	85.14	85.22	88.15
Average value	84.50	85.75	86.84	87.70

TABLE 3 | Comparison of the neural network model based on the tensor network with other models in the test set of the validation dataset.

Class	Detecting F1-score (%) for the used methods			
	Baseline system [9]	TC2DCNN [12]	INRC_2D [13]	NNMBTN
Absence	87.7	79.8	79.7	90.2
Cooking	93.0	88.7	86.9	95.0
Dishwashing	77.2	71.8	73.8	82.3
Eating	81.2	78.9	82.2	77.0
Other	35.0	17.6	42.7	55.5
Social activity	96.6	96.2	97.1	93.4
Vacuum cleaning	95.8	94.4	97.4	98.2
Watching TV	99.9	99.7	99.9	99.5
Working	81.4	64.6	75.5	82.3
Average value	83.1	76.9	81.7	85.9

TABLE 4 | Performance and parameter comparison between the proposed model and the traditional neural network.

Model	Precision/%	Recall/%	F1-score/%	Parameter quantity (M)
4CNN + Maxpool + fully connected	74.11	64.08	65.8	23.70
NNMBTN (2CNN+4MPS)	88.08	84.13	85.9	17.74

TABLE 5 | Performance comparison between the separable convolution model and the separable convolution model combined with tensor networks.

Model	F1-score/%	Parameter quantity (M)	GPU(GB)
Separable Convolutions [5] (batch size = 128)	90.78	4.20	3.85
(2SepConv+3MPS) (batch size = 128)	89.52	4.16	1.72

shows that the MPS layer has better compression ability for the network.

To further investigate the effect of combining the proposed model with the state-of-the-art model, separable convolutions network [5] are used to verify the feasibility of the proposed model. Separable convolutions network consists of four convolutional layers using 5 × 5 filters, followed by a global pooling layer and a final MLP (Multilayer Perceptron). The separate convolution network structure is improved to be combined with the MPS tensor network in which only two layers of separate convolutions network are retained. In the comparison experiments, only the network structure was changed, and the rest remained unchanged. The experimental results are shown in **Table 5**.

It can be seen from **Table 5** that the GPU occupancy in the training procedure is reduced by 55% after combining with the tensor network under the same conditions except for the network structure. This shows that the tensor network can better reduce the redundancy of the network during the training. In terms of parameter quantity, the parameter quantity of the separate convolution is slightly smaller than that of ordinary convolution. The F1-score is slightly lower than the split convolutional network. Compared with the state-of-the-art model, the combination of the tensor network can reduce the redundancy of the network to achieve a balance between efficiency and accuracy. In the future, more research practices could be carried out to find a better way when combining the tensor network with the new neural network approaches.

4 CONCLUSION

In this article, the neural network model based on the tensor network is proposed for audio tagging of domestic activities, which takes the advantage of the CNN in extracting spatial features and the MPS tensor network for better interpretability and the ability to compress the network with tensor train decomposition. The MPS is one-dimensional tensor network structure, which is based on tensor train decomposition. It uses the chain-connected small tensors to represent the high-dimensional tensors. The proposed model is composed of two convolutional layers and four MPS layers. The function of the first three MPS layers is to extract the features, and the last MPS layer is used as a classifier. The DCASE 2018 challenge task 5 datasets are considered in the experiment, and the F1-score is calculated for performance evaluation. The experimental results show that the neural network model based on the tensor network proposed in this article has a good learning ability. The results show that the average F1-Score of the proposed neural network model based on the tensor network in the test set of the development dataset and validation dataset of DCASE 2018 challenge task 5 reached 87.7 and 85.9%, which were 3.2 and 2.8% higher than the baseline system, respectively. When compared with the state-of-the-art model, the combination of the tensor network can reduce the redundancy of the network to achieve a balance between the efficiency and accuracy. It is verified that the proposed model can function better for the task of audio tagging of domestic activities.

In the future, it is necessary to extract more representative audio features in the face of a huge database. There are some other structures of tensor networks, such as PEPS and MERA, and the combination of these models with the neural networks deserves a further in-depth study. In addition, the classes of the sound events in household activities are more complex, so expanding the dataset and improving the audio tagging accuracy are also necessary.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

L-DY and R-BY mainly carried out topic selection, experiments, and completion of the first draft. JW mainly checked the logical structure and writing of the manuscript. ML was mainly responsible for the revision and polishing of the manuscript.

FUNDING

The Central Government Guides Local Science and Technology Development Fund Project of China (grant number: 2021ZY0004). This work was supported by the National Natural Science Foundation of China (Grant Nos. 62071039 and 62161040), the Natural Science Foundation of Inner Mongolia (Grant No. 2021MS06030), and the Science and Technology Major Project of Inner Mongolia (Grant No. 2021GG0023).

REFERENCES

1. Rafferty J, Nugent CD, Liu J, Chen L. From Activity Recognition to Intention Recognition for Assisted Living within Smart Homes. *IEEE Trans Human-mach Syst* (2017) 47(3):368–79. doi:10.1109/thms.2016.2641388
2. Erden F, Velipasalar S, Alkar AZ, Cetin AE. Sensors in Assisted Living: A Survey of Signal and Image Processing Methods. *IEEE Signal Process Mag* (2016) 33(2):36–44. doi:10.1109/msp.2015.2489978
3. Phan H, Hertel L, Maass M, Koch P, Mazur R, Mertins A. Improved Audio Scene Classification Based on Label-Tree Embeddings and Convolutional Neural Networks. *Ieee/acm Trans Audio Speech Lang Process* (2017) 25(6): 1278–90. doi:10.1109/taslp.2017.2690564
4. Gong Y, Chung Y-A, Glass J. Psla: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *Ieee/acm Trans Audio Speech Lang Process* (2021) 29:3292–306. doi:10.1109/taslp.2021.3120633
5. Bursuc A, Puy G, Jain H. *Separable Convolutions and Test-Time Augmentations for Low-Complexity and Calibrated Acoustic Scene Classification*. Barcelona, Spain: Detection and Classification of Acoustic Scenes and Events 2021 (2021).
6. Dekkers G, Lauwereins S, Thoen B, Adhana MW, Brouckxon H, Van den Bergh B, et al. *The Sins Database for Detection of Daily Activities in a Home Environment Using an Acoustic Sensor Network*. Munich, Germany: Detection and Classification of Acoustic Scenes and Events 2017 (2017). p. 1–5.
7. Wang J, Xie X, Kuang J. Microphone Array Speech Enhancement Based on Tensor Filtering Methods. *China Commun* (2018) 15(4):141–52. doi:10.1109/cc.2018.8357692
8. Yang L, Liu M, Wang J, Xie X, Kuang J. Tensor Completion for Recovering Multichannel Audio Signal with Missing Data. *China Commun* (2019) 16(4): 186–95. doi:10.12676/j.cc.2019.04.014
9. Dekkers G, Vuegen L, van Waterschoot T, Vanrumste B, Karsmakers P. *Dcase 2018 Challenge-Task 5: Monitoring of Domestic Activities Based on Multi-Channel Acoustics*. Surrey, United Kingdom: arXiv preprint arXiv:180711246 (2018).
10. Inoue T, Vinayavekhin P, Wang S, Wood D, Greco N, Tachibana R. *Domestic Activities Classification Based on Cnn Using Shuffling and Mixing Data Augmentation*. Surrey, United Kingdom: DCASE 2018 Challenge (2018).
11. Tanabe R, Endo T, Nikaido Y, Ichige T, Nguyen P, Kawaguchi Y, et al. *Multichannel Acoustic Scene Classification by Blind Dereverberation, Blind Source Separation, Data Augmentation, and Model Ensembling*. Surrey, United Kingdom: DCASE 2018 Challenge (2018).
12. Tiraboschi M. *Monitoring of Domestic Activities Based on Multi-Channel Acoustics: A Time-Channel {2d}-Convolutional Approach*. Surrey, United Kingdom: DCASE 2018 Challenge (2018).
13. Raveh A, Amar A. *Multi-Channel Audio Classification with Neural Network Using Scattering Transform*. Surrey, United Kingdom: Tech. Rep. DCASE (2018).
14. Kong Q, Iqbal T, Xu Y, Wang W, Plumbley MD. *Dcase 2018 Challenge Surrey Cross-Task Convolutional Neural Network Baseline*. Surrey, United Kingdom: arXiv preprint arXiv:180800773 (2018).
15. Hofmann T, Schölkopf B, Smola AJ. Kernel Methods in Machine Learning. *Ann Stat* (2008) 36(3):1171–220. doi:10.1214/009053607000000677
16. Wang J, Liu M, Xie X, Kuang J. Compression of Head-Related Transfer Function Based on Tucker and Tensor Train Decomposition. *IEEE Access* (2019) 7:39639–51. doi:10.1109/access.2019.2906364
17. Stoudenmire EM, Schwab DJ. *Supervised Learning with Quantum-Inspired Tensor Networks*. Barcelona, Spain: arXiv preprint arXiv:160505775 (2016).
18. Efthymiou S, Hidary J, Leichenauer S. *Tensornetwork for Machine Learning*. Ithaca, New York: arXiv preprint arXiv:190606329 (2019).
19. R Selvan EB Dam, editors. *Tensor Networks for Medical Image Classification*. Montreal, QC: Medical Imaging with Deep Learning (2020).
20. Evenbly G, Vidal G. Tensor Network States and Geometry. *J Stat Phys* (2011) 145(4):891–918. doi:10.1007/s10955-011-0237-4
21. Liu H, Wang F, Liu X, Guo D. *An Ensemble System for Domestic Activity Recognition*. Surrey, United Kingdom: DCASE2018 Challenge, Tech Rep (2018).
22. SK Kopparapu M Laxminarayana, editors. Choice of Mel Filter Bank in Computing Mfcc of a Resampled Speech. In: Proceedings of the 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010); 2010 May 10; Kuala Lumpur, Malaysia. IEEE (2010).
23. K Yanai Y Kawano, editors. Food Image Recognition Using Deep Convolutional Network with Pre-training and Fine-Tuning. In: Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW); 2015 June 29; Turin, Italy. IEEE (2015).
24. Kalchbrenner N, Grefenstette E, Blunsom P. *A Convolutional Neural Network for Modelling Sentences*. Ithaca, New York: arXiv preprint arXiv:14042188 (2014).
25. Schmidt-Hieber J. Nonparametric Regression Using Deep Neural Networks with Relu Activation Function. *Ann Stat* (2020) 48(4):1875–97. doi:10.1214/19-aos1875
26. Sigtia S, Benetos E, Dixon S. An End-To-End Neural Network for Polyphonic Piano Music Transcription. *Ieee/acm Trans Audio Speech Lang Process* (2016) 24(5):927–39. doi:10.1109/taslp.2016.2533858
27. Bridgeman JC, Chubb CT. Hand-Waving and Interpretive Dance: An Introductory Course on Tensor Networks. *J Phys A: Math Theor* (2017) 50(22):223001. doi:10.1088/1751-8121/aa6dc3
28. Oseledets IV. Tensor-Train Decomposition. *SIAM J Sci Comput* (2011) 33(5): 2295–317. doi:10.1137/090752286

29. Koenderink JJ, Van Doorn AJ. The Structure of Locally Orderless Images. *Int J Comput Vis* (1999) 31(2):159–68. doi:10.1023/a:1008065931878
30. Oron S, Bar-Hillel A, Levi D, Avidan S. Locally Orderless Tracking. *Int J Comput Vis* (2015) 111(2):213–28. doi:10.1007/s11263-014-0740-6
31. Miller J. Torchmps (2019). Available from: https://githubcom/jemisjoky/torchmps (Accessed March 1, 2019).
32. Fishman M, White SR, Stoudenmire EM. *The Itensor Software Library for Tensor Network Calculations*. Ithaca, New York: arXiv preprint arXiv: 200714822 (2020).
33. Novikov A, Izmailov P, Khrulkov V, Figurnov M, Oseledets IV. Tensor Train Decomposition on Tensorflow (T3f). *J Mach Learn Res* (2020) 21(30):1–7.
34. W Astuti, W Sediono, A Aibinu, R Akmeliawati, M-JE Salami, editors. Adaptive Short Time Fourier Transform (Stft) Analysis of Seismic Electric Signal (Ses): A Comparison of Hamming and Rectangular Window. In: Proceedings of the 2012 IEEE Symposium on Industrial Electronics and Applications; Bandung, Indonesia; 2012 September 23. IEEE (2012).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Digital Health

ORIGINAL RESEARCH
published: 06 June 2022
doi: 10.3389/fdgth.2022.873822

Applying Meta-Learning and Iso Principle for Development of EEG-Based Emotion Induction System

Kana Miyamoto[1,2]*, *Hiroki Tanaka*[1,2] *and Satoshi Nakamura*[1,2]

[1] *Division of Information Science, Nara Institute of Science and Technology, Nara, Japan,* [2] *Center for Advanced Intelligence Project, RIKEN, Nara, Japan*

Music is often used for emotion induction. ince the emotions felt when listening to it vary from person to person, customized music is required. Our previous work designed a music generation system that created personalized music based on participants' emotions predicted from EEG data. Although our system effectively induced emotions, unfortunately, it suffered from two problems. The first is that a long EEG recording is required to train emotion prediction models. In this paper, we trained models with a small amount of EEG data. We proposed emotion prediction with meta-learning and compared its performance with two other training methods. The second problem is that the generated music failed to consider the participants' emotions before they listened to music. We solved this challenge by constructing a system that adapted an iso principle that gradually changed the music from close to the participants' emotions to the target emotion. Our results showed that emotion prediction with meta-learning had the lowest RMSE among three methods ($p < 0.016$). Both a music generation system based on the iso principle and our conventional music generation system more effectively induced emotion than music generation that was not based on the emotions of the participants ($p < 0.016$).

Keywords: electroencephalogram (EEG), emotion induction, emotion prediction, music generation, meta-learning, iso principle

OPEN ACCESS

Edited by:
Kun Qian,
Beijing Institute of Technology, China

Reviewed by:
Jian Guo,
RIKEN Center for Computational Science, Japan
Meishu Song,
University of Augsburg, Germany

Correspondence:
Kana Miyamoto
miyamoto.kana.mk4@is.naist.jp

Specialty section:
This article was submitted to Frontiers in Health Informatics, a section of the journal Frontiers in Digital Health

Received: *11 February 2022*
Accepted: *06 May 2022*
Published: *06 June 2022*

Citation:
Miyamoto K, Tanaka H and Nakamura S (2022) Applying Meta-Learning and Iso Principle for Development of EEG-Based Emotion Induction System.
Front. Digit. Health 4:873822.
doi: 10.3389/fdgth.2022.873822

1. INTRODUCTION

Appropriate emotional induction is important for mental health (1–3). Many research attempts have used music to induce emotions. Even though such musical elements as rhythm and tempo induce emotions (4), not every person feels the same emotions when they listen to the same piece of music (5). In addition, the same person might experience different emotions depending on the situation. Therefore, it is challenging to induce emotions using music that takes into account the emotions of participants (6, 7). Using a subjective evaluation is a simple method for obtaining the emotions of participants. The Self-Assessment Mannequin (SAM) is often utilized for such emotional evaluation (8). However, since real-time subjective evaluations burden participants, using physiological signals has been proposed to predict emotions. Since electroencephalogram (EEG) has a high temporal resolution and are expected to be used for computer-human interaction, our work induces emotions by generating music with emotions predicted from them.

We developed a system that generates music based on participants' emotions using their EEG data (9). It consists of three elements. The first is a music generator. We treat emotions on two axes, valence and arousal, based on the circumplex model (10). The target emotion's valence and arousal to be induced are set in a range from 0 to 1 and input to a music generator, which creates music that induces an emotion similar to the inputted emotion. Note that depending on the individual differences in the feeling of an emotion and the participant's state, the input emotion and the actual emotion may not be identical. The second element is emotion prediction based on EEG. We previously showed that a convolutional neural network (CNN), which takes into account the positional relationships of EEG electrodes, effectively predicted emotions (9). The system uses a CNN to predict the participants' emotions while they listen to music in real-time. The third element is the control of a music generator. The system calculates the difference between the target emotion and the participants' emotion predicted by the EEG and adds it to the music generator's previous input. By changing the inputs of the music generator based on the participants' emotions, the system creates music that matches their characteristics. We previously verified our system that consists of these elements with six participants. We used a baseline method that generates music without considering the participants' emotions by continuously inputting the target emotion into the music generator. Our proposed method used the system to generate music from the participants' emotions in real-time. After comparing these two methods, the distance between the target emotion and the emotion that was finally induced was smaller in the proposed method, suggesting the effectiveness of the system. However, it has two problems, which we address in this paper.

The first problem is that it takes a long time to record EEG data for training the emotion prediction models because a sufficient amount of EEG data is required to train them. In our experiment, the EEG data were recorded for only 30 min, considering the burden placed on the participants. Because of this time factor, we trained an emotion prediction model using EEG data for 30 min and used the system on a different day. Even though the EEG recording time must be shortened to improve the system's usability, a lack of EEG data negatively impacts model training (11). Previous studies solved this problem by proposing transfer learning (9, 12), which adapts a pre-trained model from one domain to another (13). With a small amount of EEG data to retrain a pre-trained model which was trained on a large amount of EEG data, more accurate predictions can be made than with just a small amount of EEG data. However, in previous studies, the pre-training model mixed the EEG data of multiple people and treated them as one big amount of data (9, 12). Perhaps individual EEG characteristics cannot be taken into account because no individuals are recognized. Meta-learning, which is an effective solution to this problem, has been used for few-shot learning, fast many-shot optimization, robustness to domain-shift, and so on (14). It helps models acquire experience through multiple tasks with which to improve future learning performance. There are gradient-descent, reinforcement learning, and evolutionary search as its optimizer. Some previous studies on EEG predictions trained models with one person's EEG data as a single task and demonstrated the effectiveness of meta-learning. Model-agnostic meta-learning (MAML) (15) is one popular type of meta-learning with gradient-descent. MAML trains an initial model that easily adapts to any task from multiple tasks. Therefore, the initial model can adapt to new tasks with a small amount of data. Previous studies predicted sleep levels (16) and motor imagery (17) using EEG with MAML. MAML was also used for emotion prediction using EEG, and its effectiveness was investigated using the DEAP dataset with music video stimuli and the SEED dataset with movie stimuli (18). However, we use only music, which is an auditory stimulus. Since the type of stimulus influences emotion elicitation (19), we believe that the effectiveness of applying MAML must be tested for emotion prediction while listening to music. We have two baseline methods: one trains a model using multiple participants' EEG data without MAML, and the other trains a model with a small amount of a single participant's EEG data. We compared the emotion prediction performance of the proposed method with two baseline methods and constructed an emotion induction system using a model trained with MAML.

The second problem is that the music generator creates music without identifying the participants' emotions before they listen to it. We showed that the inputs to the music generator and the emotions felt by the participants are similar. Therefore, we tried to increase/decrease the music generator's inputs based on the participants' emotions using empirically-determined formulas. Here is an example of a case where we want to induce a high valence. First, a high valence, which is the target emotion, is input into the music generator. The participants listen to the generated music. If they experience a low valence, the next valence input will be higher, and the music generator will try to induce a higher valence. As shown above, we proposed a method that makes music that rapidly moves a participants' emotion toward the target emotion, starting from the beginning of listening to a piece of music, and adjusts the music generator's inputs based on their states. The proposed method more effectively induced emotions than the baseline method in which the target emotion is continuously input to the music generator. In music therapy, the iso principle, which is used in emotion induction (20, 21), plays music that is close to the participant's emotion and gradually leads them to the target emotion. In a previous study (22), participants with sadness listened to two pieces of music: sad music or happy music. The results showed that listening to happy music after sad music induced more positive emotions. Our previously proposed method rapidly induced emotions to target emotions without considering the emotions of the participants before they listened to music. In this paper, we develop a system based on this iso principle and investigate whether music generation with it and our conventional music generation effectively induce emotions.

Our paper provides the following two contributions:

1. It verifies EEG-based emotion prediction using meta-learning.

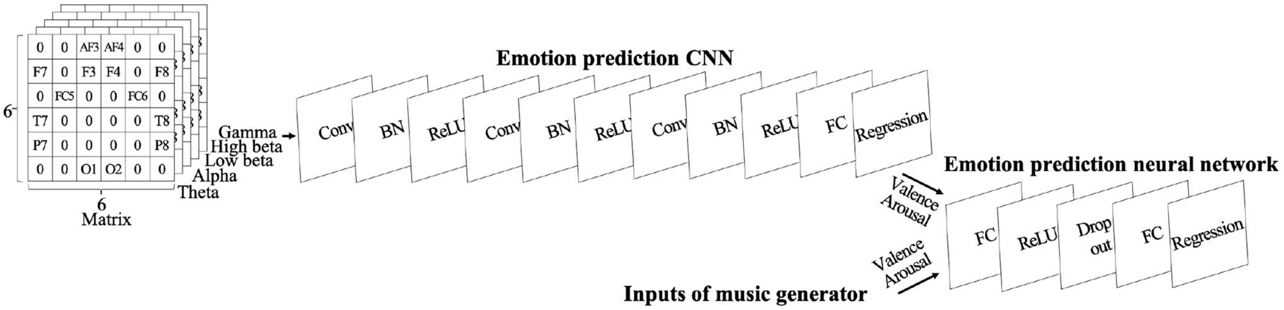


FIGURE 1 | Model structure: Upper part is a CNN that predicts emotion from EEG data. Right part is a neural network that uses predicted emotions from CNN and music generator's inputs for emotion prediction.

2. It evaluates an emotion induction system using meta-learning and the iso principle.

This paper is an extension of our previously proposed emotion induction system (9). To improve it, we utilized emotion prediction with the meta-learning of our previous work (23) and newly investigated the relationship between the amount of training data and the performance of models trained by meta-learning. We also adopted the iso principle as a new music generation method and evaluated a new emotion induction system that applied meta-learning and the iso principle.

2. EEG EMOTION PREDICTION WITH META-LEARNING

In Section 2, we train a highly accurate emotion prediction model with a small amount of a participant's EEG data while listening to music. First, we describe the EEG dataset and the features and the model structure for training models. We then introduce our proposed method using meta-learning and two baseline methods and emphasize its effectiveness by comparing the performances of the three methods.

2.1. Dataset

In our previous work, we created a dataset containing EEG data and subjective evaluations of emotions felt by participants while they listened to music (9). The experiment was approved by the Ethics Committee of the Nara Institute of Science and Technology. Its participants were 10 males and 10 females. The music was created using a music generator designed based on a previous study (6) that made music that induced emotions similar to its valence and arousal inputs. We created 41 pieces of music by inputting 41 various emotions into the music generator. The detailed input values are described in our previous work (9). The sample music can be heard here: https://sites.google.com/view/music-generator. We used a 3-s EEG before listening to the music for a baseline correction based on a previous study (24). The participants silently gazed at a cross mark in the center of the monitor for 5 s for the baseline correction because the initial silent state may contain body movement noise. They then listened to a 20 s piece of music while continuing to gaze at the cross mark. In studies using music to induce emotion, using 30–60 s of music is appropriate (25). However, we tried to record EEG using a variety of music. To incorporate the burden on the participants who wore the electroencephalograph, the music was set to 20 s, referring to previous studies (6, 26). After listening to the music, they evaluated the valence they felt using SAM on a 9-point scale between 0 and 1 and then evaluated their arousal in the same manner. This procedure was repeated for all 41 pieces of music. The EEG data were recorded using a Quick-30 headset manufactured by CGX.

2.2. Features and Model Structure

For each piece of music, we recorded 5 s of EEG data before they listened to the music and 20 s while they listened. We used the last 3 s before listening for a baseline correction and divided these 23 s of EEG data into 1 s pieces without overlap. Then we used band-pass filters and divided them into five frequency bands: theta (4–7 Hz), alpha (8–13 Hz), low beta (14–21 Hz), high beta (22–29 Hz), and gamma (30–45 Hz). We calculated the logarithms of the variances of the EEG waveforms as features and subtracted the average feature values of the three samples before listening to the music from each feature of the 20 samples for baseline correction. Although Quick-30 provides 29 EEG channels, emotion prediction using a selection of 14 EEG channels provided higher performance in our previous work (9). We also used 14 EEG channels in this paper and calculated 20 samples of features with a total of 70 dimensions per piece of music. The features calculated as described above were mapped to matrices, shown in the upper left corner of **Figure 1**. The matrices took into account the position of the EEG channels and the characteristics of the frequency bands. The grid is 6 × 6 × five matrices. The areas without electrodes are embedded with zeros.

We used a CNN for the emotion prediction using EEG. The structure is shown in **Table 1** and at the top of **Figure 1**. We used an SGD optimizer.

2.3. Emotion Prediction Methods

We compared the following three methods for predicting emotions using a small amount of EEG data while listening to music:

TABLE 1 | Structures of CNN and neural network: Conv is convolutional layer, BN is batch normalization layer, FC is fully connected layer, and Drop is drop-out layer.

Model	Layer	Kernel	Channels	Stride	Drop-out rate
CNN	Conv+BN+ReLU	2×2	8	1	–
	Conv+BN+ReLU	2×2	8	1	–
	Conv+BN+ReLU	2×2	8	1	–
	FC	–	2	–	–
Neural network	FC+ReLU+Drop	–	8	–	0.2
	FC	–	2	–	–

Algorithm 1 : MAML for emotion prediction using EEG data.

```
Require: $p(\mathcal{T})$: distribution over tasks
Require: $\alpha, \beta$: learning rate
Randomly initialize $\theta$
Sample training tasks $\mathcal{T}_i \sim p(\mathcal{T})$
for each iteration do
    for each $\mathcal{T}_i$ do
        Select data of 20 pieces of music $\mathcal{D}_i = \{x, y\}$ from $\mathcal{T}_i$
        Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}i}(f_\theta)$ using $\mathcal{D}_i$ and $\mathcal{L}_{\mathcal{T}i}$
        Update parameters: $\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}i}(f_\theta)$
        Select data of about 21 pieces of music $\mathcal{D}_i^{'} = \{x^{'}, y^{'}\}$ from $\mathcal{T}_i$
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}i}(f_{\theta_i^{'}})$ using each $\mathcal{D}_i^{'}$ and $\mathcal{L}_{\mathcal{T}i}$
end for
```

- Method A: multiple participants' EEG with MAML;
- Method B: multiple participants' EEG without MAML;
- Method C: a single participant's EEG.

We set one participant as a target. Methods A and B were trained by the pre-training models with/without MAML using the EEG data of multiple participants. The pre-trained models were trained without the target participant's EEG data. Then the pre-training models were fine-tuned by the target participant's EEG data. Method C was trained with just the target participant's EEG data.

2.3.1. Method A: Multiple Participants' EEG With MAML

We first describe method A, which is our proposed scheme. We used **Algorithm 1**, and the training procedure is shown at the top of **Figure 2**.

For pre-training, we randomly extracted 10 participants' data from our dataset. We considered one participant's data as one task and selected the data of 20 pieces of music from each task i. We set EEG data x and labels y of the emotions felt by the participant as support set $\mathcal{D}_i = \{x, y\}$ and EEG data $x^{'}$ and labels $y^{'}$ of the remaining 21 pieces of music as query set $\mathcal{D}_i^{'} = \{x^{'}, y^{'}\}$. We first used initialized model parameters θ and updated them using the support set in each task. These updated parameters were evaluated with the query set, and the loss was calculated in each task. After all the tasks were computed, model parameters θ were updated to minimize the loss for all of them. This process was repeated. The hyperparameter sets were $\alpha \in \{10^{-1}, 10^{-2}\}$, and $\beta = 10^{-1}$. The number of hyperparameters was small because we needed to train 20 pre-trained models for each target participant to reduce the computation time. We used all of the data from the remaining nine participants in the dataset and set them as the validation data. The hyperparameters were determined using the validation data. The model was trained until the validation loss did not decrease for five consecutive iterations.

We fine-tuned our pre-trained model using the target participant's data. To reduce the preparation time of using the emotion induction system, a small amount of data must be collected from the participants before the emotion induction. Therefore, we examined how much to reduce the amount of data for fine-tuning. We prepared four different kinds of training data to investigate the relationship between the amount of training data and the model performance. We created the music used for the training data by inputting the following emotions into the music generator: five pieces of music ({val,aro}={0,0}; {0,1}; {0.5,0.5}; {1,0}; {1,1}), nine pieces of music (five pieces of music + {val,aro}={0,0.5}; {0.5,0}; {0.5,1}; {1,0.5}), 13 pieces of music (nine pieces of music + {val,aro}={0.25,0.25}; {0.25,0.75}; {0.75,0.25}; {0.75,0.75}), and 25 pieces of music (13 pieces of music + {val,aro}={0,0.25}; {0,0.75}; {0.25,0}; {0.25,0.5};{0.25,1};{0.5,0.25};{0.5,0.75}; {0.75,0};{0.75,0.5};{0.75,1};{1,0.25};{1,0.75}). We selected these music generator inputs to be taken uniformly on the valence and arousal coordinates. The learning rate was set to 0.1 and the iterations to 13. These parameters were determined based on our previous work (23). The fine-tuned model was evaluated using 16 pieces of music, which were not included in the training data.

2.3.2. Method B: Multiple Participants' EEG Without MAML

Next we describe method B, which is a baseline method whose training procedure is shown in the middle of **Figure 2**.

For pre-training, we randomly extracted the data of 10 participants from our dataset. Multiple participants were regarded as one large amount of data. Initial parameters θ were trained with a batch size of 1,024 using the data. The hyperparameter sets were *learning rate* $\in \{10^{-1}, 10^{-2}\}$. We used all of the data from the remaining nine participants in the dataset and set them as the validation data from which the hyperparameters were determined. The model was trained until the validation loss did not decrease for five consecutive iterations. The pre-trained model was fine-tuned using the target participant's data. We prepared four different kinds of training data to investigate the relationship between the number of training data and the performance as well as the proposed method. The learning rate was set to 0.1 and the iterations to 10. We evaluated the fine-tuned model using the test data as well as the proposed method.

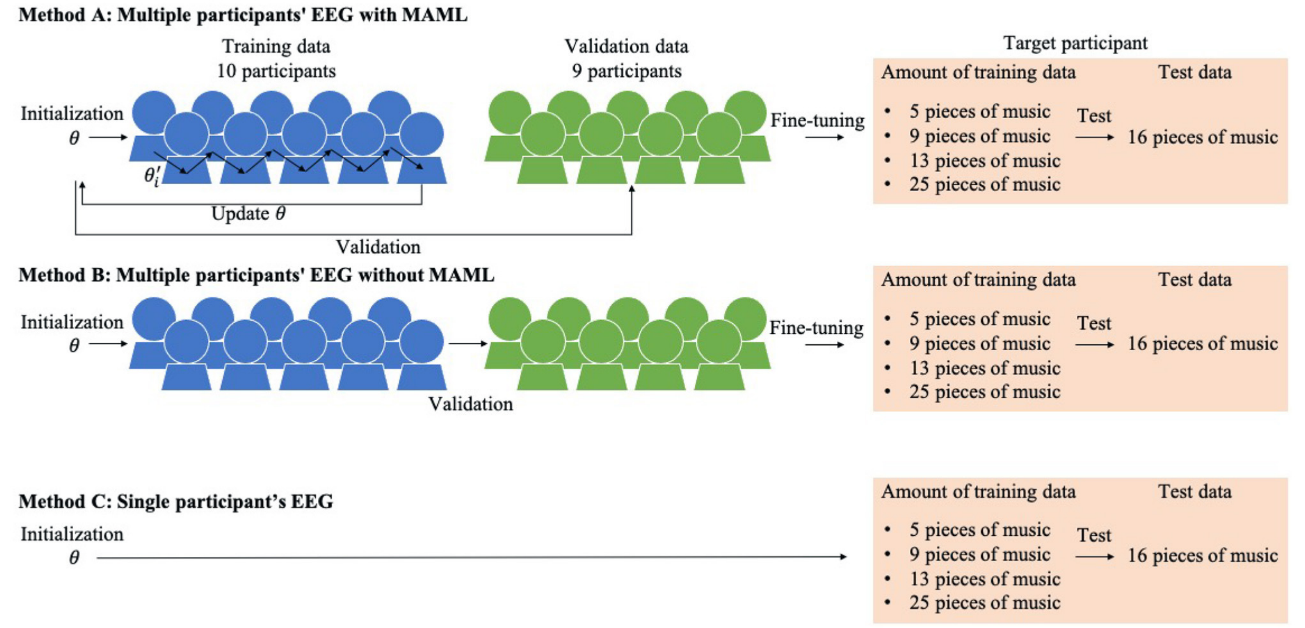


FIGURE 2 | Three emotion prediction methods using a single target participant's small amount of EEG data.

2.3.3. Method C: Single Participant's EEG

Next we describe method C, which is a baseline method whose training procedure is shown at the bottom of **Figure 2**. It has no pre-training. The model was trained from initial parameters θ by the same procedure as in the fine-tuning of the proposed method and the other baseline method. We prepared four different kinds of training data to investigate the relationship between the amount of training data and the performance as well as the other methods. The learning rate was set to 0.1 and the iterations to 25. The fine-tuned model was evaluated using the test data like the other methods.

2.4. Comparison of Three Methods for Predicting Emotions Using EEG

We trained 20 models with different target participants in each method. The emotion prediction results are shown in **Table 2** and **Figure 3**, which show the RMSE between the label values of the dataset and the values predicted by the CNN. We found a significant difference among the three methods using the same amount of data in both valence and arousal in the Friedman test ($p < 0.05$). We used Wilcoxon signed-rank tests with Bonferroni correction to compare the three methods. In the valence results, there was a significant difference between methods A and B and methods A and C with any amount of data ($p < 0.016$). However, for methods B and C, there was a significant difference when nine pieces of music were used ($p < 0.016$). In the arousal results, there was a significant difference between methods A and B, between methods A and C, and between methods B and C with any amount of data ($p < 0.016$). Proposed method A had a significantly lower RMSE than the two baseline methods for both valence and arousal. We also found a significant difference in the RMSE trained by four different training data amounts of proposed method A of both valence and arousal in the Friedman test ($p < 0.05$). The results indicated that the performance of the emotion prediction depended on the amount of training data.

These results showed that in the case of arousal prediction with a small amount of EEG data while listening to music, methods A and B had lower RMSE than method C. Moreover, method A had a lower RMSE than method C for the prediction valence. In the case of using multiple participants' EEG data, method A had a lower RMSE than method B in the predictions of both the valence and arousal. Furthermore, the RMSE was lower when the amount of training data was larger in proposed method A, indicating that the amount of training data is important for highly accurate emotion prediction.

2.5. Predicting Emotions Using EEG and Music Generator Inputs

Our previous work argued that a neural network using emotions predicted from EEG and a music generator's inputs can predict participants' emotions with high performance (9). Since the music generator makes music to induce emotions that resemble its inputs, we considered its inputs the predicted emotions felt by the participants when they listened to music. We also used an emotion prediction neural network in this paper to stabilize the predictions by using two types of information as its inputs: the emotion predicted by the CNN with MAML using EEG and the music generator's inputs. We compared the prediction performance of the following two models:

TABLE 2 | Participants' mean and standard deviation of RMSEs of felt and predicted emotions using EEG data: Bold indicates RMSE of proposed method with a significant difference from baseline methods.

Method	5 pieces		9 pieces		13 pieces		25 pieces	
	Val	Aro	Val	Aro	Val	Aro	Val	Aro
A	0.298	0.298	0.275	0.290	0.262	0.285	0.256	0.274
	(0.121)	(0.071)	(0.101)	(0.071)	(0.096)	(0.066)	(0.098)	(0.058)
B	0.347	0.328	0.325	0.323	0.318	0.320	0.312	0.308
	(0.122)	(0.082)	(0.103)	(0.080)	(0.099)	(0.077)	(0.101)	(0.067)
C	0.378	0.391	0.355	0.366	0.338	0.354	0.331	0.344
	(0.080)	(0.079)	(0.084)	(0.068)	(0.071)	(0.070)	(0.070)	(0.069)

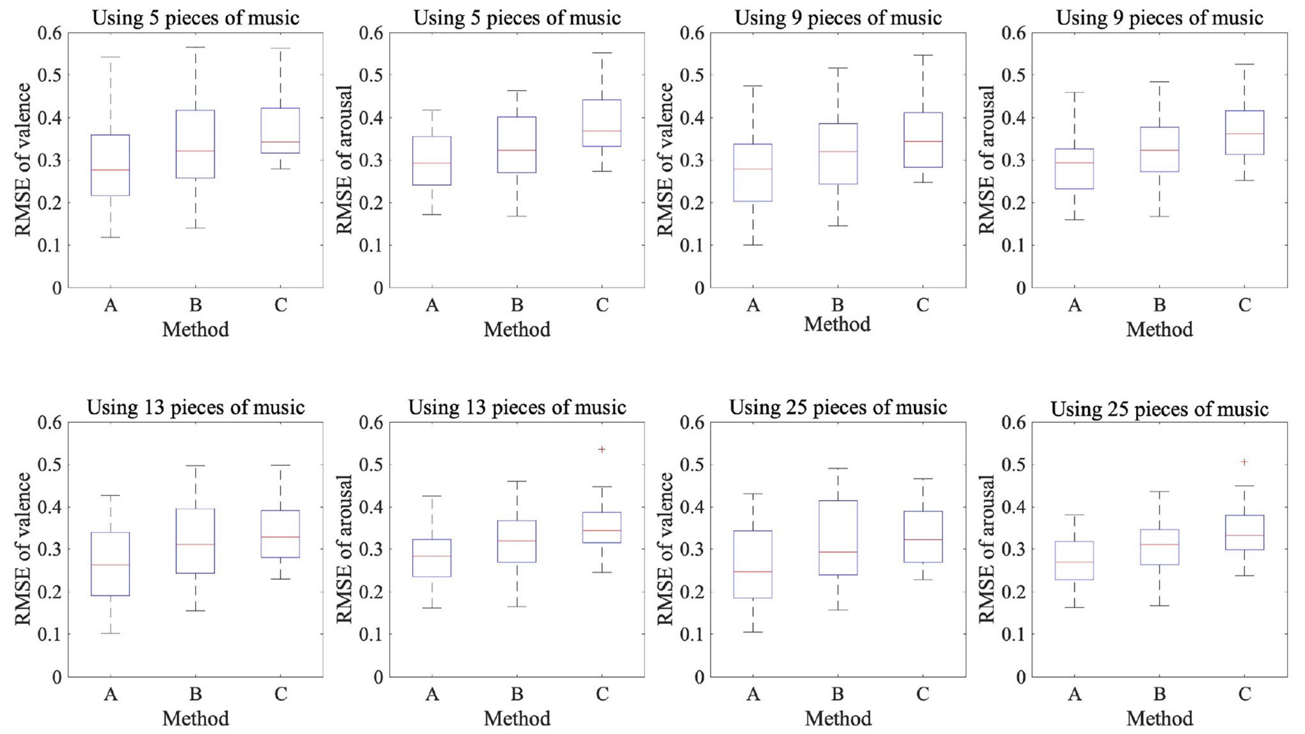


FIGURE 3 | Box plots of 20 participants' RMSEs of felt and predicted emotions using EEG data.

- Model A: CNN
- Model B: CNN + neural network.

The neural network's structure is shown in **Table 1** and in the right part of **Figure 1**. We used an SGD optimizer, fine-tuned the CNN pre-trained by MAML, and trained the neural network using the target participant's data. CNN's fine-tuning method was identical as in Section 2.3.1. The learning rate was set to 0.1 and the iterations to 100 for training the neural network.

The emotion prediction results are described in **Table 3** and **Figure 4**, which show the RMSE between the label values of the dataset and the predicted values using model B or the music generator's inputs of each target participant. Compared with **Table 2**, we found a significant difference among the following three predictions using the same amount of data in both the valence and the arousal in the Friedman test ($p < 0.05$): using model A, model B, and the music generator's inputs. We used Wilcoxon signed-rank tests with Bonferroni correction to compare the three predictions. In the valence results, there was a significant difference between models A and B when any amount of data was used ($p < 0.016$). For model B and using the music generator's inputs, there was a significant difference when 13 or more pieces of music were used ($p < 0.016$). In the arousal results, there was a significant difference between models A and B when any amount of data was used ($p < 0.016$). For model B and using the music generator's inputs, there was a significant difference

TABLE 3 | Participants' mean and standard deviation of RMSEs of felt and predicted emotions using EEG data and music generator's inputs: Music gen. indicates emotion prediction using music generator's inputs.

Model B: CNN + neural network								Music gen.	
5 pieces		9 pieces		13 pieces		25 pieces			
Val	Aro	Val	Aro	Val	Aro	Val	Aro	Val	Aro
0.202	0.204	0.194	**0.196**	**0.181**	**0.192**	**0.171**	**0.184**	0.251	0.258
(0.088)	(0.044)	(0.093)	(0.043)	(0.078)	(0.041)	(0.075)	(0.039)	(0.084)	(0.086)

Bold indicates RMSE with a significant difference from both predictions using model A and music generator's inputs.

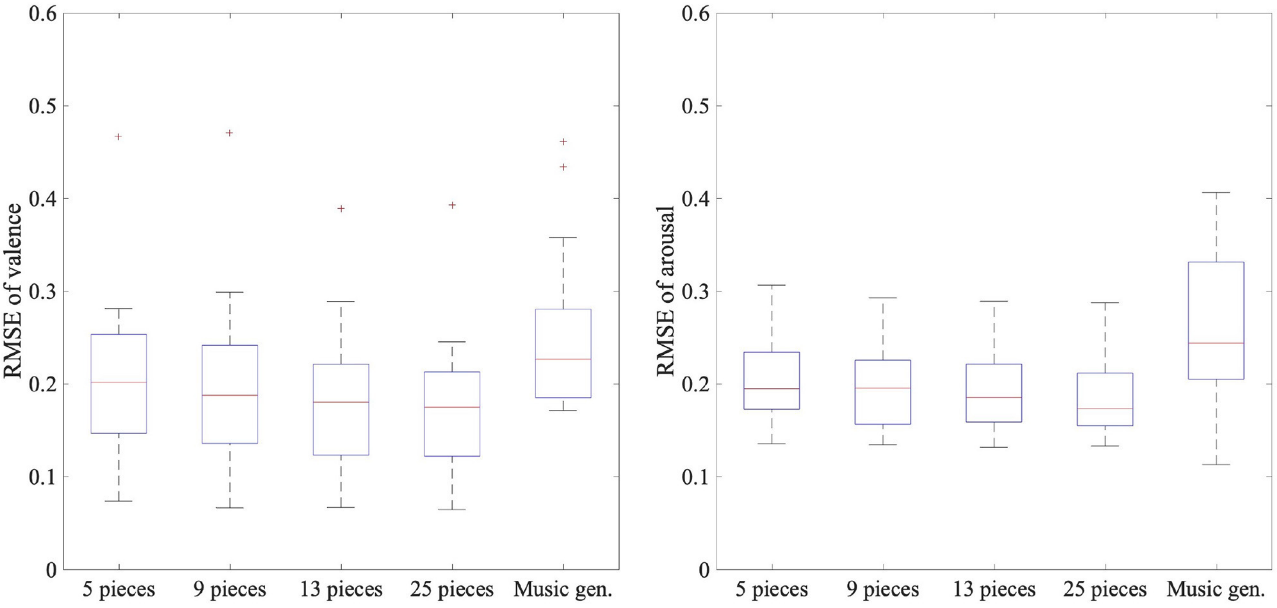


FIGURE 4 | Box plots of 20 participants' RMSEs of felt and predicted emotions using EEG data and music generator's inputs: Music gen. indicates emotion prediction using music generator's inputs.

when nine or more pieces of music were used ($p < 0.016$). We also found a significant difference in the RMSE trained by four different amounts of training data of model B in both the valence and the arousal in the Friedman test ($p < 0.05$).

These results showed that for emotion prediction with a small amount of EEG while listening to music, model B had lower RMSE than model A. Furthermore, model B had lower RMSE than using the music generator's inputs in the predictions of both the valence and arousal when 13 or more pieces of music were used. The RMSE values were lower when the amount of training data was larger in model B. This result indicates that the amount of training data is important for highly accurate emotion prediction, as in the results of **Table 2**.

In this section, we experimentally used a small amount of EEG data while our participants listened to music to train the emotion prediction models. The results showed that MAML was effective for emotion prediction. We also developed a neural network using the emotions predicted by a CNN trained by MAML and the music generator's inputs. A neural network using both the EEG data and the music generator's inputs improved the performance of the emotion prediction. In the next section, we construct and validate an emotion induction system using the CNN trained by MAML and a neural network as an emotion prediction model.

3. EMOTION INDUCTION USING MODELS TRAINED BY META-LEARNING

In Section 3, we construct an emotion induction system using a CNN trained by MAML and a neural network for emotion prediction (**Figure 5**). Since the music generation method used in the conventional system ignored the emotions of the participants before they listened to music, we developed a music generation method based on the iso principle. Our system generates music that resembles a participant's emotion before he listened to music and gradually generated music that was close to the target emotion. We investigated whether a

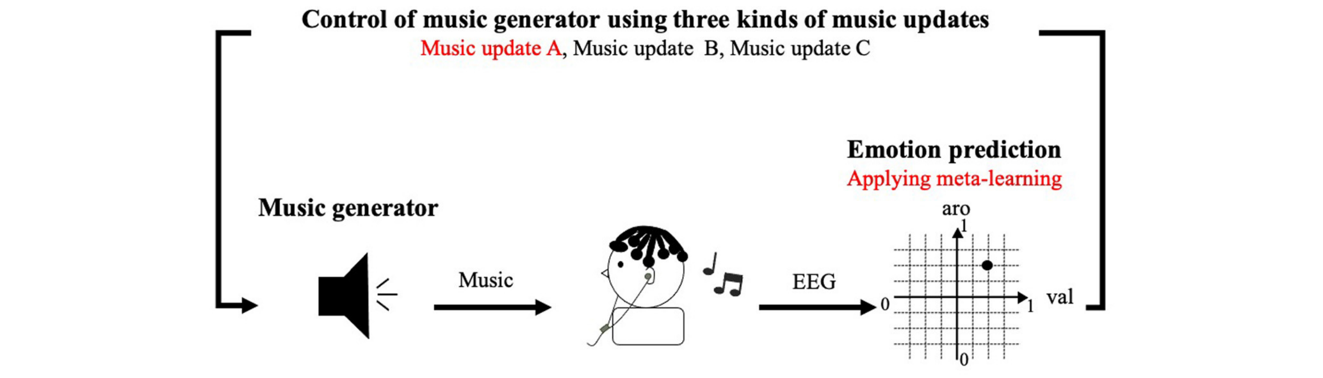

Control of music generator using three kinds of music updates
Music update A, Music update B, Music update C
Emotion prediction
Applying meta-learning
Music generator
Music
EEG
aro
1
0
1
val
0

FIGURE 5 | Emotion induction system using meta-learning: Red text is newly implemented methods in this paper.

system using meta-learning and the iso principle effectively induced emotion.

3.1. Utilization of Emotion Induction System

We used our system into which we embedded an EEG-based emotion prediction model trained by MAML to generate music in real-time. We next present information on the data collection during emotion induction.

3.1.1. Participants

Ten healthy people (age: 26.6 years; eight males, two females) participated in this experiment, which was approved by the ETHICS Committee of the Nara Institute of Science and Technology. They did not participate in the previous experiments described in Section 2. We used Section 2's dataset to train a pre-training model for the emotion prediction. If the same participant's data were used for pre-training and fine-tuning, we believe that the emotion prediction performance might be distorted by using the same participant's data for pre-training and fine-tuning. For this reason, we carefully recruited these participants.

3.1.2. Target Emotions

We set the following five types of target emotions to be induced in the participants: {val,aro} = {0.125,0.125}; {0.125,0.875}; {0.5,0.5}; {0.875,0.125}; {0.875,0.875}. Although in our previous work we set nine target emotions (9), here we reduced them to five due to experimental time limitations. These five emotions were taken from the nine target emotions of our previous work.

3.1.3. Pre-training Model

In Section 2, we showed that training a CNN with MAML predicted emotions best when using a small amount of EEG data. Therefore, we used MAML to train the pre-training CNN with EEG data. We used our dataset with 10 participants for the training data and 10 for the validation data and tuned the learning rate and the iterations. We fine-tuned the pre-training model with the data of a target participant who joined the experiment in Section 3. The neural network's effectiveness is also shown in Section 2 using the emotions predicted by the CNN and the music generator's inputs. We only trained the neural network with the target participant's data.

3.1.4. Experimental Protocol

At the experiment's beginning, the participants wore earphones at a desk with a monitor and listened to five 15 s samples with the following input values to the music generator: {val,aro}={0,0};{0,1};{0.5,0.5};{1,0};{1,1}.

Then we conducted a practice session. In this experiment, we trained the models for predicting the participants' emotions using a pre-training model in task 1 and induced emotions by generating music in a system embedded with the models in task 2. The details of each experiment are shown in **Figure 6**. Our participants practiced each task once to understand how to perform both tasks. First, we introduce task 1, which trained the models for predicting the participants' emotions. They silently gazed at a cross mark in the center of the monitor for 5 s and then listened to each 20 s music sample while continuing to gaze at the cross mark. After listening to the music, they separately evaluated their emotions using SAM on a 9-point scale from 0 to 1 for valence and arousal. They practiced the experiment with one of two pieces of music: {val,aro}={0.125,0.25} or {0.875,0.75}. Next we introduce task 2, which is the emotion induction procedure of music generation in the system. Before listening to the music, our participants separately evaluated their emotions using SAM on a continuous value from 0 to 1 for valence and arousal. They again silently gazed at the cross mark for 10 s and listened to each 20 s music sample that has 20 measures while continuing to gaze at the cross mark. After listening, they evaluated their emotions using SAM; then they took a 10 s break. They practiced the experiment with one of two pieces of music: {val,aro}= {0.875,0.75} or {0.125,0.25}.

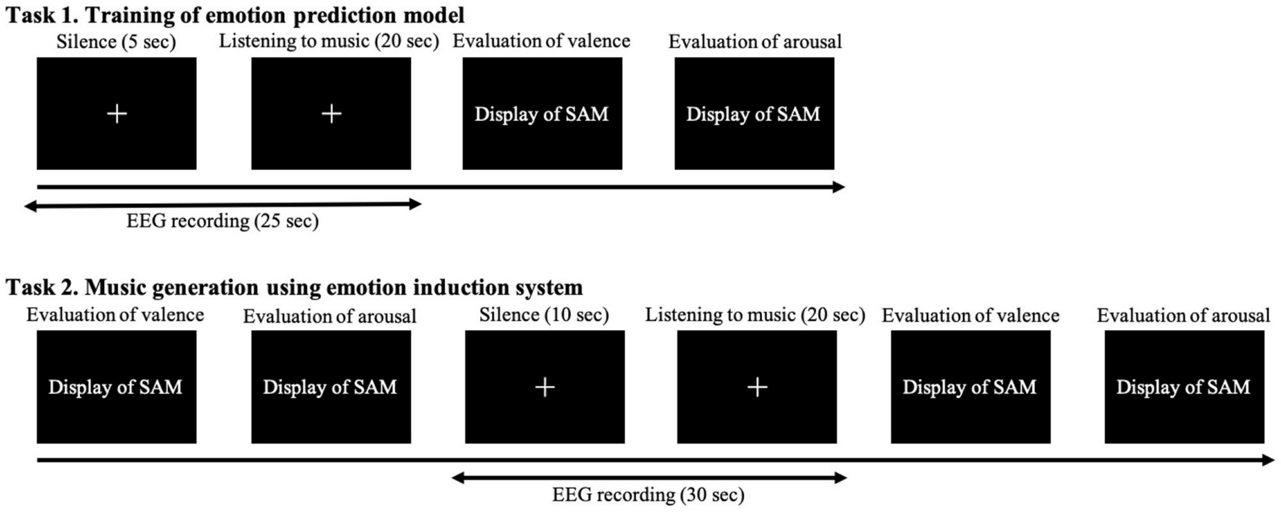


FIGURE 6 | Experimental protocol.

After the practice, the participants put on a Quick-30 headset manufactured by CGX. We repeated the same procedures from the practice session to record the EEG data and the subjective evaluations of their experienced emotions while they listened. Section 2 showed how the emotion prediction's performance improved with more training data. We used the 13 pieces of music to record the EEG data and the subjective evaluations so that the participants continued to wear the electroencephalograph for <30 min. Next we fine-tuned the pre-training model using the recorded data. The preprocessing method is the same as described in Section 2. The learning rate was set to 0.1 and the iterations to 13 for fine-tuning model A's CNN. The learning rate was set to 0.1 and the iterations to 100 for training model B's neural network. Then we conducted emotion induction by music generation for the system embedded with the model in task 2. The participants listened to 15 pieces of music using three different music generation methods. Each method generated music that was intended to invoke five target emotions. EEG data from 2 to 5 s after the onset of silence were used as a baseline correction. The emotion before listening to music was predicted using the EEG data from 5 to 6 s after the onset of silence just using model A. Emotions while listening to music were predicted once every four measures using model B. For this prediction, we used a 1 s EEG after the beginning of the first measure in four measures. The EEG's sampling frequency in the whole experiment was 100 Hz, and the tools used in the experiment included MATLAB (2021b), Lab Streaming Layer, Psychtoolbox (8, 27, 28), Cakewalk, and LoopBe1.

3.2. Music Update Methods

We applied the following three methods for the music updates using **Algorithm 2**:

- Music update A: music updates with the iso principle;
- Music update B: music updates without the iso principle;
- Music update C: fixed music without participants' emotions.

Algorithm 2 : Update music generator's inputs.

```
Record 1 s EEG during the silent state
Predict emotion before listening to music using EEG
if Music update A then
    Set a music generator's inputs as a participant's emotion before listening to music
else if Music update B or C then
    Set a music generator's inputs as a target emotion
end if
for each update do
    Start generating music using the music generator's inputs
    Record a 1 s EEG
    Predict the current emotion using EEG
    if Music update A then
        Update the music generator's inputs using formulas (2) and (4)
    else if Music update B then
        Update the music generator's inputs using formulas (5) and (6)
    else if Music update C then
        Update the music generator's inputs using formulas (7) and (8)
    end if
end for
```

3.2.1. Music Update A: Music Update With Iso Principle

Neither method from our previous work took into account the participants' emotions before they listened to music (9). However, the iso principle showed that using music that is close to the participant's emotion at the beginning and gradually changing it to induce the target emotion effectively induces emotion. In this method, the music generator's inputs were changed based on the iso principle once every four measures using participant's emotion and determined by the following formulas:

$$mid\_target_{\mathrm{val}}(s+1) = \begin{cases} pred_{\mathrm{val}}(s) & \textit{if } \text{s=0}, \\ pred_{\mathrm{val}}(0) + s * (target_{\mathrm{val}} - pred_{\mathrm{val}}(0))/(s_{max} - 1) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (1)$$

$$input_{\mathrm{val}}(s+1) = \begin{cases} mid\_target_{\mathrm{val}}(s+1) & \textit{if } \text{s=0}, \\ mid\_target_{\mathrm{val}}(s+1) + 0.5 * (mid\_target_{\mathrm{val}}(s) - pred_{\mathrm{val}}(s)) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (2)$$

$$mid\_target_{\mathrm{aro}}(s+1) = \begin{cases} pred_{\mathrm{aro}}(s) & \textit{if } \text{s=0}, \\ pred_{\mathrm{aro}}(0) + s * (target_{\mathrm{aro}} - pred_{\mathrm{aro}}(0))/(s_{max} - 1) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (3)$$

$$input_{\mathrm{aro}}(s+1) = \begin{cases} mid\_target_{\mathrm{aro}}(s+1) & \textit{if } \text{s=0}, \\ mid\_target_{\mathrm{aro}}(s+1) + 0.5 * (mid\_target_{\mathrm{aro}}(s) - pred_{\mathrm{aro}}(s)) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (4)$$

In the formulas, s represents the number of times the inputs are updated, $s = 0$ denotes the period before the music generator starts making music, and $s = 1$ denotes when the music generator starts making music. Updates were made up to $s = 5$. s_{max} represents the number of times the music was updated, and $s_{max} = 5$. *input* represents the input emotion to the music generator, *target* represents the target emotion in the induction, *mid_target* represents the intermediate target emotion determined by the number of times the music was updated, and *pred* represents the emotion predicted from the EEG while listening to music. First, the system predicts the participant's emotion before listening to the music using only model A. The difference between the target and predicted emotions was divided by four, which is the maximum number of times the inputs to the music generator were updated using the participant's emotion; the intermediate target emotion was set for each update. In the first loop, the participant's emotion before listening was input directly to the music generator. In the next loop, the system added half of the difference between the intermediate target emotion and the participant's emotion predicted by model B to the next intermediate target emotion. We used a half value because the music generator's inputs were between 0 and 1 for both the valence and arousal. Inputs outside the range were set to 0 or 1. If the difference value is large, the music generator will continue to receive a constant input, such as 0 or 1, and the music will not change. For these reasons, half of this difference was added. In this way, the system generated music that gradually induced emotions while taking into account how the participants were feeling. We show a conceptual scheme of the music generator's control in the yellow dotted line (**Figure 7**).

3.2.2. Music Update B: Music Update Without Iso Principle

In this method, the system first created music by inputting the target emotion into the music generator and adjusting the inputs once every four measures using the participant's emotion. The inputs were determined by the following formulas:

$$input_{\mathrm{val}}(s+1) = \begin{cases} target_{\mathrm{val}} & \textit{if } \text{s=0}, \\ input_{\mathrm{val}}(s) + 0.5 * (target_{\mathrm{val}} - pred_{\mathrm{val}}(s)) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (5)$$

$$input_{\mathrm{aro}}(s+1) = \begin{cases} target_{\mathrm{aro}} & \textit{if } \text{s=0}, \\ input_{\mathrm{aro}}(s) + 0.5 * (target_{\mathrm{aro}} - pred_{\mathrm{aro}}(s)) & \textit{if } 0<\text{s}<s_{max}. \end{cases} \quad (6)$$

First, the system predicted the emotion of the participants before they listened, although the predicted emotion was not used for the music generation. In the first loop, the target emotion was input directly to the music generator. In the next loop, the system added half of the difference between the target and the predicted emotions of the participant to the previous inputs of the music generator. In this way, the system generated music that rapidly induced emotions while taking into account how the participants felt. We show a conceptual scheme of the music generator's control in the red dotted line in **Figure 7**.

3.2.3. Music Update C: Fixed Music Without Participants' Emotions

In this method, the system kept inputting the target emotion to the music generator. The inputs were determined by the following formulas:

$$input_{\mathrm{val}}(s) = target_{\mathrm{val}} \;\; if \;\; 0 < s < s_{max}. \quad (7)$$

$$input_{\mathrm{aro}}(s) = target_{\mathrm{aro}} \;\; if \;\; 0 < s < s_{max}. \quad (8)$$

The system predicts the emotions of the participants before they listened to the music and while they listened to it, although the predicted emotions were not used for music generation. We show a conceptual scheme of the control of the music generator in the blue line in **Figure 7**.

3.3. Evaluation of Emotion Induction System

We fine-tuned the emotion prediction model of the emotion induction system for each participant in this experiment described in Section 3. The model's performance is important because two music generation methods used the emotions predicted by it. We first confirmed the performance of the trained emotion prediction models of the 10 participants. The emotion prediction results are shown in **Table 4**, which shows the RMSE between the label values evaluated by the participants of their emotions and the predicted values by models A or B before/after listening to all the music. As a reference of the conventional system, the following are the means of the RMSE of the emotion predictions after listening to music with model B for all the participants: valence: 0.201 and arousal: 0.180. The conditions of the conventional system and the current system are different: the number of participants and target emotions, the structure of the emotion prediction model, the emotion evaluation method, and the length of silence before listening to the music. Therefore, comparing the conventional and current systems is impossible. However, from the conventional system's results as a reference, no large difference seems to exist in the RMSE of emotion prediction.

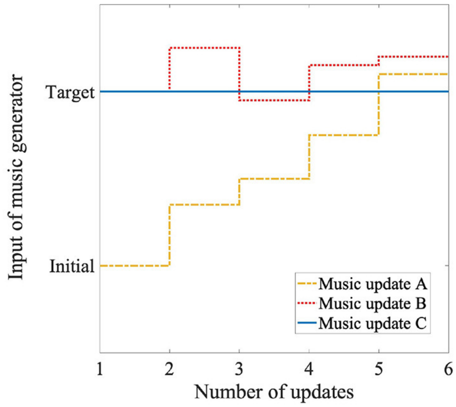


FIGURE 7 | Conceptual scheme of control of music generator using three methods: Target denotes target emotion for emotion induction. Initial denotes participant's emotion before listening to music.

We also investigated the effect of emotion induction by the system. We evaluated the emotion induction performance by calculating the distance between the target emotion and the final predicted emotion by model B using following the formula:

$$distance = \sqrt{(target_{\mathrm{val}} - pred_{\mathrm{val}}(s_{max}))^2 + (target_{\mathrm{aro}} - pred_{\mathrm{aro}}(s_{max}))^2}. \quad (9)$$

The calculated means of the distances of the five types of emotional induction are shown in **Table 5**. In the conventional system of our previous work, the following are the means of the distances for all the participants: music update B: 0.248 and music update C: 0.296. The results showed that both the current and conventional systems effectively induced emotions by taking into account the participants' emotions.

We not only compared the current system with the conventional one but also the performances of the three methods

TABLE 4 | RMSE of felt and predicted emotions before or after listening to music in current system: Bold indicates performance of CNN and neural network used by system to generate music.

	Before listening to music		After listening to music			
Par.	Model A		Model A		Model B	
	Val	Aro	Val	Aro	Val	Aro
1	0.183	0.117	0.159	0.162	0.144	0.193
2	0.170	0.317	0.259	0.198	0.126	0.155
3	0.117	0.203	0.297	0.329	0.265	0.311
4	0.163	0.196	0.221	0.239	0.168	0.164
5	0.301	0.199	0.458	0.379	0.321	0.191
6	0.200	0.183	0.233	0.163	0.132	0.163
7	0.251	0.264	0.490	0.358	0.200	0.225
8	0.317	0.167	0.318	0.322	0.253	0.326
9	0.190	0.248	0.192	0.210	0.152	0.155
10	0.242	0.203	0.332	0.356	0.196	0.200
Mean	0.213	0.210	0.296	0.272	**0.196**	**0.208**
SD	0.063	0.055	0.109	0.086	0.065	0.062

in the current system. We found a significant difference among them in the Friedman test using the distances calculated for all the pieces of music for all the participants ($p < 0.05$). We used Wilcoxon signed-rank tests with Bonferroni correction for comparisons of the three methods. There was a significant difference between music updates A and C and between music updates B and C ($p < 0.016$). From the above results, we conclude that music updates A and B, which generated music according to the participants' emotions, more effectively induced emotions than music update C that didn't generate music according to their emotions. However, we found no significant difference between music updates A and B. We show plots of the music generator's inputs and the emotions predicted from model B in **Figure 8**. This is the result for participant eight; the target emotion is {val,aro} = {0.875,0.125}, and music update A provided more effective emotion induction than the other two methods. The number of updates is zero before listening to music, and the music generator created music from five updates. Music updates B and C suddenly generated music that induced the target emotion, and music update A generated music that gradually induced the target emotion, starting from music close to the participant's emotion before listening to the music. Music update A led to an emotion closer to the target than the other two methods (**Figure 8**).

TABLE 5 | Distance between target and induced emotions: Bold indicates distance with a significant difference from baseline method.

Par.	Music update A	Music update B	Music update C
1	0.483	0.496	0.450
2	0.399	0.371	0.401
3	0.387	0.445	0.348
4	0.318	0.345	0.418
5	0.385	0.353	0.378
6	0.299	0.296	0.393
7	0.234	0.190	0.362
8	0.267	0.284	0.318
9	0.309	0.287	0.340
10	0.224	0.303	0.338
Mean	**0.331**	**0.337**	0.375
SD	0.082	0.087	0.041

4. CONCLUSION AND FUTURE WORKS

Our conventional emotion induction system using music and EEG suffered from two problems. It took a long time to record EEG to train the emotion prediction model, which is a required step for constructing our system. The second problem was that

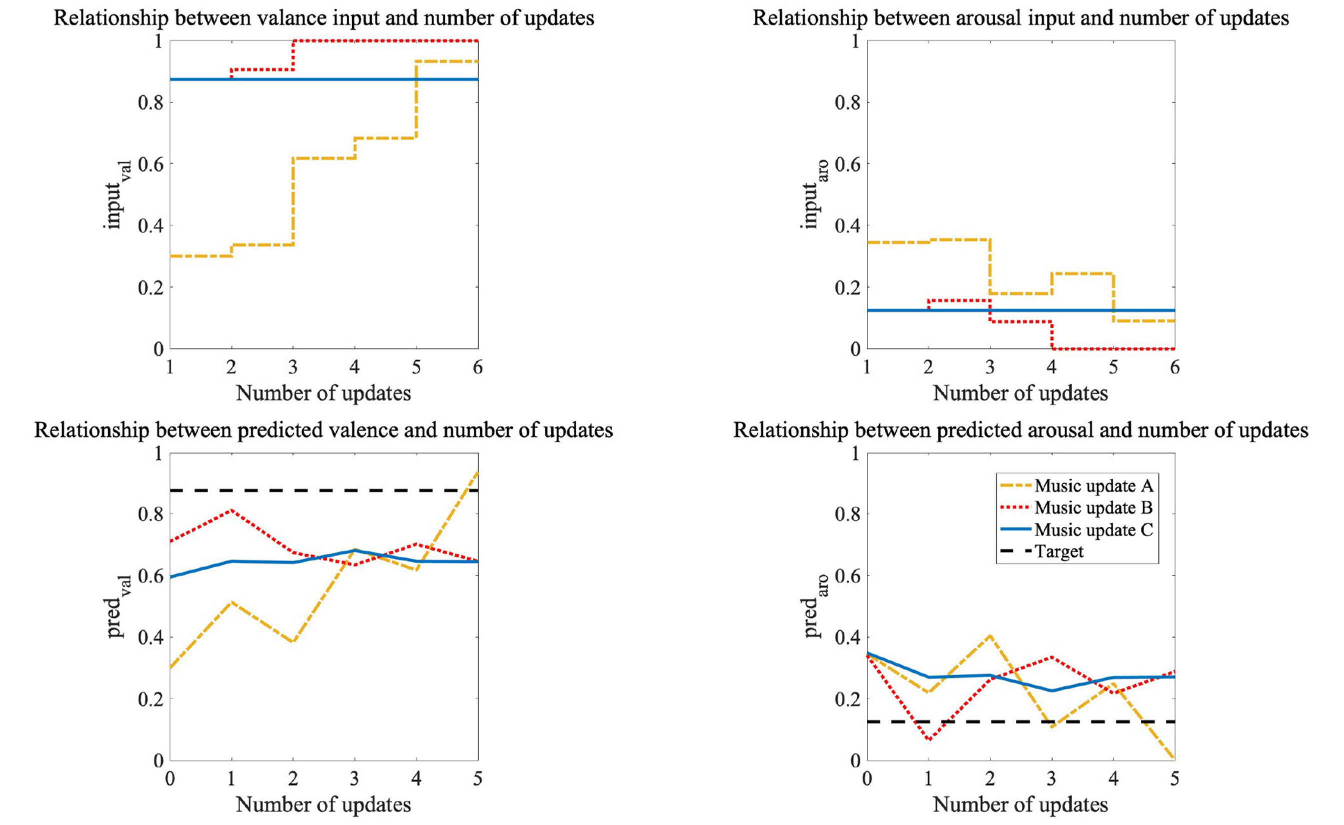


FIGURE 8 | Plots of inputs of music generator and emotions predicted from CNN and neural network in participant eight: Target emotion is {val,aro} = {0.875,0.125}.

the music generator's control method created music without the participants' emotions before they listened to music. We solved these problems by developing a new system that uses meta-learning and the iso principle. To solve the first problem, we proposed a meta-learning method using a small amount of EEG data while listening to music. The proposed method predicted emotions with higher performance than the baseline methods without meta-learning. In addition, the system into which the trained model with meta-learning was embedded effectively induced emotions. Therefore, we conclude that meta-learning reduced the EEG recording time and increased the usability of our emotion induction system.

To solve the second problem, our system induced emotions through music generation using the iso principle. The methods with/without it, which took the participants' emotions into account, more effectively induced emotions than the methods that did not consider them. We found no significant difference between the methods with/without the iso principle. In previous studies on it, emotions opposite to the target emotion were induced in the participants beforehand, and then the participants were led to the target emotion (22). In our experiment, we did not induce emotions opposite to the target emotion before our participants listened to music. We believe that music generation with the iso principle may be more effective than the other two music generation methods when the participants are induced to the target emotion from an opposite emotion. We set the length of the music sample to 20 s. The results are limited in terms of the music duration. We need to consider how many seconds of music to use for more effective emotion induction in the future.

Our future works will investigate two problems. The first is to improve meta-learning for more efficient emotion prediction. Meta-learning has been actively studied in recent years, and improvements are being developed (29, 30). Improvements in meta-learning that address the EEG characteristics will raise the accuracy of emotion prediction. The second problem is the investigation of more diverse music generation methods. We used predefined formulas to control the music generator. In the future, we will develop a method using deep learning to control it based on the participants' characteristics.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the datasets contain some private information. Requests to access the datasets should be directed to KM, miyamoto.kana.mk4@is.naist.jp.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Nara Institute of Science and Technology (reference number: 2019-I-23). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KM, HT, and SN designed and conducted the experiment. KM wrote the manuscript. SN is the project's principal investigator and directs all the research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by JST CREST Grant Number JPMJCR19A5, Japan and the RIKEN Junior Research Associate Program.

REFERENCES

1. Joormann J, Stanton CH. Examining emotion regulation in depression: a review and future directions. *Behav Res Ther*. (2016) 86:35–49. doi: 10.1016/j.brat.2016.07.007
2. Compare A, Zarbo C, Shonin E, Van Gordon W, Marconi C. Emotional regulation and depression: a potential mediator between heart and mind. *Cardiovasc Psychiatry Neurol*. (2014) 2014:324374. doi: 10.1155/2014/324374
3. Santos V, Paes F, Pereira V, Arias-Carrión O, Silva AC, Carta MG, et al. The role of positive emotion and contributions of positive psychology in depression treatment: systematic review. *Clin Pract Epidemiol Ment Health*. (2013) 9:221–37. doi: 10.2174/1745017901309010221
4. Schubert E. Emotion felt by the listener and expressed by the music: literature review and theoretical perspectives. *Front Psychol*. (2013) 4:837. doi: 10.3389/fpsyg.2013.00837
5. Larsen RJ, Ketelaar T. Personality and susceptibility to positive and negative emotional states. *J Pers Soc Psychol*. (1991) 61:132. doi: 10.1037/0022-3514.61.1.132
6. Ehrlich SK, Agres KR, Guan C, Cheng G. A closed-loop, music-based brain-computer interface for emotion mediation. *PLoS ONE*. (2019) 14:e213516. doi: 10.1371/journal.pone.0213516
7. Sourina O, Liu Y, Nguyen MK. Real-time EEG-based emotion recognition for music therapy. *J Multim User Interfaces*. (2012) 5:27–35. doi: 10.1007/s12193-011-0080-6
8. Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. *J Behav Ther Exp Psychiatry*. (1994) 25:49–59. doi: 10.1016/0005-7916(94)90063-9
9. Miyamoto K, Tanaka H, Nakamura S. Online EEG-based emotion prediction and music generation for inducing affective states. *IEICE Trans Inform Syst*. (2022) 105:1050–63. doi: 10.1587/transinf.2021EDP7171
10. Russell JA. A circumplex model of affect. *J Pers Soc Psychol*. (1980) 39:1161. doi: 10.1037/h0077714
11. Wang F, Zhong Sh, Peng J, Jiang J, Liu Y. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In: *International Conference on Multimedia Modeling*. Bangkok: Springer (2018). p. 82–93. doi: 10.1007/978-3-319-73600-6_8
12. Lan Z, Sourina O, Wang L, Scherer R, Müller-Putz GR. Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans Cogn Dev Syst*. (2018) 11:85–94. doi: 10.1109/TCDS.2018.2826840
13. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. (2009) 22:1345–59. doi: 10.1109/TKDE.2009.191
14. Hospedales TM, Antoniou A, Micaelli P, Storkey AJ. Meta-learning in neural networks: a survey. *IEEE Trans Pattern Anal Mach Intell*. (2021) 1. doi: 10.1109/TPAMI.2021.3079209
15. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: *International Conference on Machine Learning*. Sydney, NSW (2017). p. 1126–35.

16. Banluesombatkul N, Ouppaphan P, Leelaarporn P, Lakhan P, Chaitusaney B, Jaimchariya N, et al. Metasleeplearner: a pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE J Biomed Health Inform*. (2020) 25: 1949–63. doi: 10.1109/JBHI.2020.3037693
17. Li D, Ortega P, Wei X, Faisal A. Model-agnostic meta-learning for EEG motor imagery decoding in brain-computer-interfacing. *arXiv preprint arXiv:210308664*. (2021). doi: 10.1109/NER49283.2021.9441077
18. Duan T, Shaikh MA, Chauhan M, Chu J, Srihari RK, Pathak A, et al. Meta learn on constrained transfer learning for low resource cross subject EEG classification. *IEEE Access*. (2020) 8:224791–802. doi: 10.1109/ACCESS.2020.3045225
19. Baumgartner T, Esslen M, Jäncke L. From emotion perception to emotion experience: emotions evoked by pictures and classical music. *Int J Psychophysiol*. (2006) 60:34–43. doi: 10.1016/j.ijpsycho.2005.04.007
20. Altshuler IM. four years'experience with music as a therapeutic agent at Eloise hospital. *Am J Psychiatry*. (1944) 100:792–4. doi: 10.1176/ajp.100.7.792
21. Heiderscheit A, Madson A. Use of the iso principle as a central method in mood management: a music psychotherapy clinical case study. *Mus Ther Perspect*. (2015) 33:45–52. doi: 10.1093/mtp/miu042
22. Starcke K, Mayr J, von Georgi R. Emotion modulation through music after sadness induction-the iso principle in a controlled experimental study. *Int J Environ Res Publ Health*. (2021) 18:12486. doi: 10.3390/ijerph182312486
23. Miyamoto K, Tanaka H, Nakamura S. Meta-learning for emotion prediction from EEG while listening to music. In: *Companion Publication of the 2021 International Conference on Multimodal Interaction.* Montreal, QC (2021). p. 324–8. doi: 10.1145/3461615.3486569
24. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans Affect Comput*. (2011) 3:18–31. doi: 10.1109/T-AFFC.2011.15
25. Ali SO, Peynirciou ZF. Intensity of emotions conveyed and elicited by familiar and unfamiliar music. *Mus Percept*. (2010) 27:177–82. doi: 10.1525/mp.2010.27.3.177
26. Eerola T, Vuoskoski JK. A review of music and emotion studies: approaches, emotion models, and stimuli. *Mus Percept*. (2012) 30:307–40. doi: 10.1525/mp.2012.30.3.307
27. Pelli DG, Vision S. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis*. (1997) 10:437–42. doi: 10.1163/156856897X00366
28. Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. What's new in Psychtoolbox-3? *Perception.* (2007) 36:1–16.
29. Raghu A, Raghu M, Bengio S, Vinyals O. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In: *International Conference on Learning Representations.* (2020). Available online at: https://openreview.net/forum?id=rkgMkCEtPB (accessed February 9, 2022).
30. Oh J, Yoo H, Kim C, Yun SY. BOIL: towards representation change for few-shot learning. In: *International Conference on Learning Representations.* (2021). Available online at: https://openreview.net/forum?id=umIdUL8rMH (accessed February 9, 2022).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fpsyg.2022.878581

Comparison and Analysis of Timbre Fusion for Chinese and Western Musical Instruments

Jingyu Liu[1,2,3,4], *Shuang Wang*[1,2,3,4], *Yanyin Xiang*[4], *Jian Jiang*[5], *Yujian Jiang*[1,2,3,4]* and *Jing Lan*[6]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China, [2] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Communication University of China, Beijing, China, [3] Beijing Key Laboratory of Modern Entertainment Technology, Communication University of China, Beijing, China, [4] School of Information and Communication Engineering, Communication University of China, Beijing, China, [5] China Digital Culture Group Co., Ltd, Beijing, China, [6] Center for Ethnic and Folk Literature and Art Development, Ministry of Culture and Tourism, Beijing, China

OPEN ACCESS

Edited by:
Diego Minciacchi,
University of Florence, Italy

Reviewed by:
Sven-Amin Lembke,
De Montfort University,
United Kingdom
Sushen Rameshpant Gulhane,
D. Y. Patil College of Engineering, India

***Correspondence:**
Yujian Jiang
yjjiang@cuc.edu.cn

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Psychology

Received: *18 February 2022*
Accepted: *31 May 2022*
Published: *07 July 2022*

Citation:
Liu J, Wang S, Xiang Y, Jiang J, Jiang Y and Lan J (2022) Comparison and Analysis of Timbre Fusion for Chinese and Western Musical Instruments.
Front. Psychol. 13:878581.
doi: 10.3389/fpsyg.2022.878581

Timbre fusion is the theoretical basis of instrument acoustics and Chinese and Western orchestral acoustics. Currently, studies on timbre fusion are mainly focused on Western instruments, but there are some studies on the timbre fusion of Chinese instruments. In this paper, the characteristics of timbre fusion for Chinese and Western instruments are explored, focusing on the subjective attributes and objective acoustic parameters, and a series of experiments is carried out. First, a database containing 518 mixed timbre stimuli of Chinese and Western instruments was constructed to provide basic data that are necessary for the subjective and objective analyses of timbre fusion. We designed and conducted a subjective evaluation experiment of timbre perception attributes based on the method of successive categories. The experimental data were processed using statistical approaches, such as variance analysis, multidimensional preference analysis, and correlation analysis, and we studied the influence of the temporal envelopes and instrument types on fusion, segregation, roughness, and pleasantness. In addition, the differences between Chinese and Western instruments were compared based on these four perception attributes. The results show that fusion and segregation are the most important attributes for Chinese instrument timbre, while roughness is the most important attribute for Western instrument timbre. In addition, multiple linear regression, random forest, and multilayer perceptron were used to construct a set of timbre fusion models for Chinese and Western instruments. The results show that these models can better predict the timbre fusion attributes. It was also found that there are some differences between the timbre fusion models for Chinese and Western instruments, which is consistent with the analysis results of subjective experimental data. The contribution of acoustic objective parameters to the fusion model is also discussed.

Keywords: timbre, fusion, auditory perception, acoustic parameters, Chinese and Western instruments, instrument acoustics, cross-cultural

INTRODUCTION

Background

Since the twentieth century, Western music culture has been gradually introduced in China. Western symphony orchestras, with their rich instrument timbre, powerful expressive force, and standardized orchestral arrangements and sound effects, have led to new inspirations and musical forms of Chinese folk music. The "fusion" of all musical instruments is the basic aesthetic principle of Western symphony. In the long-term development of symphonic creation, both harmony and orchestration have formed relatively mature acoustic theory and have been relatively successful with respect to practical experience. "Fusion" refers to a relationship that occurs after the combination of timbre, that is, the combined effect produced by the simultaneous sound of different instruments. From acoustic theory, fusion can be understood as the degree of harmonic integration of musical instruments (Li, 2020). Most instruments in Western symphony orchestras have a high degree of consonance, which means that each harmonic is, basically, an integer multiple of the fundamental frequency, making the overall sound effect is integrated and unified (Wang et al., 2016).

Instruments in Western symphony orchestras have many timbre characteristics that are not prominent, so it is easy to achieve the effect of fusion by producing compound timbre in orchestration, which refers to the timbre composed of two musical instruments whose timbre is very similar (Li, 2020). When these two instruments play the same or octave melody at the same time, it is difficult to distinguish them. For example, the sound of a violin and viola can be considered a "compound timbre." Here, "compound timbre" corresponds to "timbral emergence," which was proposed by Sandell (1995); that is, all sounds are blended and unidentifiable (McAdams, 2019). Due to the differences between Chinese and Western cultural backgrounds, symphony orchestras and Chinese orchestras have different orchestration ideas. For a long time, musicians have explored the diversification of Chinese musical instruments' timbre in practice and philosophy, borrowed compositional techniques from traditional music, and formed their own aesthetic principles. It is difficult to produce the so-called "compound timbre" between Chinese instruments, but a "mixed timbre" can be produced between them. The so-called "mixed timbre," which is both harmonious and independent and both related and separated, refers to the timbre combined and superimposed by different musical instruments (Li, 2020). Here, "mixed timbre" corresponds to "timbral heterogeneity," which was proposed by Sandell (1995). Timbral heterogeneity is a unique timbre characteristic of Chinese orchestral music, where a beautifully mixed sound with both individuality and combination is formed. In fact, this paper states that Western instruments more easily achieve the fusion effect in orchestration, which means that the maximum fusion effect between Western instruments is better than that of Chinese instruments. The fusion mentioned in this manuscript does not refer to the composer's orchestration. In other words, for a Western orchestra, if the composer wants a fusion effect, he or she can achieve it through a combination of existing Western instruments. For a Chinese orchestra, it is difficult for the composer to find a combination of two or more instruments to achieve the effect of fusion. We have previously performed experiments on the timbre contrast between Chinese instruments and Western instruments and found that the timbre of Chinese instruments is, overall, rougher than that of Western instruments. Moreover, the distribution of Chinese instruments is more dispersed in the three-dimensional timbre space, and the timbre similarity is lower (Jiang et al., 2020).

Currently, the orchestration theory and music practice of Chinese orchestras are still in the exploration stage, and the problem of the fusion between instruments in Chinese orchestra still exists. The Chinese orchestra is composed of folk instruments, most of which have evolved from ancient Chinese instruments and have a distinctive timbre. Therefore, the sound of the Chinese orchestra as a whole has the auditory feeling of "sharp, dry, messy, and noisy." In addition, there are some problems in the Chinese orchestra, such as volume imbalance of the vocal part, low degree of integration between timbres, and uncertain composition of the orchestra. Therefore, on the basis of the aesthetic principles of Chinese music and timbre characteristics of Chinese musical instruments, we should further explore the timbre combination rules for musical instruments.

This paper takes musical instrument combinations as the research object to discuss the differences in the timbre fusion of different musical instrument combinations. A comparative analysis of the fusion of timbre combinations in Western symphony and Chinese orchestra is also presented, and references and theory for Chinese orchestra orchestration are provided. Next, the current research status is summarized from three aspects: the definition of fusion, perception experiments of fusion, and subjective and objective parameters that affect fusion.

Definition of Fusion

Currently, there is more than one definition of timbre fusion within the academic circle. For example, McAdams (2019) proposed that the result of combining sounds concurrently in orchestration is a timbral blend, when events fuse together, or timbral heterogeneity, when they remain separate. Concurrent grouping determines how components of sounds are grouped together into musical events, a process referred to in psychology as auditory fusion. In describing sound quality as a whole, the sense of fusion is one of the important attributes that are used to express the degree of acoustic integration between the whole band or chorus, solo instruments or collaborative instruments, and singing or accompaniment.

The concept of fusion may have been first proposed by Stumpf (DeWitt and Crowder, 1987), who proposed the principle of tonal fusion, defining the fusion as the degree to which two simultaneous monophonic tones are perceived acoustically as one sound. He believed that fusion was the basis of tonal consonance (Apel, 2003). Subsequently, DeWitt and Crowder (1987) further developed Stumpf's theory. They performed three experiments and investigated Stumpf's fusion principle of tonal consonance. The results of this experiment showed that fusion may represent the tendency for people to interpret pitch combinations that could represent harmonics, resulting from a single fundamental as timbres rather than as chords.

Timbre emerges from the perceptual fusion of acoustic components into a single auditory event, including the blending of sounds produced by separate instruments in which the illusion of a "virtual" sound source is created (McAdams, 2019). Bregman and Pinker demonstrated the interplay of concurrent fusion and sequential stream formation and conceived a sort of competition between the two auditory organization processes. Therefore, the attack asynchrony and the decomposition of simultaneities into separate auditory streams, whose events are timbrally similar, work together to reduce the degree of perceptual fusion (Bregman and Pinker, 1978). Timbre is a property of fused auditory events.

Perceptual Experiments on Fusion

Concerning the term "fusion" and its different interpretations, we structured perceptual experiments on fusion into (1) its involvement in concurrent groupings, as in spectral fusion, which forms a timbral identity and (2) the special case of instrument combinations, where it has commonly been referred to as the timbral blend.

McAdams proposed the concept of spectral fusion (McAdams, 1982), which belongs to the first category. An important perceptual aspect of the formation of auditory images evoked by acoustic phenomena is the distinguishing of different sound sources. To form images of sound sources, the auditory system must be able to perceptually fuse the concurrent elements that come from the same source and separate the elements that come from different sources. Then, the relationship between spectral fusion auditory sensory cues was further studied (McAdams, 1984). The results showed that the acoustic cues that contribute to the formation and distinction of multiple, simultaneous source images that are investigated include the harmonicity of the frequency content, the coherence or low-frequency frequency modulation, and the stability and/or recognizability of spectral form when coupled with frequency modulation. Shields and Roger (2004) studied the relationship of timbre to dissonance and spectral fusion. In this experiment, listeners rated dissonance and blend levels for a set of dyads involving fourteen interval sizes and twenty-five orchestral combinations. The researchers related dissonance and spectral fusion to the timbre of time-variant steady-state dyads. The experimental results show that interval size and orchestration are significantly influenced by both dissonance and blend ratings.

At approximately the same time, Carterette and Kendall (1989) and Kendall and Carterette (1991) also conducted similar experiments on timbre. Sandell (1989a,b) reported preliminary work on the "blend" of "concurrent timbres" using 15 of Grey's (1975) line-segment approximations of brief real instrument tones. The results of interest demonstrated that a blend is related to the summed distribution of energy in the harmonic series of the two tones, with a less blend correlated with more energy in higher harmonics compared to lower harmonics. Sandell's (1991) doctoral thesis provided a detailed overview of the concept of fusion. This study investigated the acoustical correlates of a blend for 15 natural-sounding orchestral instruments presented in concurrently sounding pairs. Sandell's acoustically based guidelines for a blend, which augment instance-based methods of traditional orchestration teaching, provided underlying abstractions that are helpful for evaluating the blend of arbitrary combinations of instruments. Sandell (1995) also proposed three possible perceptual results of instrument combinations: timbral heterogeneity, timbral augmentation, and timbral emergence. Kendall and Carterette (1993) reported on a series of experiments directed toward questions concerning the timbres of simultaneous orchestral wind instruments. In this study, researchers ascertained the degree of a blend and identifiability of soprano orchestral winds. It was found that the degree of a blend corresponded with the positions of instruments in a two-dimensional similarity space.

Subjective and Objective Parameters Affecting Fusion

Regarding the subjective perception attributes describing the fusion of timbre, different studies have provided representative terms from different perspectives. In the experiment of Bregman and Pinker (1978), compound sounds composed of two pure tones with different frequencies were used as experimental stimuli. Compound sounds are somewhat dissonant and are described as "rough" or "complex." DeWitt and Crowder (1987) further supplemented Stumpf's theory and proposed three pairs of evaluation terms to describe musical intervals: consonance-dissonance, smoothness-roughness, and pleasant-unpleasant. Kim (2018) investigated how musicians perceive and compensate for the interacting effects of timbre, blend and sensory dissonance when tuning and rating harmonic intervals. In this experiment, timbre terminology, such as rough, unpleasant, smooth, and pleasant, was used to describe the timbre perception properties of the trumpet and vibraphone. Sounds that differ acoustically are organized by the auditory system into separate percepts called auditory streams (Bregman and Campbell, 1971). A physical sound source can produce a sequence of successive acoustic events. To examine this phenomenon, Fischer et al. (2021) used naturalistic orchestral excerpts from the symphonic repertoire to examine perceptual segregation.

Beating is an important factor causing roughness. In this experiment, dyads are in pitch unison or octave. These dyads thus exhibit a very low degree of roughness. In our previous pre-experiment, we found a certain negative correlation between roughness and a degree of fusion. Previous research studies have, indeed, shown that dyads in pitch unison are perceived to be more blended than dyads involving non-unison pitches (Kendall and Carterette, 1993; Jingyu, 2013; Lembke et al., 2019). Combining all of these studies, we have chosen four timbre perception attributes, fusion, roughness, segregation, and pleasantness for subjective evaluation experiments.

Researchers have also explored the relationship between fusion and objective acoustic parameters. Fusion is affected by sensory cues, such as whether the acoustic components begin synchronously, whether they are related by a common period, and whether there is coherent frequency and amplitude behavior (McAdams, 1984). The coherent behavior cues are related to the

Gestalt principle of common fate. In other words, sounds that change in a similar manner are likely to have originated from the same source (Bregman, 1994).

The degree of fusion also depends on spectrotemporal relations among the concurrent sounds (Siedenburg et al., 2019). Sandell (1995) demonstrated that sounds blend better when they have similar attack envelopes and spectral centroids, as well as when their composite spectral centroid is lower. This experiment also found that the more similar these parameters are for the two combined sounds, the greater their blend. Tardieu and McAdams (2012) performed two experiments on combinations of pitched impulsive and sustained sounds. They highlighted the audio descriptors, underlying the perception of a blend and the perception of emergent timbre for dyads composed of one impulsive and one sustained sound. In both experiments, the attack time was very important, as it was one of the two most important factors in predicting both a blend and emergent timbre perception. Chon and Huron (2014) proposed the concept of timbre salience. In this paper, they examined the identification of an instrument sound in concurrent unison dyads. As a salient timbre is defined as one that captures listeners' attention easily and tends not to blend well with concurrent sounds (Chon and McAdams, 2012), we can logically expect that a salient timbre will be easily identified.

In addition to the acoustic parameters mentioned above, researchers have proposed other features that describe the fusion of timbre. Rossetti (2016) discussed timbre and sound morphology in live electroacoustic and instrumental music from a compositional standpoint and convergence issues in live electroacoustic music. They proposed that timbre fusion should be addressed based on the concepts of jitter, permeability, and timbre of movement. In addition to the study of timbre fusion for Western symphonies, some researchers have studied the timbre characteristics of African music (Fales and McAdams, 1994). The authors presented the results of perceptual and acoustic investigations of the fusion and "layering" of noise and tone. The results also exemplified the fusion of two extremely different timbres with implications for the blending of instrumental timbres in an orchestral setting.

In addition to global descriptors, such as the spectral centroid, research has been conducted on the role of local descriptors of formant structure (Siedenburg et al., 2019). Goodwin (1980) studied the acoustic parameters of individual voices in choral blends. The phenomenon of a choral blend was investigated by identifying spectral differences between vocal sounds produced in solo singing and in unison ensemble singing to achieve the optimum blend. Reuter (2003) studied the relationship between stream segregation and formant areas. The results are as follows: Alternating timbres with equivalent main formant areas tend to produce one sole, continuous melody in perception. Alternating timbres with non-matching formant areas tend to produce two distinct melodies in perception. Lembke and McAdams (2012, 2015) investigated the acoustical and perceptual factors involved in timbre blending between orchestral wind instruments based on a pitch-invariant acoustical description of wind instruments. A possible perceptual relevance for these formants was tested in their experiments, employing different behavioral tasks. The results showed that the relative frequency location and magnitude differences between formants can be shown to bear a pitch-invariant perceptual relevance to blend for several instruments.

In the context of perceptual blending between orchestral timbres, holistic acoustical descriptions of instrument-specific traits can assist in the selection of suitable instrument combinations (Lembke et al., 2013). Researchers have proposed several parameters, such as spectral maxima or formants, which have been shown to influence timbre blending involving frequency relationships between local spectral features, their prominence as formants, and constraints imposed by the human auditory system. Computational approaches to predict a timbre blend have been proposed that are based on these factors and explain ∼85% of the variance in behavioral timbre-blend data.

In summary, research on timbre fusion has mostly focused on Western instruments, and there is, currently, no study on Chinese instruments. To explore the rules of timbre fusion for Chinese instrument combinations and compare the differences between Chinese and Western instruments, a dataset has been constructed for this study that contains a combination of Chinese and Western instruments. Through the statistical processing of experimental data, the differences between Chinese and Western musical instruments in timbre fusion are analyzed, and the subjective and objective acoustic parameters affecting timbre fusion are also analyzed. In addition, a timbre fusion model for Chinese and Western instruments, which provides basic theory and data support for the orchestration of Chinese and Western instruments, is constructed for this study.

The following sections of this paper are arranged as follows. In Section Methods, the second part, the four-part method, which includes the participants, stimuli, apparatus and procedure, is introduced. In Section Subjective Evaluation Experiment and Data Analysis, the statistical analysis of the experimental data, including the factors affecting the fusion and the comparison of the timbre fusion between Chinese and Western instruments, is presented. In Section Construction of the Timbre Fusion Model, the construction of the timbre fusion model is described. Multiple linear regression, random forest, and multilayer perceptron methods were used to construct the fusion model of Chinese instruments and Western instruments. In Section Discussion, the discussion and summary are presented.

METHODS

Participants

Thirty-two participants, including 15 males and 17 females (between 18 and 35 years of age), took part in this test. All the participants had received routine listening training for more than 1 year (M = 1.62, SD = 0.38). All the participants listened to different types of Chinese music, such as Jiangnan Sizhu, Fujian Nanyin, Guangdong music, and Chinese orchestral symphony, and Western music, such as pop, rock, classical, blues, and R&B. Among the participants, 22 of them listened to Chinese and Western music in a concert hall. All the participants met the required hearing threshold of 20 dB HL by a pure-tone audiometric test with octave-spaced frequencies from 125 to

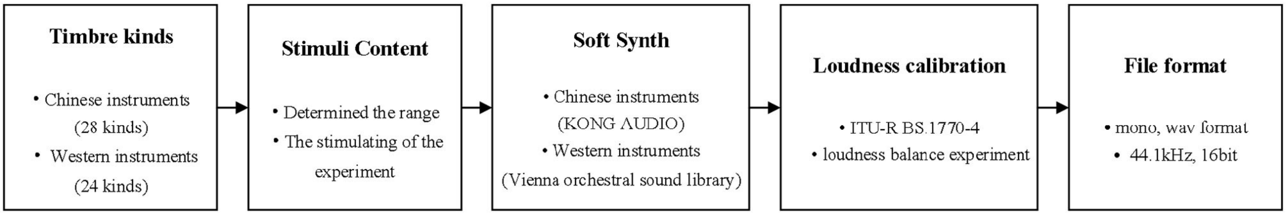

Timbre kinds
• Chinese instruments (28 kinds)
• Western instruments (24 kinds)
Stimuli Content
• Determined the range
• The stimulating of the experiment
Soft Synth
• Chinese instruments (KONG AUDIO)
• Western instruments (Vienna orchestral sound library)
Loudness calibration
• ITU-R BS.1770-4
• loudness balance experiment
File format
• mono, wav format
• 44.1kHz, 16bit

FIGURE 1 | Production of single-tone stimuli.

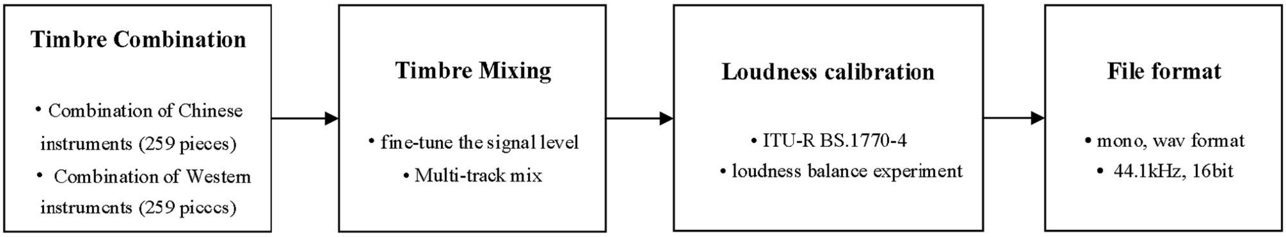

Timbre Combination
• Combination of Chinese instruments (259 pieces)
• Combination of Western instruments (259 pieces)
Timbre Mixing
• fine-tune the signal level
• Multi-track mix
Loudness calibration
• ITU-R BS.1770-4
• loudness balance experiment
File format
• mono, wav format
• 44.1kHz, 16bit

FIGURE 2 | Production of mixed timbre experimental stimuli.

8 kHz (Martin and Champlin, 2000). The participants, who were university students raised in China, were recruited in Beijing. The participants signed an informed consent form and were compensated for their participation.

Stimuli

There were 518 mixed timbre (composed of two timbre) stimuli. The process of making mixed timbre stimuli consisted of two steps. The first step was to determine the types of single-tone instruments to be mixed. Then, single-tone stimuli were created. The second step was to combine the single-tone stimuli to form some mixed stimuli based on two tones. The following is a detailed description of the stimuli production process, which is shown in **Figures 1**, **2**.

Production of Single-Tone Stimuli

Fifty-two kinds of instrument timbres were selected in this experiment. There were 24 Western instrument timbres, including wood wind instruments, brass wind instruments, bowed string instruments, hammered string instruments, and percussion instruments, and 28 Chinese instrument timbres, including wind instruments, bowed string instruments, plucked instruments, and percussion instruments. The experiment stimuli comprised four phrases in the lyric paragraph of the second part of the Spring Festival prelude. The music score is shown in Figure 1 of the **Appendix**. The range of each instrument is used most often by composers. The experimental stimuli were made by the combination of MIDI and a sampling sound source. Among them, the stimuli of Western instruments were produced by the Vienna Symphonic Library,[1] and the stimuli of Chinese instruments were produced by the Kong Audio Sound Library[2]. The audio file format was saved in the WAV format, the sampling frequency was 44.1 kHz, and the quantization accuracy was 16 bits. The timbre types of the Chinese and Western instruments and the specific range of each instrument are shown in Table 1 of the **Appendix**.

According to the definition of timbre, it is necessary to exclude the influence of pitch and loudness when studying it. Previous studies have shown that, in some cases, timbre and tone are inseparable (Melara and Marks, 1990). Therefore, the timbre perception features extracted in this paper also included the pitch factor. To avoid the influence of loudness on the perception results, all stimuli were first calibrated based on a loudness measurement algorithm (ITU-RBS 1770-4, 2015). Then, three audio engineers with music backgrounds fine-tuned the signal level based on the results of the music loudness balance experiment. The specific process of the loudness balance experiment and the statistical analysis of the experimental results have been detailed in previous research results (Zhu et al., 2018).

Production of Mixed Timbre Experimental Stimuli

Twenty-eight single-tone stimuli of Chinese instruments were divided into four groups: wind instruments, bowed string instruments, plucked instruments, and percussion instruments. By combining these stimuli in pairs within and between groups, we obtained 259 mixed timbre stimuli of Chinese instruments. Twenty-four single-tone stimuli of Western instruments were divided into five groups: wood wind instruments, brass wind instruments, bowed string instruments, hammered string instruments, and percussion instruments. By combining these stimuli in pairs within and between groups, we obtained 259 mixed timbre stimuli of Western instruments. In fact, the 24 Western musical instruments and 28 Chinese musical instruments each have more than 259 kinds of combinations.

[1] Website of the Vienna Symphonic Library: https://www.vsl.co.at/en/.
[2] Website of the Kong Audio Sound Library: http://www.kongaudio.com/.

Considering the amount of experimental data, we only evaluated the common timbre combination methods in the orchestration. These dyads are often used by composers. The combination of these dyads references the Chinese National Orchestra Practical Orchestration Manual. Similar to the single-tone stimuli, to avoid the influence of loudness on the perception results, all mixed experimental stimuli were first calibrated based on a loudness measurement algorithm (ITU-RBS 1770-4, 2015). Finally, a collection of 518 mixed stimuli was obtained, as shown in Table 2 of the **Appendix.**

Apparatus

The experiment was carried out in a listening room, conforming to standards (EBU-TECH 3253, 2008). The reverberation time of the listening room was 0.3 s, the sound field distribution was uniform, and there was no bad acoustic phenomenon or body noise. A Genelec 1038B three-way active midfield monitoring speaker was used to replay the experimental signals. Its parameters, which conform to international standards, are shown in Table 3 of the **Appendix**.

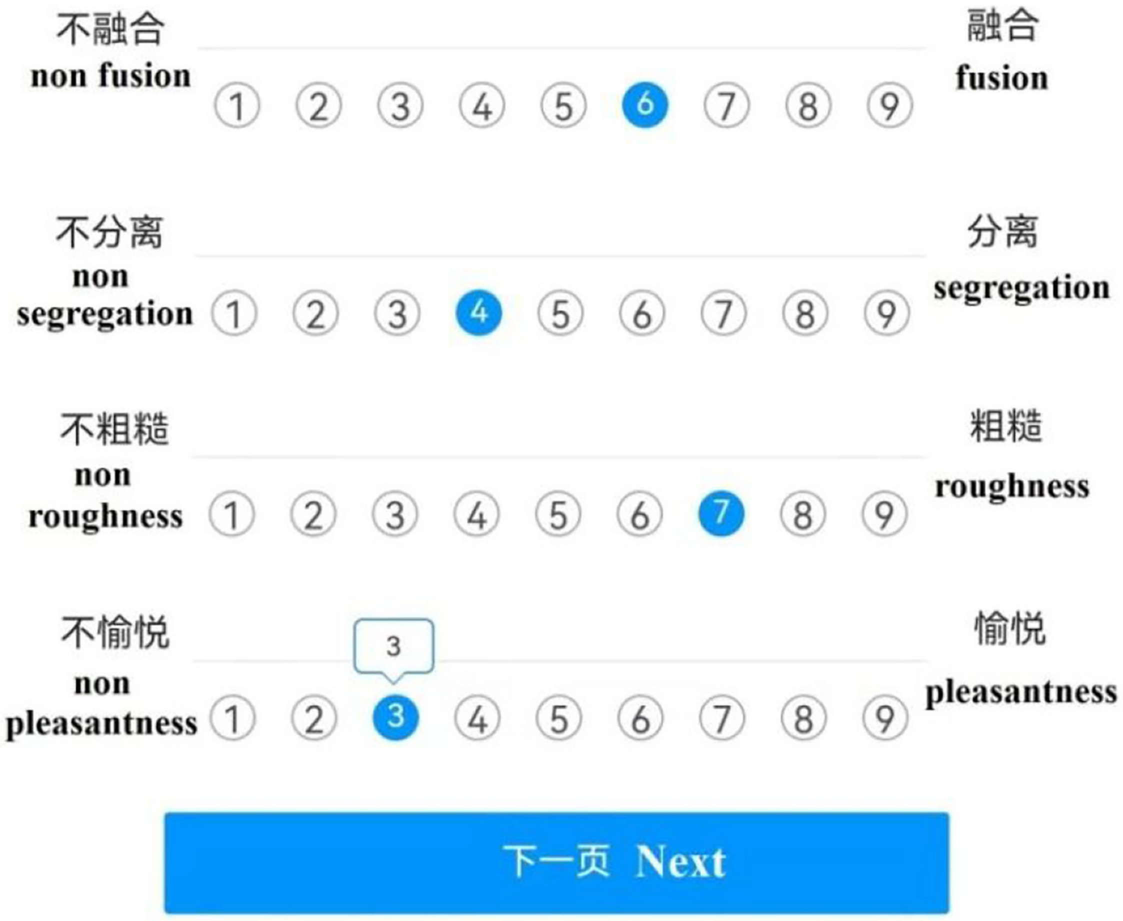
Because the experimental results are affected by the listening sound pressure level, it is necessary to ensure that the participants listen at the standard level (EBU-TECH 3253, 2008), and that this level remains unchanged throughout the experiment. The equipment used in the calibration test system is a Lenovo T460 notebook computer, a BK4231 sound calibrator, a BK2250 sound-level meter, and a YAMAHA 01V96i digital mixer. The actual listening pressure level is 74 dBA, which conforms to the international listening standard (EBU-TECH 3253, 2008). Experimental stimuli were played using Adobe Audition software. The seats in the listening room were arranged in triangles. That is, in the listening area, one listener sat in the first row, two listeners sat in the second row, and so on. To avoid presentation level changes caused by the shielding of the front seats from the back seats, the back seats were all 15 cm higher than the front seats. In the process of listening, the ear height of the participants should be at the same level as the midpoint of the vertical line in the high and low sounds of the speakers. We calibrated the test system with the sound-level meter. After the system was calibrated, white noise was used as the test signal. The sound-level meter was located in the center of the listening seat triangle. The system volume was adjusted so that the A-meter sound level of the system was 74 dBA (as read from the sound-level meter).

Procedure

The experimental steps included four stages: the experimental introduction stage, the pre-experimental stage, the training stage, and the formal experimental stage. The experimental introduction stage: The background of the experiment was introduced, and the participants were informed of the purpose of the experiment to enhance their cognition. Then, we explained the concept of the timbre perception attribute to the participants and used the example audio stimuli as an aid so that the participants could accurately understand the meaning of each attribute. The pre-experimental stage: We explained the corresponding relationship between the value of the 9-level evaluation scale and the degree of integration to the participants. Then, we randomly played all audio stimuli to familiarize the participants with their variation range. The training stage: Three timbre stimuli were randomly selected. The participants were asked to evaluate the fusion, segregation, roughness, and pleasantness of the stimuli according to their subjective feelings using the 9-level evaluation scale (1–9). The purpose of this step was to familiarize the participants with the experimental process and to avoid any experiment impacts related to unfamiliarity with the experimental process in the formal experimental stage. These data were not used for the analysis of the final results. The formal experiment stage: Thirty-two subjects were randomly divided into four groups. Each group consisted of eight subjects. A total of 518 stimuli were randomly divided into 52 stimuli groups. The order within a stimuli group was fixed, which was generated by a random program. And the order of the groups was random. To avoid the possibility of participant fatigue from listening to the sounds for a long amount of time, all experimental stimuli were divided into three sets. The experimental time of each set was no more than 30 min, and the rest, between each set, was 15 min. Each participant used a smartphone app to provide his or her responses. Then, we played the stimuli groups for each subject group. The participants were not allowed to communicate with one another during the test. The participants evaluated the fusion, segregation, roughness, and pleasantness of the timbre stimuli they heard and filled out the forms accordingly. The experimental data were collected by the app, in which we can select scores and export data. The app interface is shown in the figure below (**Figure 3**). The experiment was carried out according to the above steps, and the data collection was completed. The above steps were followed to conduct subjective evaluation experiments and complete the data collection.

SUBJECTIVE EVALUATION EXPERIMENT AND DATA ANALYSIS

The steps for the subjective evaluation of the experimental data analysis are as follows. First, the original experimental data were tested for reliability and validity, in which the reliability test was conducted by calculating Cronbach's alpha, and the validity test was conducted by calculating the standard deviation. Then, the method of successive categories was used to statistically analyze the experimental data, and the psychological scales of all samples in the four dimensions of fusion, segregation, roughness and pleasantness were obtained. These data were used for the comparative analysis of the timbre fusion of Chinese and Western instruments and the construction of timbre fusion modeling. Third, analysis of variance was used to statistically analyze the fusion, segregation, roughness, and pleasantness, and the differences in the four-dimensional attributes of different musical instrument timbre combinations were obtained. Finally, correlation analysis and multidimensional preference analysis were used to explore the relationship between the timbre perception dimension, musical instrument types,



不融合 non fusion 1 2 3 4 5 [✓ 6] 7 8 9 融合 fusion

不分离 non segregation 1 2 3 [✓ 4] 5 6 7 8 9 分离 segregation

不粗糙 non roughness 1 2 3 4 5 6 [✓ 7] 8 9 粗糙 roughness

不愉悦 non pleasantness 1 2 [✓ 3] 4 5 6 7 8 9 愉悦 pleasantness

FIGURE 3 | The interface for the listening test.

and temporal envelopes, and conclusions from the analyses were given.

Reliability and Validity Tests

Cronbach's alpha is used to evaluate the internal consistency of questionnaires and is applicable to the reliability analysis of attitudes, questionnaires or scales. Cronbach's alpha value is between 0 and 1. The higher the alpha coefficient is, the higher the reliability and the better the internal consistency of the questionnaire. Generally, a questionnaire with an alpha coefficient above 0.8 has value that is useful, and a questionnaire with an alpha coefficient above 0.9 shows that the reliability of the questionnaire is very good. The calculated Cronbach's alpha values of the four timbre perception attributes are shown in Table 4 of the **Appendix**. As seen from this table, the Cronbach alpha values were 0.932 for fusion, 0.941 for segregation, 0.926 for roughness, and 0.918 for pleasantness. These measures indicated that all scales had very good internal consistency among the 32 participants.

The validity test was designed to examine the validity of the experimental results. The higher the validity is, the better the measure shows the characteristics it is intended to measure. Different experiments have different purposes and require different levels of validity. The validity test for this experiment was to calculate the standard deviation of the experimental data for the 32 subjects for each experimental stimulus and to consider the experimental data beyond 1.5 times the standard deviation to be invalid and to eliminate them. Since the statistical model of the method of successive categories to be used next requires that there be no missing values in the experimental data, the data within 1.5 times the standard deviation of each stimulus were averaged, and the mean was used to fill in the missing values that were eliminated. After the reliability and validity tests, the data were statistically analyzed using the method of successive categories.

Data Statistics Based on the Method of Successive Categories

The experimental data were counted by the method of successive categories (Zihou, 2008). The theoretical basis of this method is to assume that the psychological quantity is a random variable that is subject to a positive Pacific distribution, and the boundary of each category in the method of successive categories is not a predetermined value but a random variable determined according to the experimental data. According to the Thurstone model, the preference of object a_i is the probability variable X_i on the preference scale, which follows the normal distribution, and its preference psychological scale $f\ (a_i)\ =\ S_i$. The dividing line between category g and category $g + 1$ is the random variable T_g on the subjective preference scale, which also follows the normal distribution $(T_g,\ \sigma^2)$. T_g and $f\ (a_i)$ satisfy the following relationship. The category judgment model is shown in **Table 1**.

The category judgment model was used to calculate statistics of the fusion, segregation, roughness, and pleasantness experimental data, and the psychological scale of all samples on each timbre perception attribute was obtained, as shown in **Figures 4–7**. N + N refers to non-sustaining instruments and non-sustaining instruments, S + N refers to sustaining instruments and non-sustaining instruments, and S + S refers to sustaining instruments and sustaining instruments. The abscissa represents the psychological scale distribution of each dimension, and the ordinate represents the serial number of each stimulus. It can be seen intuitively from the figures that there are certain differences in the distribution of timbre perception attributes with different temporal envelopes. The next section further analyzes these specific differences by combining one-way ANOVA and two-way ANOVA.

TABLE 1 | The category judgment model.

C_i / a_i	C_1	C_2	…	C_{m-1}
a_1	$t_1 - f\,(a_1) = z_{11}$	$t_2 - f\,(a_1) = z_{21}$	…	$t_{m-1} - f\,(a_1) = z_{m-11}$
a_2	$t_1 - f\,(a_2) = z_1$	$t_2 - f\,(a_2) = z_{22}$	…	$t_{m-1} - f\,(a_2) = z_{m-12}$
a_n	$t_1 - f\,(a_n) = z_{1n}$	$t_2 - f\,(a_n) = z_{2n}$	…	$t_{m-1} - f\,(a_n) = z_{m-1n}$
Sum	$nt_1 - \sum f\left(a_j\right)$	$nt_2 - \sum f\left(a_j\right)$	…	$nt_{m-1} - \sum f\left(a_j\right)$
Average	$t_1 - \frac{1}{n}\sum f\left(a_j\right)$	$t_2 - \frac{1}{n}\sum f\left(a_j\right)$	…	$t_{m-1} - \frac{1}{n}\sum f\left(a_j\right)$

Analysis and Discussion of Experimental Results

The listening test was a 3-×-2 mixed-measures design with two between subjects, the temporal-envelope factors and the instrument-type factors. There were three types of temporal-envelope factors: a combination of sustaining instruments and sustaining instruments (S + S), a combination of sustaining instruments and non-sustaining instruments (S + N), and a combination of non-sustaining instruments and non-sustaining instruments (N + N). There were two types of instrument type factors: Western instruments (W) and Chinese instruments (C).

In this part, the fusion, segregation, roughness, and pleasantness experimental data were analyzed. The analysis idea of each timbre perception dimension was as follows. First, based on the analysis results of the method of successive categories, the timbre combination forms of different temporal-envelope factors were statistically analyzed to compare the timbre perception attributes of different temporal envelopes. Second, one-way ANOVA was used to explore the effects of the temporal-envelope factors and instrument-type factors on each timbre perception attribute. Finally, two-way ANOVA was used to explore the differences in timbre perception attributes under the interaction between the temporal envelopes and instrument types.

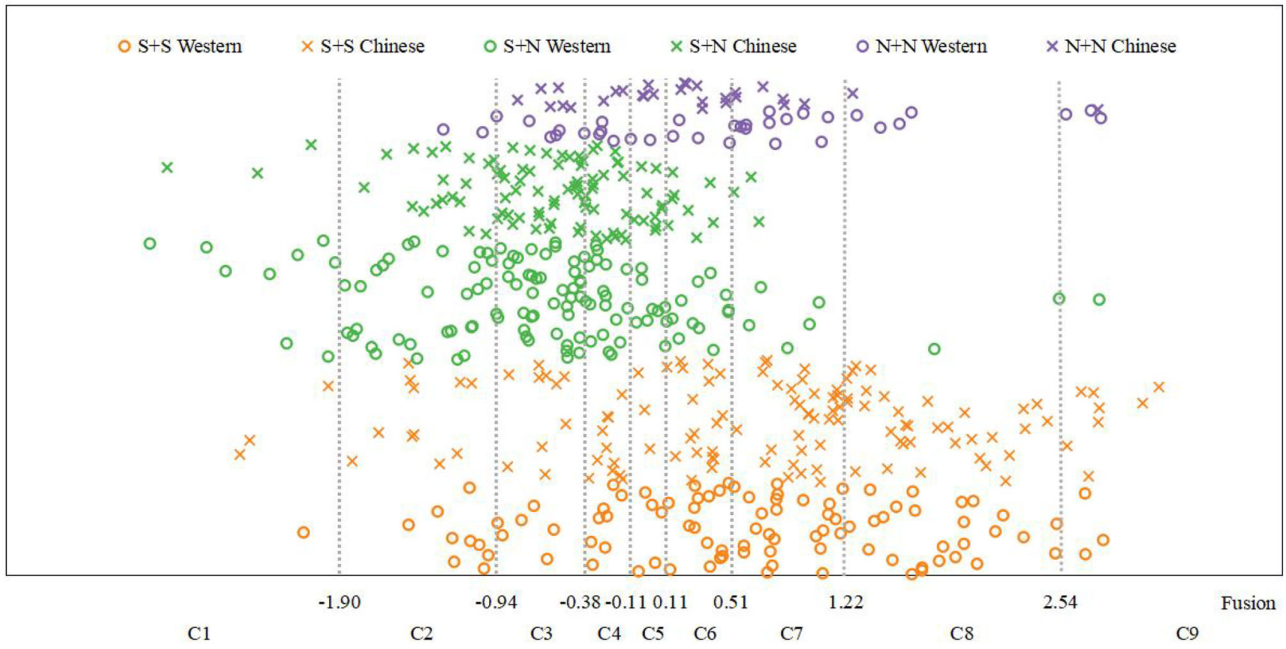


FIGURE 4 | Fusion psychological scales.

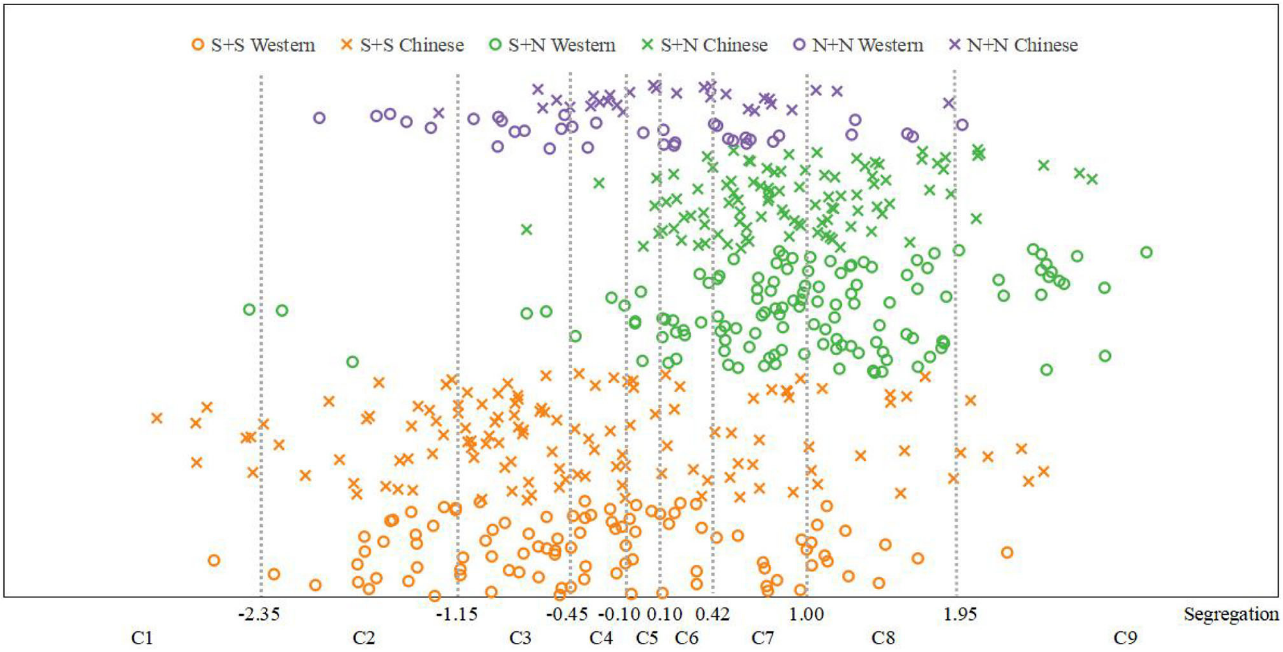


FIGURE 5 | Segregation psychological scales.

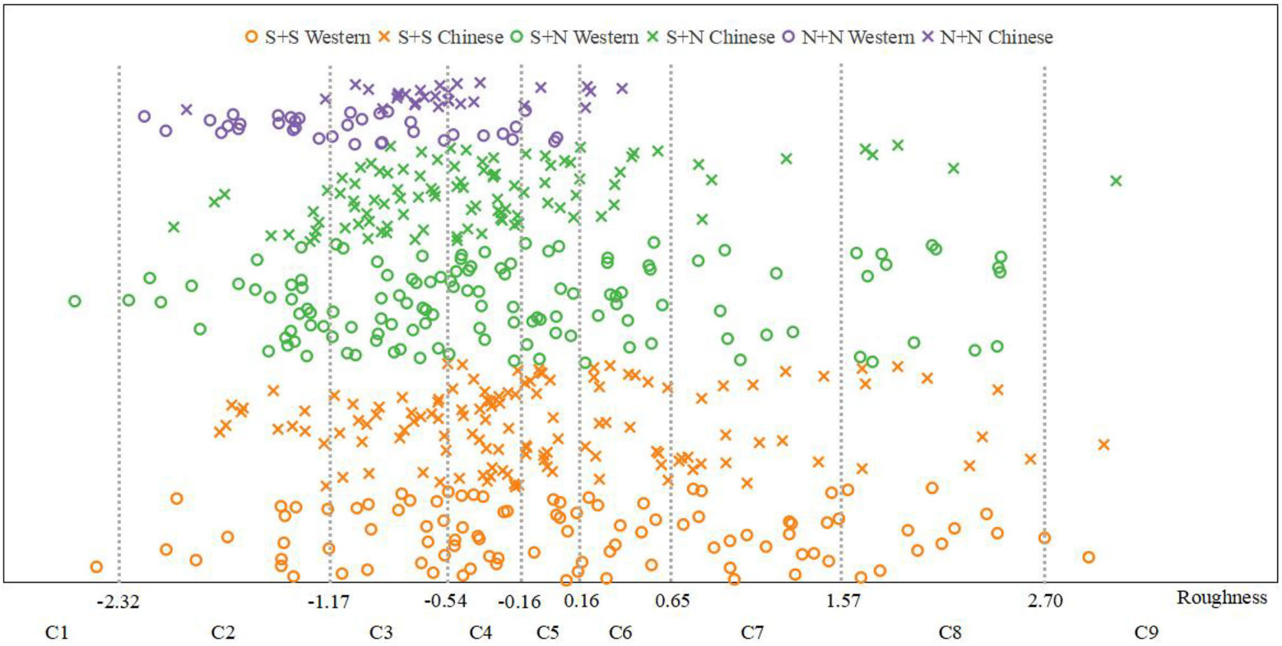


FIGURE 6 | Roughness psychological scales.

Fusion

Fusion Data Distribution Statistics

To further explore the relationship between the temporal envelope factor and fusion, we calculated the number and percentage of audio stimuli with different temporal envelopes in each category of fusion (Table 5 of the **Appendix**). Then, the frequency statistical histogram of each category of fusion was drawn according to the data in Table 5 of the **Appendix**, and the results are shown in **Figure 8**.

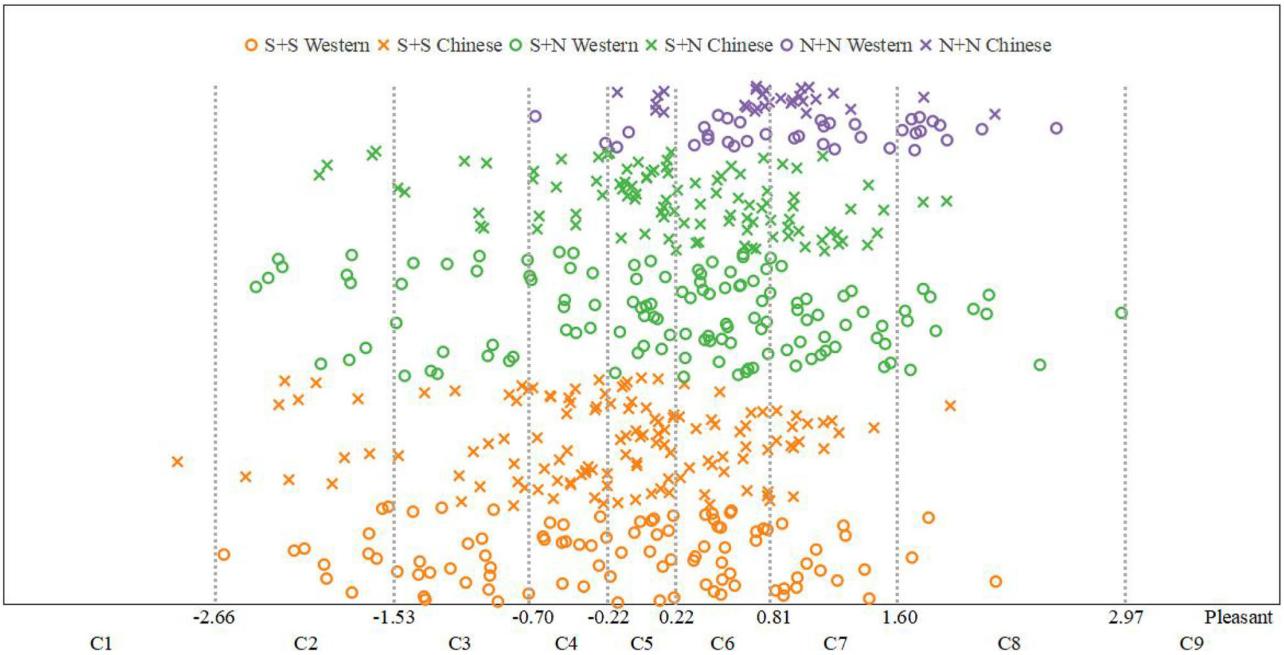


FIGURE 7 | Pleasantness psychological scales.

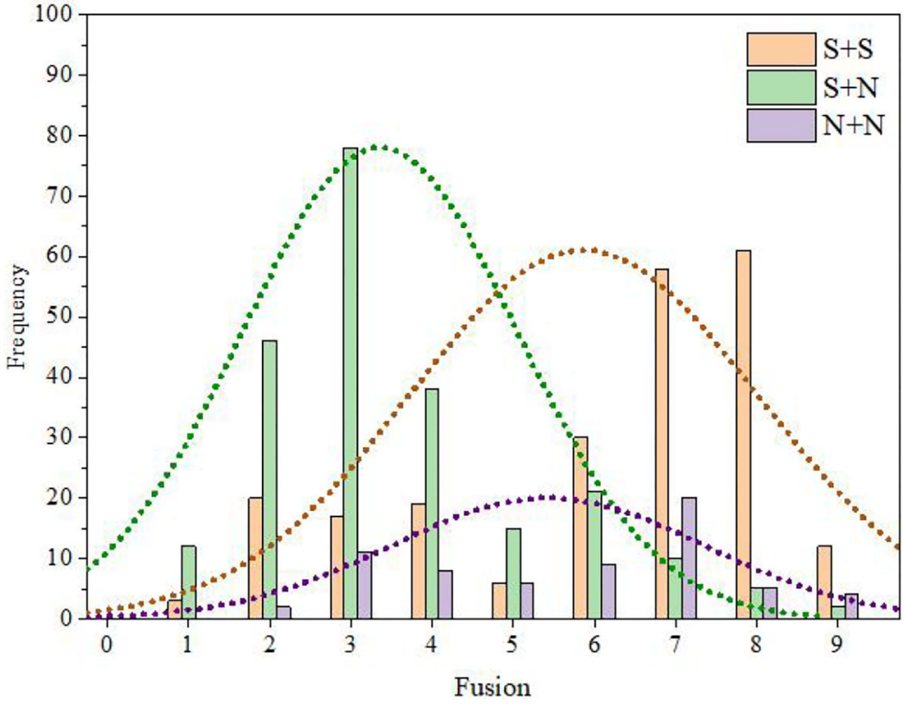


FIGURE 8 | Distribution statistical histogram of fusion category.

It can be seen from Table 5 of the **Appendix** and **Figure 8** that the fusion of S + S was mainly distributed in categories C6, C7, and C8 and was relatively high (M = 5.97, SD = 3.44). In contrast, the fusion of S + N was mainly distributed in categories C2, C3, and C4 and was relatively low (M = 3.98, SD = 1.61). There were no obvious distribution characteristics for the fusion of N + N

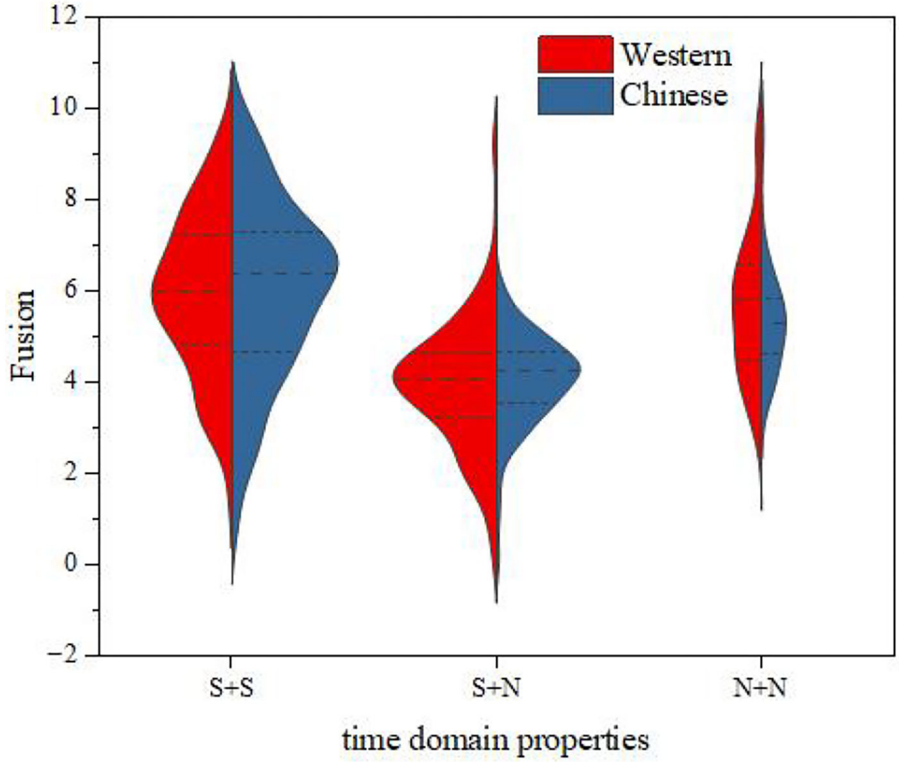


FIGURE 9 | Fusion comparisons between Western and Chinese instruments.

instruments, and there was a certain distribution in the range of categories C3–C7 (M = 5.51, SD = 2.07).

The relationship between the instrument type factor and the distribution of fusion was further discussed. We divided the experimental data into two groups according to the type of a musical instrument (i.e., Chinese instruments and Western instruments) and calculated the statistical characteristics of the fusion for each group. The results are shown in **Figure 9**. It can be seen from this figure that, for the same temporal envelope, the average fusion values for Chinese instruments and Western instruments were close, and the change law of the fusion from time to time was consistent; that is, the order of fusion from largest to smallest was S + S > N + N > S + N.

Influence of Instrument Types and Temporal Envelopes

To explore the influence of instrument types and temporal envelopes on the fusion, the one-way ANOVA model was used for statistical analysis. Before one-way ANOVA, the normality of the experimental fusion data was tested. The experimental data were grouped according to the instrument types and temporal envelopes, the normality of each group of data was tested, and the normal P-P was drawn, as shown in the Figure 2 of the **Appendix**. In this figure, the ordinate represents the cumulative probability of prediction, and the abscissa represents the cumulative probability of the actual data. If the measured curve is closer to the predicted cumulative probability (i.e., a line with a slope of 1), the actual data distribution is closer to the normal distribution. It can be seen from this figure that the data distribution of fusion met normality for both instrument type and temporal-envelope factors.

First, the experimental data were divided into two groups: Chinese instruments and Western instruments. The data of each group were analyzed by one-way ANOVA. The results are shown in the Table 6 of the **Appendix**. It can be seen from this table that the temporal envelope had an impact on both Western instruments [$p < 0.0001$, $F_{(2,256)} = 48.080$] and Chinese instruments [$p < 0.0001$, $F_{(2,256)} = 44.694$] with respect to fusion.

We further analyzed the influence of temporal envelopes on the fusion of Chinese and Western instruments and the differences between different temporal envelopes. Here, the Student–Newman–Keuls (SNK) method was used for pairwise comparisons between groups. The results are shown in the Tables 7, 8 of the **Appendices**.

For Western instruments, the three temporal envelopes were divided into two subgroups. The fusion scores of S + S and N + N were similar, and they were divided into the same subgroup. The significance $p = 0.326 > 0.05$ indicated that there was no difference between the average values of various types in the subgroup. The mean value of fusion in the second subgroup was greater than that in the first subgroup. For Chinese instruments, the three temporal envelopes were divided into three subgroups: S + S, S + N, and N + N. The fusion scores of the three subgroups were different, and the order of fusibility from largest to smallest was S + S > N + N > S + N.

From the above results, it can be concluded that the three temporal envelopes of both Chinese and Western instruments have an impact on the fusion. Moreover, temporal envelopes have a greater impact on the fusion of Chinese instruments but lesser impact on the fusion of Western instruments.

Then, we analyzed the factors of instrument type. The experimental data were divided into three groups: S + S, S + N, and N + N. The results are shown in the Table 9 of the **Appendix**. It can be seen from this table that under, any time domain property condition, the significant *P*-value of instrument types was >0.05 [S + S: $p = 0.863 > 0.05$, $F_{(1,224)} = 0.030$; S + N: $p = 0.564 > 0.05$, $F_{(1,225)} = 0.334$; N + N: $p = 0.319 > 0.05$, $F_{(1,63)} = 1.008$], indicating that there is no difference in the fusion score between Chinese and Western instruments under the three time domain property conditions; that is, the instrument type has no effect on the fusion.

Interaction Between Temporal Envelopes and Instrument Types

The above one-way ANOVA only considered the difference in fusion under the same factor. Next, we further studied the analysis model, considering both temporal envelopes and instrument types. Here, the two-way ANOVA model was used to analyze the fusion. Similar to one-way ANOVA, two-way ANOVA also requires normality testing. The P-P diagram of the normal probability distribution (normal P–P) was calculated and drawn, as shown in the Figure 3 of the **Appendix**. It can be seen from this figure that the measured curve was close to the predicted cumulative probability, indicating that the distribution of the experimental data met normality.

The results of two-way ANOVA for the experimental data are shown in the Table 10 of the **Appendix**. It can be seen from this table that the significance of instrument types and temporal envelopes was >0.05 ($p = 0.581 > 0.05$), indicating that the interaction between instrument types and temporal envelopes was not statistically significant. To make the model more concise, this interaction can be removed from the model, and the model can be fitted with only the main effect. The results are shown in the Table 11 of the **Appendix**. This table shows that the instrument types [$p = 0.906 > 0.05$, $F_{(5,512)} = 0.014$] had no effect on fusion, while the temporal envelope [$p < 0.0001$, $F_{(5,512)} = 92.469$] had an effect on fusion. That is, whether Chinese or Western instruments are utilized, the temporal envelope impacts the fusion. This conclusion is the same as that of one-way ANOVA, which further explains the relationship between the temporal envelope and fusion.

Combining the results of the descriptive statistical analyses, one-way ANOVA and two-way ANOVA, we can draw the following conclusions: (1) The temporal envelopes have a certain influence on the fusion; that is, the fusion of different temporal envelopes is different. The instrument types have no effect on fusion. For both Chinese and Western instruments, the order of fusion from largest to smallest is S + S > N + N > S + N. There is no significant difference in the ranking trend of fusion between Chinese and Western instruments. (2) An interaction between temporal envelopes and instrument types has not been found; that is, the difference in fusion between different temporal envelopes is, basically, the same in different instrument types.

Segregation

Segregation Data Distribution Statistics

Using the same methods as those used for fusion degree analysis, we obtained the frequency distribution statistics of each category of segregation (Table 12 of the **Appendix**) and the distribution statistical histogram of the segregation category (**Figure 10**).

It can be seen from Table 12 of the **Appendix** and **Figure 10** that the segregation of S + S was mainly distributed in categories C2, C3, and C4 and was relatively low (M = 4.35, SD = 3.22). In contrast, the segregation of S + N was mainly distributed in categories C7 and C8 and was relatively high (M = 6.53, SD = 1.67). The fusion of N + N did not show obvious distribution characteristics, and there was a certain distribution in the range of categories C2–C8 (M = 5.01, SD = 1.97).

The relationship between the instrument-type factor and the distribution of segregation was further discussed by the same method as that used for fusion degree analysis. The results are shown in **Figure 11**. It can be seen from this figure that, for the same temporal envelope, the average values of the segregation for Chinese instruments and Western instruments were close, and the change law of the segregation from time to time was consistent; that is, the order of the segregation from largest to smallest was S + N > N + N > S + S.

Influence of Instrument Types and Temporal Envelopes

We tested the normality of the experimental data of segregation and then analyzed them by one-way ANOVA. The results are shown in Figure 4 and Table 13 of the **Appendices**. It can be seen from this table that the temporal envelope had an impact on both Western instruments [$p < 0.0001$, $F_{(2,256)} = 58.651$] and Chinese instruments [$p < 0.0001$, $F_{(2,256)} = 54.167$] with respect to segregation.

We further analyzed the influence of temporal envelopes on the segregation of Chinese and Western instruments and the differences between different temporal envelopes. Here, the Student–Newman–Keuls (SNK) method was used for pairwise comparisons between groups. The results are shown in the Tables 14, 15 of the **Appendices**.

For Western instruments, the three temporal envelopes were divided into two subgroups. The segregation scores of S + S and N + N were similar, and they were divided into the same subgroup. There was no difference between the average values of various types in the subgroup ($p = 0.108 > 0.05$). The mean value of segregation in the second subgroup was greater than that in the first subgroup. For Chinese instruments, the three temporal envelopes were divided into three subgroups: S + S, S + N, and N + N. The segregation scores of the three were different, and the order of segregation from largest to smallest was S + N > N + N > S + S.

From the above results, it can be concluded that the three temporal envelopes of both Chinese and Western instruments have an impact on segregation. Moreover, temporal envelopes

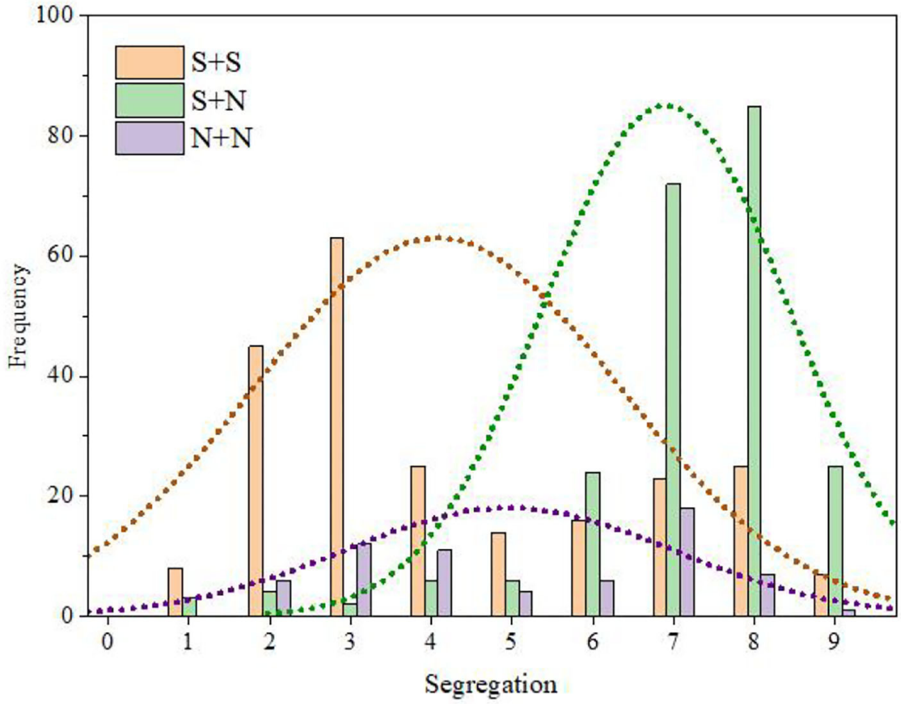


FIGURE 10 | Distribution statistical histogram of segregation category.

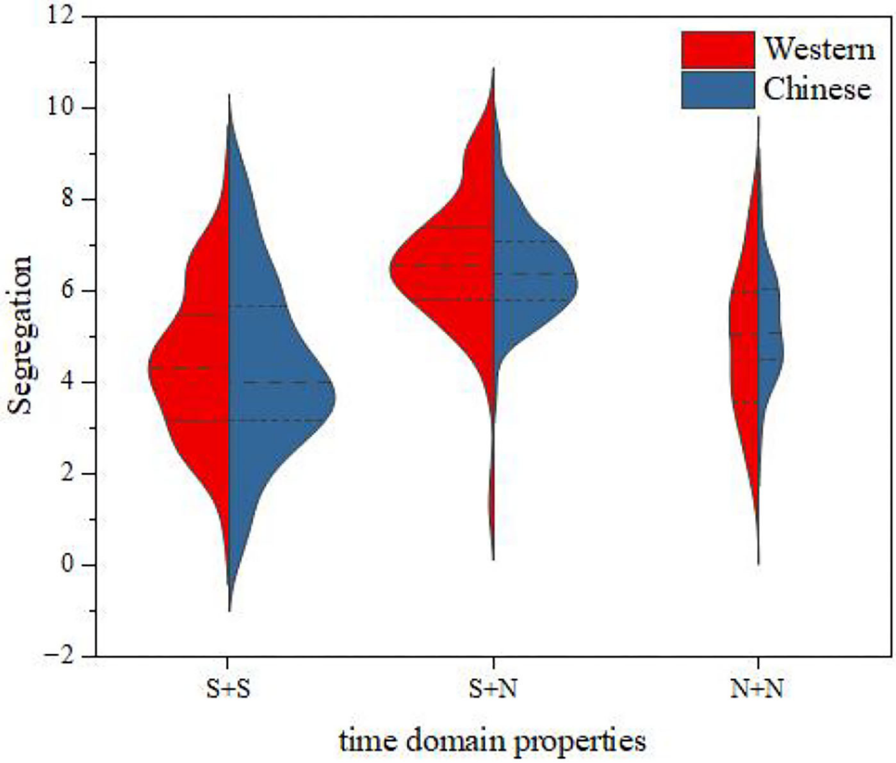


FIGURE 11 | Segregation comparisons between Western and Chinese instruments.

have a greater impact on the segregation of Chinese instruments but a lesser impact on the segregation of Western instruments.

Then, we analyzed the factors of instrument type. The experimental data were divided into three groups: S + S, S + N, and N + N. The results are shown in the Table 16 of the **Appendix**. It can be seen from the table that, under any time domain property condition, the significant P value of instrument types was > 0.05 [S + S: $p = 0.732 > 0.05$, $F_{(1,224)} = 0.117$; S + N: $p = 0.505 > 0.05$, $F_{(1,225)} = 0.445$; N + N: $p = 0.268 > 0.05$, $F_{(1,63)} = 1.249$], indicating that there is no difference in the segregation score between Chinese and Western instruments under the three time domain property conditions; that is, the instrument type has no effect on the segregation.

Interaction Between Temporal Envelopes and Instrument Types

The above one-way ANOVA only considered the difference in segregation under the same factor. Next, we further studied the analysis model, considering both temporal envelopes and instrument types. Here, the two-way ANOVA model was used to analyze segregation. Similar to one-way ANOVA, two-way ANOVA also requires normality testing. The P–P diagram of the normal probability distribution (normal P–P) was calculated and drawn, as shown in the Figure 5 of the **Appendix**. It can be seen from this figure that the measured curve was close to that of the predicted cumulative probability, indicating that the distribution of the experimental data met normality.

The results of two-way ANOVA for the experimental data are shown in the Table 17 of the **Appendix**. It can be seen from this table that the significance of instrument types and temporal envelopes was >0.05 ($p = 0.492 > 0.05$), indicating that the interaction between instrument types and temporal envelopes was not statistically significant. To make the model more concise, this interaction can be removed from the model, and the model can be fitted with only the main effect. The results are shown in the Table 18 of the **Appendix**. This table shows that the instrument types [$p = 0.785 > 0.05$, $F_{(5,512)} = 0.075$] had no effect on segregation, while the temporal envelope [$p < 0.0001$, $F_{(5,512)} = 112.211$] had an effect on segregation. That is, whether Chinese or Western instruments are utilized, the temporal envelope impacts segregation. This conclusion is the same as that of one-way ANOVA, which further explains the relationship between the temporal envelope and segregation.

Combining the results of the descriptive statistical analyses, one-way ANOVA and two-way ANOVA, we can draw the following conclusions: (1) The temporal envelopes have a certain influence on segregation; that is, the segregation of different temporal envelopes is different. The instrument types have no effect on segregation. For both Chinese and Western instruments, the order of segregation from largest to smallest is S + N > N + N > S + S. There is no significant difference in the ranking trend of segregation between Chinese and Western instruments. (2) An interaction between temporal envelopes and instrument types has not been found; that is, the difference in segregation between different temporal envelopes is, basically, the same for different instrument types.

Roughness

Roughness Data Distribution Statistics

Using the same methods as those used for fusion degree analysis, we obtained the frequency distribution statistics of each category of roughness (Table 19 of the **Appendix**) and the distribution statistical histogram of the roughness category (**Figure 12**).

It can be seen from Table 19 of the **Appendix** and **Figure 12** that the roughness of S + S was mainly distributed in categories C3 and C4 and was relatively high (M = 4.96, SD = 2.94). In contrast, the roughness of S + N was mainly distributed in categories C2, C3, and C4 and was higher (M = 4.42, SD = 2.60). The roughness of N + N was mainly distributed in categories C3, C4, and C5 and was relatively low (M = 3.54, SD = 0.93).

The relationship between the instrument-type factors and the distribution of roughness was further discussed by the same method as that used for fusion degree analysis. The results are shown in **Figure 13**. It can be seen from this figure that, for S + S and S + N, the mean roughness values for Chinese instruments and Western instruments were close, and the variation law of roughness from time to time was consistent; for N + N, the mean roughness of Chinese instruments was larger than that for Western instruments. However, for all western and Chinese instruments, the roughness still had the same law from largest to smallest, i.e., S + S > S + N > N + N.

Influence of Instrument Types and Temporal Envelope

Similarly, we tested the normality of the experimental roughness data and then analyzed them by one-way ANOVA. The results are shown in Figure 6 and Table 20 of the **Appendices**. It can be seen from this table that the temporal envelope had an impact on both Western instruments [$p < 0.0001$, $F_{(2,256)} = 15.902$] and Chinese instruments [$p = 0.0001$, $F_{(2,256)} = 6.833$] with respect to roughness.

We further analyzed the influence of temporal envelopes on the roughness of Chinese and Western instruments and the differences between different temporal envelopes. Here, the Student–Newman–Keuls (SNK) method was used for pairwise comparisons between groups. The results are shown in the (Tables 21, 22 of the **Appendices**).

For Western instruments, the three temporal envelopes were divided into three subgroups: S + S, S + N and N + N. The roughness scores of the three subgroups were different, and the order of roughness from largest to smallest was S + S > S + N > N + N. For Chinese instruments, the three temporal envelopes were divided into two subgroups: the roughness scores of S + N and N + N were similar, so they were divided into the same subgroup. There was no difference between the average values of various types in the subgroup ($p = 0.209 > 0.05$). The mean value of roughness in the second subgroup was greater than that in the first subgroup.

From the above results, it can be concluded that the three temporal envelopes of both Chinese and Western instruments have an impact on the roughness. Moreover, temporal envelopes have a greater impact on the roughness of Western instruments but a lesser impact on the roughness of Chinese instruments.

Then, we analyzed the factors of instrument type. The experimental data were divided into three groups: S + S, S +

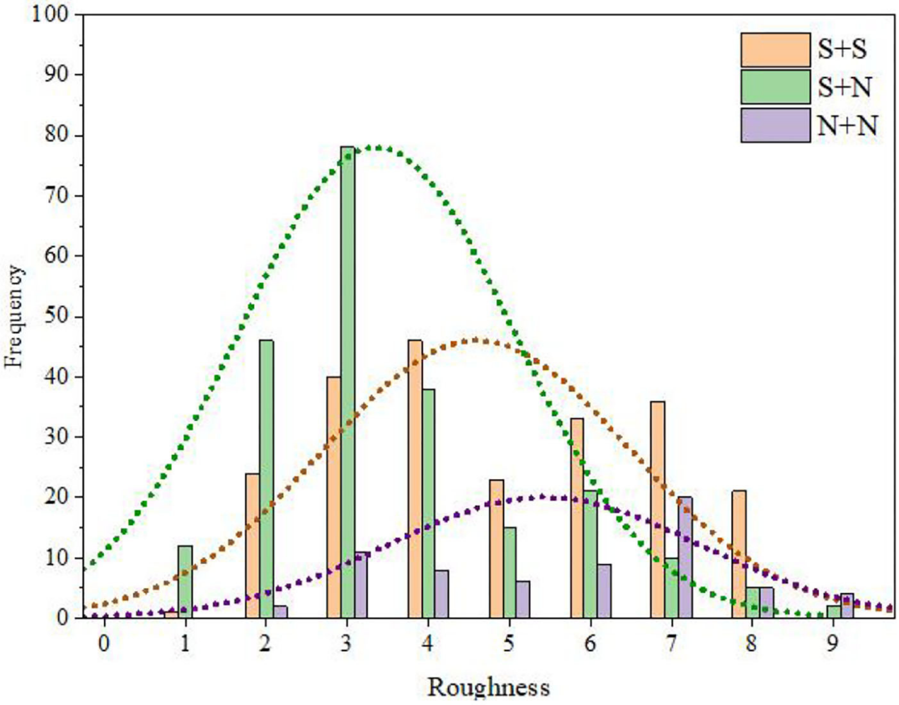


FIGURE 12 | Distribution statistical histogram of roughness category.

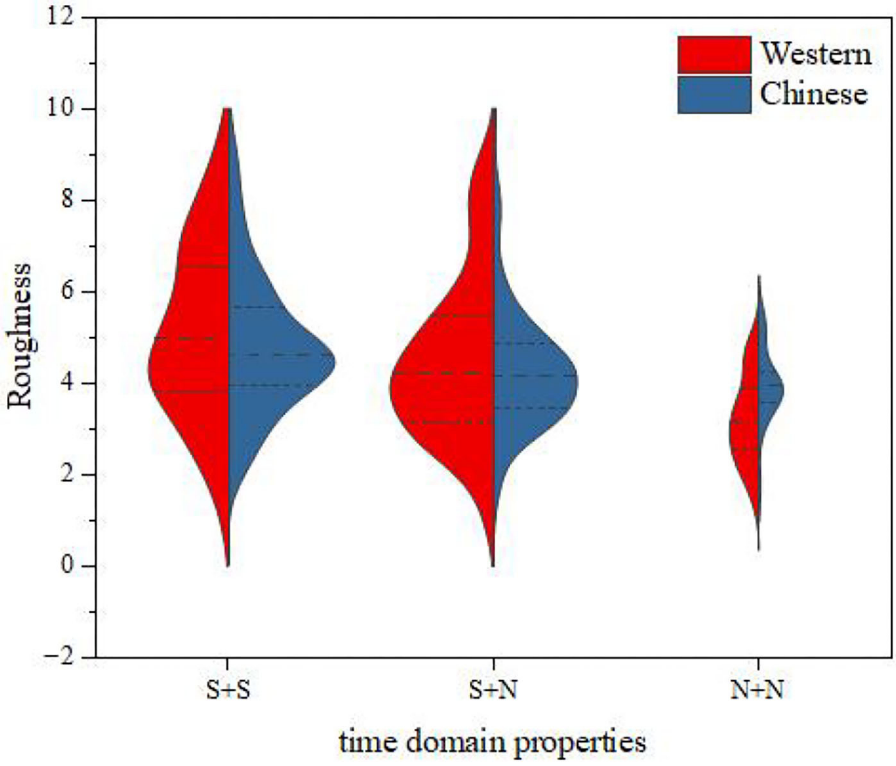


FIGURE 13 | Roughness comparisons between Western and Chinese instruments.

N, and N + N. The results are shown in the (Table 23 of the **Appendix**). It can be seen from this table that, when the time domain property condition was S + S or S + N, the significant P-value of instrument type was >0.05 [S + S: $p = 0.249 > 0.05$, $F_{(1,224)} = 1.335$; S + N: $p = 0.369 > 0.05$, $F_{(1,225)} = 0.809$], indicating that there was no difference in the roughness score between Chinese and Western instruments. However, when the time domain property condition was N + N [$p = 0.001 < 0.05$, $F_{(1,63)} = 13.010$] there were some differences between Chinese and Western instruments under the three time domain property conditions; that is, the instrument type affected the roughness when the time domain property condition was N + N.

Interaction Between Temporal Envelope and Instrument Types

The above one-way ANOVA only considered the difference in roughness under the same factor. Next, we further studied the analysis model, considering both temporal envelopes and instrument types. Here, the two-way ANOVA model was used to analyze roughness. Similar to one-way ANOVA, two-way ANOVA also requires normality testing. The P–P diagram of the normal probability distribution (normal P–P) was calculated and drawn, as shown in the Figure 7 of the **Appendix**. It can be seen from this figure that the measured curve was close to the predicted cumulative probability, indicating that the distribution of the experimental data met normality.

The results of two-way ANOVA of the experimental data are shown in the Table 24 of the **Appendix**. It can be seen from this table that the significance of instrument types and temporal envelopes was greater than 0.05 ($p = 0.053 > 0.05$), indicating that the interaction between instrument types and temporal envelopes was not statistically significant. To make the model more concise, this interaction can be removed from the model, and the model can be fitting with only the main effect. The results are shown in the Table 25 of the **Appendix**. The table shows that the instrument types [$p = 0.478 > 0.05$, $F_{(5,512)} = 0.505$] had no effect on the roughness, while the temporal envelope [$p < 0.0001$, $F_{(5,512)} = 21.654$] had an effect on the roughness. That is, whether Chinese or Western instruments are utilized, the temporal envelope has an impact on the roughness. This conclusion is the same as that of one-way ANOVA, which further explains the relationship between temporal envelopes and roughness.

Combining the results of the descriptive statistical analyses, one-way ANOVA and two-way ANOVA, we can draw the following conclusions: (1) The temporal envelopes have a certain influence on the roughness, that is, the roughness of different temporal envelopes are different. For both Chinese and Western instruments, the order of roughness from largest to smallest is S + S > N + N > S + N. There is no significant difference in the ranking trend of roughness between Chinese and Western instruments. (2) Interactions between temporal envelopes and instrument types have not been found; however, when the time domain property is N + N, the instrument type has an effect on the roughness. That is, when the time domain property is N + N, the roughness of Chinese instruments is larger than that of Western instruments.

Pleasantness

Pleasantness Data Distribution Statistics

Using the same methods as those used for fusion degree analysis, we obtained the frequency distribution statistics of each category of pleasantness (Table 26 of the **Appendix**) and the distribution statistical histogram of the pleasantness category (**Figure 14**).

It can be seen from Table 26 of the **Appendix** and **Figure 14** that the pleasantness of S + S was mainly distributed in categories C3–C6 and was relatively low (M = 4.64, SD = 2.23). In contrast, the pleasantness of S + N was mainly distributed in categories C5–C7 and was relatively high (M = 5.26, SD = 2.43). The pleasantness of N + N was mainly distributed in categories C6–C8 and was higher (M = 6.36, SD = 1.07).

The relationship between the instrument-type factors and the distribution of pleasantness was further discussed by the same method as that used for fusion degree analysis. The results are shown in **Figure 15**. It can be seen from this figure that, for the same temporal envelope, the average values of pleasantness for Chinese instruments and Western instruments were close, and the change law of pleasantness from time to time was consistent; that is, the order of pleasantness from largest to smallest was N + N > S + N > S + S.

Influence of Instrument Types and Temporal Envelope

Similarly, we tested the normality of the experimental pleasantness data and then analyzed them by one-way ANOVA. The results are shown in Figure 8 and Table 27 of the **Appendices**. It can be seen from this table that the temporal envelope had an impact on both Western instruments [$p < 0.0001$, $F_{(2,256)} = 17.417$] and Chinese instruments [$p < 0.0001$, $F_{(2,256)} = 18.807$] with respect to pleasantness.

We further analyzed the influence of temporal envelopes on the pleasantness of Chinese and Western instruments and the differences between different temporal envelopes. Here, the Student–Newman–Keuls (SNK) method was used for pairwise comparisons between groups. The results are shown in the Tables 28, 29 of the **Appendices**.

For Western instruments, the three temporal envelopes were divided into three subgroups: S + S, S + N, and N + N. The pleasantness scores of the three subgroups were different, and the order of pleasantness from largest to smallest was N + N > S + N > S + S. For Chinese instruments, the three temporal envelopes were also divided into three subgroups: S + S, S + N, and N + N. The pleasantness scores of the three subgroups were different, and the order of pleasantness from largest to smallest was N + N > S + N > S + S.

From the above results, it can be concluded that the three temporal envelopes of both Chinese and Western instruments have the same impact on pleasantness.

Then, we analyzed the factors of instrument type. The experimental data were divided into three groups: S + S, S + N, and N + N. The results are shown in the Table 30 of the **Appendix**. It can be seen from this table that, under any time domain property condition, the significant P-value of instrument type was >0.05 [S + S: $p = 0.926 > 0.05$, $F_{(1,224)} = 0.009$; S + N: $p = 0.900 > 0.05$, $F_{(1,255)} = 0.016$; N + N: $p = 0.148 > 0.05$, $F_{(1.63)} = 2.147$], indicating that there is no difference in the pleasantness

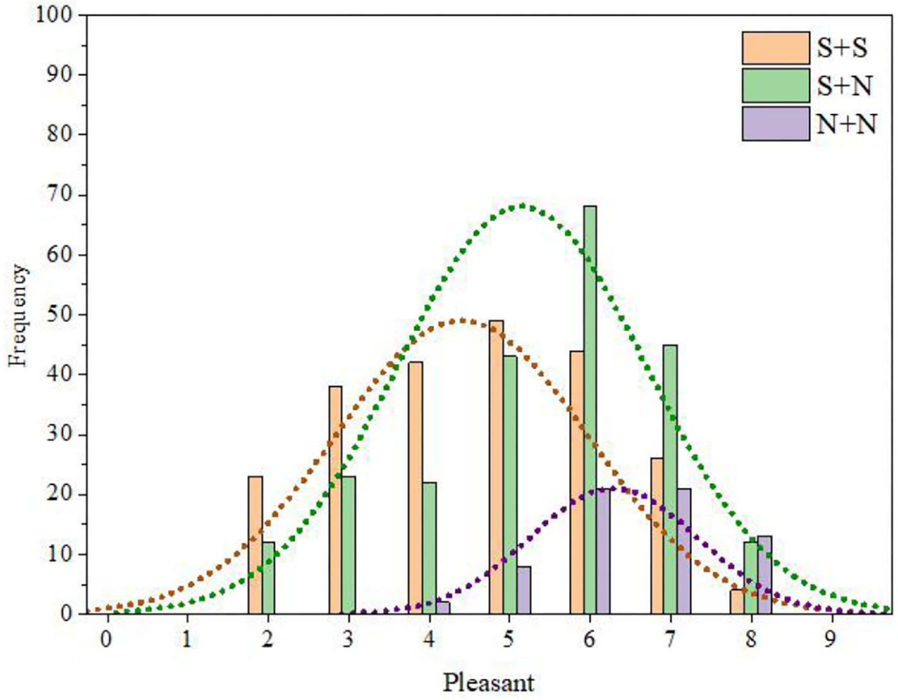


FIGURE 14 | Distribution of statistical histogram of pleasantness category.

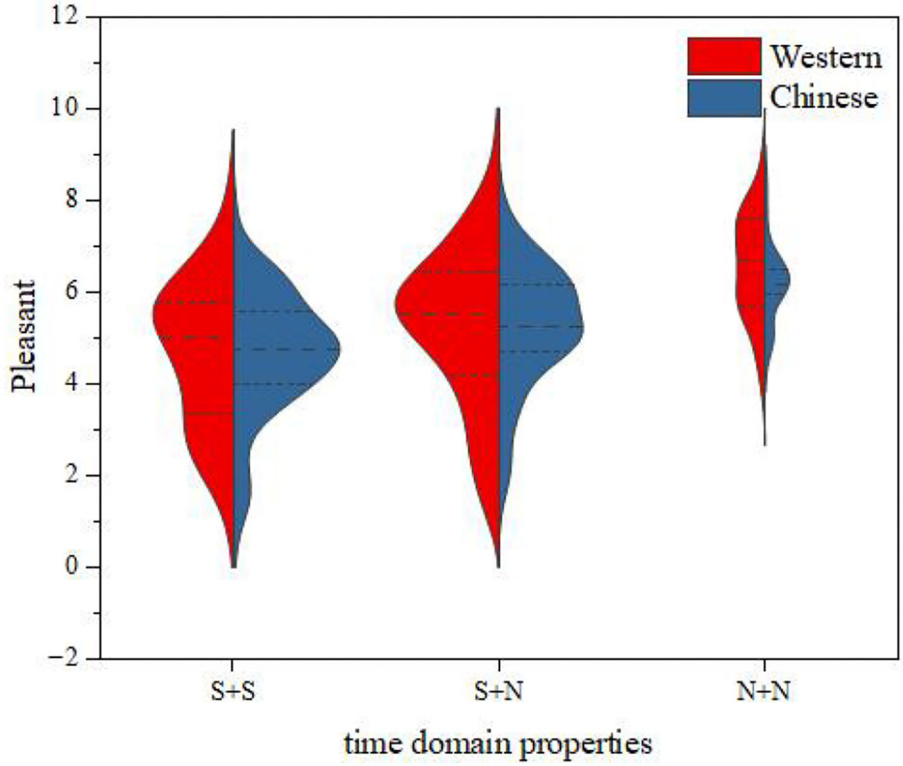


FIGURE 15 | Pleasantness comparisons between Western and Chinese instruments.

TABLE 2 | Correlation matrix and test results of timbre perception attributes.

		Fusion	Segregation	Roughness	Pleasantness
Fusion	Pearson correlation	1	−0.945**	−0.471**	0.471**
	Sig. (2-tailed)		$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
Segregation	Pearson correlation	−0.945**	1	0.371**	−0.370**
	Sig. (2-tailed)	$p < 0.0001$		$p < 0.0001$	$p < 0.0001$
Roughness	Pearson correlation	−0.471**	0.371**	1	−0.892**
	Sig. (2-tailed)	$p < 0.0001$	$p < 0.0001$		$p < 0.0001$
Pleasantness	Pearson Correlation	0.471**	−0.370**	−0.892**	1
	Sig. (2-tailed)	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	

***Correlation is significant at the 0.01 level (2-tailed).*

score between Chinese and Western instruments under the three time domain property conditions, that is, the instrument type has no effect on the pleasantness.

Interaction Between Temporal Envelope and Instrument Types

The above one-way ANOVA only considered the difference in pleasantness under the same factor. Next, we further studied the analysis model, considering both temporal envelopes and instrument types. Here, the two-way ANOVA model was used to analyze pleasantness. Similar to one-way ANOVA, two-way ANOVA also requires normality testing. The P–P diagram of the normal probability distribution (normal P-P) was calculated and drawn, as shown in the Figure 9 of the **Appendix**. It can be seen from this figure that the measured curve was close to that of the predicted cumulative probability, indicating that the distribution of the experimental data met normality.

The results of two-way ANOVA of the experimental data are shown in the Table 31 of the **Appendix**. It can be seen from this table that the significance of instrument types and temporal envelopes was >0.05 ($p = 0.624 > 0.05$), indicating that the interaction between instrument types and temporal envelopes was not statistically significant. To make the model more concise, this interaction can be removed from the model, and the model can be refitting with only the main effect. The results are shown in the Table 32 of the **Appendix**. This table shows that the instrument types [$p = 0.738 > 0.05, F_{(5,512)} = 0.112$] had no effect on pleasantness, while the temporal envelope [$p < 0.0001, F_{(5,512)} = 35.505$] had an effect on pleasantness. That is, whether Chinese or Western instruments are utilized, the temporal envelope has an impact on pleasantness. This conclusion is the same as that of one-way ANOVA, which further explains the relationship between temporal envelope and pleasantness.

Combining the results of the descriptive statistical analyses, one-way ANOVA and two-way ANOVA, we can draw the following conclusions: (1) The temporal envelopes have a certain influence on pleasantness; that is, the pleasantness of different temporal envelopes is different. The instrument types have no effect on pleasantness. For both Chinese and Western instruments, the order of pleasantness from largest to smaller is N + N > S + N > S + S. There is no significant difference in the ranking trend of pleasantness between Chinese and Western instruments. (2) An interaction between temporal envelopes and instrument types has not been found; that is, the difference in pleasantness between different temporal envelopes is, basically, the same in different instrument types.

Interaction of Timbre Perception Attributes

Here, the correlation analysis was carried out by using the Pearson's correlation coefficient for four timbre perception attributes, fusion, segregation, roughness, and pleasantness. The correlation matrix and test results can be calculated by a two-tailed test, as shown in **Table 2**. As seen from this table, there is a strong negative correlation between fusion and segregation ($r = -0.94$, Sig < 0.01), indicating that these two attributes tend to move in opposite directions. There is also a strong negative correlation between roughness and pleasantness ($R = -0.0.94$, Sig < 0.01), indicating that these two attributes tend to move in opposite directions. In addition to the above two pairs of strong correlations, other correlations among the four timbre perception attributes were weak. To further analyze the relationship between these attributes, multidimensional preference analysis was used to process the experimental data.

Multidimensional preference analysis is also called principal component analysis of classified data. The principle of this algorithm is to combine the idea of optimal scaling transformation and principal component analysis. In essence, this method is an extension of factor analysis and principal component analysis (Bechtel, 2019). Compared with principal component analysis and factor analysis, multidimensional preference analysis has several advantages. Considering various possible factors in data collection, the optimal scaling technique was introduced in multidimensional preference analysis. This allows the analysis of distance (continuous) variables and order (discrete) variables (such as rating and ranking), thus greatly broadening the application scope of this method. In addition, the results of multidimensional preference analysis can be intuitively presented in the form of a perception map. In other words, the sample and variable loadings can be plotted directly on a single diagram, making it easier to read information from it.

The experimental data can be statistically processed by multidimensional preference analysis. The component loadings of the four timbre perception attributes of the perception map and preference space can be obtained (**Table 3**, **Figure 16**). In

TABLE 3 | Component loads of timbre perception attributes in two dimensions.

	Dimension 1	Dimension 2
Fusion	0.839	−0.511
Segregation	−0.789	0.588
Roughness	−0.787	−0.569
Pleasantness	0.793	0.561

the preference space, the origin represents the average level of the whole sample. Starting from the origin, the further the scatter is from the origin, the stronger its tendency is. Points falling in the same direction from the origin in roughly the same region are related to each other. Variable scatter may represent a potential factor.

As seen from **Figure 16**, the four variables, which represent the fusion, segregation, roughness, and pleasantness timbre perception attributes, are distributed in the four quadrants of the preference space, which shows that these four attributes are representative for evaluating the combined timbre. In addition, fusion and segregation show opposite distributions, and roughness and pleasantness also show opposite distributions, indicating that these two pairs of attributes tend to move in opposite directions, which further verifies the results of the correlation analysis.

For the time domain characteristic factors, the timbre of the N + N type is distributed near the loading component of the pleasantness attribute, indicating that pleasantness is the main factor affecting this type of timbre. The timbre of the S + N type is distributed above the loading component of segregation, indicating that segregation is the main factor affecting this type of timbre, while the pleasantness attribute also has a slight influence on this type of timbre. The timbre of the S + S type is distributed in the middle of the fusion and roughness components, indicating that both fusion and roughness have a certain influence on the timbre of the S + S type, and the influence of fusion is slightly greater than the influence of roughness. For instrument-type factors, the scatter points of Chinese and Western instruments are very close to the origin, indicating that instrument type is not the main factor affecting timbre perception attributes. In summary, the results based on multidimensional preference analysis are consistent with the previous results of variance analysis, which further demonstrates the reliability of the conclusion.

CONSTRUCTION OF THE TIMBRE FUSION MODEL

To explain the influencing factors of perception fusion, this paper draws on the analysis ideas of existing research and uses audio information processing methods to extract mixed audio features from time-domain waveforms and frequency spectra. Then, we attempted to establish the correlation between objective acoustic parameters and subjective perception.

Extracting Acoustic Characteristic Parameters

Timbre is a multidimensional perception attribute that is closely related to the time-domain waveform and spectral structure of sound (Jiang et al., 2020). Objective acoustic parameters refer to any values acquired using a mathematical model, representing a normal sound signal in the time and frequency domains. To establish a timbre fusion model, an objective acoustic parameter set was constructed using 27 parameters extracted from the 518 stimuli in the timbre fusion database. These 27 parameters can be divided into 6 categories (Peeters et al., 2011):

(1) *Temporal shape features:* calculated from the waveform or the signal energy envelope (e.g., attack time, temporal increase or decrease, and effective duration).
(2) *Temporal features:* autocorrelation coefficients with a zero-crossing rate.
(3) *Energy features:* referring to various energy contents in the signal (e.g., global energy, harmonic energy, or noise energy).
(4) *Spectral shape features:* calculated from the short-time Fourier transform (STFT) of the signal (e.g., centroid, spread, skewness, kurtosis, slope, roll-off frequency, or Mel-frequency cepstral coefficients).
(5) *Harmonic features:* calculated using sinusoidal harmonic modeling of the signal (e.g., the harmonic/noise ratio, the odd-to-even and tristimulus harmonic energy ratio, and harmonic deviation).
(6) *Perceptual features:* calculated using a model for human hearing (e.g., relative specific loudness, sharpness, and spread).

Considering that the parameters of the audio stimuli change with time, we calculated the time-varying statistical of these parameters, including the maximum, minimum, mean, variance, standard deviation, interquartile range, skewness coefficient, and kurtosis coefficient so as to produce an objective acoustic parameter set, containing 216 parameters. We screened these parameters before establishing the regression equation. The correlation between 216 parameters and fusion was analyzed, and it was found that the correlation between mean, interquartile range, and fusion was relatively high. Therefore, the mean and interquartile range of 27 parameters were retained, and a total of 54 parameters were retained. Considering the nine parameters attack time, log attack time, decrease time, effective duration, release time, attack slope, decrease slope, frequency modulation, and amplitude modulation are mainly calculated for a single note, whereas stimuli featured a melody, these parameters were relatively unimportant, for which reason the interquartile range was omitted. This produced an objective acoustic parameter set, containing 45 parameters (see **Table 4**).

The calculation methods of some important acoustic parameters are as follows:

(1) *Zero-crossing rate:* It is defined as the number of times the audio signal waveform crossing the zero amplitude level during a 1-s interval, and it provides a rough estimator of the dominant frequency component of the signal (Alías et al., 2016).

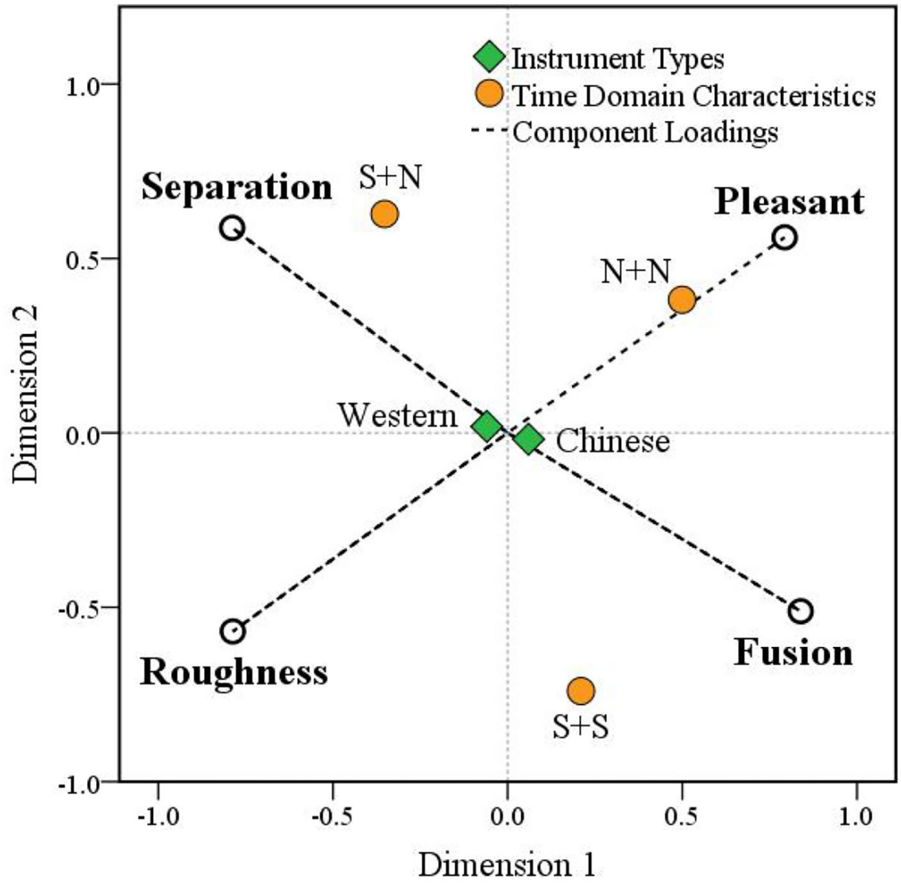


FIGURE 16 | A preference space location map.

(2) *Spectrum centroid:* SC for short, defined as the centroid of spectral energy (Marchetto and Peeters, 2015). It can be defined as the first moment of the amplitude spectrum of the signal frame (the mean value of the frequency position), which represents the geometric center of the spectrum, and the unit is hertz. $f(n)$ is the frequency after ERBfft transformation, and $P[E(n)]$ is the probability value of the spectral energy of each point on the total energy. N is the length of the DFT transform.

$$\mathbf{SpecCent}=\sum_{\mathbf{n=1}}^{\mathbf{N}}\mathbf{f\,(n)\,P\,(E\,(n))} \tag{1}$$

(3) *Spectrum flatness:* It is a measure of the uniformity of the power spectrum frequency distribution. It can be calculated as the ratio of the sub-band geometric average to the arithmetic average (equivalent to the MPEG-7 audio frequency spectrum flatness (ASF) description Character (Grzywczak and Gwardys, 2014).

$$\mathbf{SFM\,(t_m)}=\frac{\left(\prod_{\mathbf{k=1}}^{\mathbf{K}}\mathbf{a_k\,(t_m)}\right)^{\frac{\mathbf{1}}{\mathbf{K}}}}{\frac{\mathbf{1}}{\mathbf{K}}\sum_{\mathbf{k=1}}^{\mathbf{K}}\mathbf{a_k\,(t_m)}} \tag{2}$$

(4) *Harmonic energy:* Harmonic energy is the energy of the signal explained by the harmonic partials (Sharma et al., 2020). It is obtained by summing the energy of the partials detected at a specific time. In the equation, $a_h(t_m)$ is the amplitude and frequency of partial h at time t_m. H partials are ranked by increasing frequency.

$$\mathbf{E_H\,(t_m)}=\sum_{\mathbf{h=1}}^{\mathbf{H}}\mathbf{a_h^2\,\,(t_m)} \tag{3}$$

(5) *Spectral roll-off:* This parameter was proposed by Scheirer and Slaney (1997). It is defined as the frequency $f_c(t_m)$ below, which 95% of the signal energy is contained, where sr/2 is the Nyquist frequency and af is the spectral amplitude at frequency f. In the case of harmonic sounds, it can be shown experimentally that spectral roll-off is related to the harmonic or noise cutoff frequency. The spectral roll-off also reveals an aspect of spectral shape as it is related to the brightness of a sound.

$$\sum_{\mathbf{f=0}}^{\mathbf{f_c(t_m)}}\mathbf{a_f^2\,\,(t_m)}=\mathbf{0.95}\sum_{\mathbf{f=0}}^{\frac{\mathbf{sr}}{\mathbf{2}}}\mathbf{a_f^2\,\,(t_m)} \tag{4}$$

TABLE 4 | An acoustic parameter list.

Classification	Parameter name	Statistics
Time domain	Temporal centroid	Mean, IQR
	Attack time	Mean
	Log attack time	Mean
	Decrease time	Mean
	Effective duration	Mean
	Release time	Mean
	Attack slope	Mean
	Decrease slope	Mean
	Frequency modulation	Mean
	Amplitude modulation	Mean
	Zero-Crossing rate	Mean, IQR
Frequency domain	Spectral centroid	Mean, IQR
	Spectral spread	Mean, IQR
	Spectral decrease	Mean, IQR
	Spectral skewness	Mean, IQR
	Spectral kurtosis	Mean, IQR
	Spectral roll-off	Mean, IQR
	Spectral-flatness measure	Mean, IQR
	Spectral crest measure	Mean, IQR
	Spectral flux	Mean, IQR
	Root mean square energy	Mean, IQR
Harmonic domain	Harmonic energy	Mean, IQR
	Noisiness energy	Mean, IQR
	Tristimulus	Mean, IQR
	Harmonic spectral deviation	Mean, IQR
	odd-to-even ratio	Mean, IQR
	Noisiness	Mean, IQR

In this paper, the Timbre Toolbox (Peeters et al., 2011) and MIRtoolbox (Lartillot and Toiviainen, 2007) were used for feature extraction. The corresponding acoustic parameters were extracted from stimuli in the timbre fusion database, and the acquired data were used to construct a model of timbre fusion.

Model Parameter Fitting

Subjective and objective correlations were adopted in the construction of the fusion model. The subjective label is the mean value of the fusion value, and the objective data are 45 dimensional objective acoustic parameters. This study uses multiple linear regressions, random forest, and multilayer perceptron to predict the subjective degree of fusion. The following is an introduction of the model and parameter settings.

(1) *Multiple linear regression:* We used multiple linear regression (Olive, 2017) to fit the data of the independent variable's 45 dimensional objective acoustic parameters and the degree of fusion of the dependent variable. The criterion of minimizing the mean square error and the gradient descent method are used to determine the linear regression coefficients. Adding Lasso regularization on the basis of standard multiple linear regression makes it easier to make the weight close to 0, which can be used for feature selection (Fonti and Belitser, 2017).

(2) *Random forest:* The random forest is composed of multiple decision trees (Pal, 2005). The root node of the decision tree is randomly selected from the training sample. The objective acoustic parameters of the sample are randomly selected by tree splitting. There is no correlation between multiple decision trees. Sklearn is used in this paper to fit the random forest model (Feurer et al., 2019). In this model, the adjustable parameters include bootstrapping, the maximum number of features for one decision tree, the maximum number of leaf nodes, and the number of decision trees.

We adjusted the parameters for the number of decision trees and the maximum number of features for one decision tree, and we adopted default values for other parameters. The increasing number of decision trees makes the model perform better, but too many trees may cause overfitting. The objective acoustic parameters in this paper are 45 attributes, and there are obvious category divisions and feature correlations between the features, so we set the maximum number of features of the decision tree as 6. The number "6" is determined by experience. If the number of features is too large, the accuracy of the model will be affected. The number of decision trees is determined according to the empirical value and the number of samples, which ranges from 9 to 11 in this paper. A total of 10 decision trees are optimally selected by testing the integration of Chinese and Western instruments and the results of the integrated model. The output result is determined jointly by each decision tree, which is the mean value of the predicted results of the test samples by the 10 decision trees.

(3) *Multilayer perceptron:* The multilayer perceptron consists of an input layer, a hidden layer, and an output layer. The layers are fully connected (Ramchoun et al., 2016). The units between the layers are connected as weight coefficients and biases, and ReLU is used as the activation function. The optimization of model training uses stochastic gradient descent (SGD) (Wu et al., 2020), and the gradient parameter update learning rate was set to 0.001.

To evaluate the accuracy of the prediction results of the model constructed by different algorithms, the goodness of fit R^2 was used as the evaluation index, which is defined as follows (Brook and Arnold, 2018). The *SSR* is the regression sum of squares, *SSE* is the residual sum of squares, and *SST* is the total deviation of squares. In addition, $0 < \text{R} < 1$, the closer R is to 1, the better the prediction result.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum \left(y - \hat{y}\right)^2}{\sum \left(y - \bar{y}\right)^2} \tag{5}$$

Four-fold cross-validation was performed on 518 audio data stimuli. Each time the model was built, 3-folds were taken, and the remaining fold was used for verification. The average value of R^2 was taken as the prediction accuracy of the model. The 259 pieces of Chinese and Western audio data were divided for 4-fold cross-validation, which was the same processing method as described above. The results are as follows (see **Table 5**).

The constructed linear regression model is expressed as follows. Using this model, objective acoustic parameters with an absolute value of regression coefficient >4 are selected

TABLE 5 | Comparison of accuracy of fusion models.

Name	R^2 (Chinese and Western)	R^2 (Chinese)	R^2 (Western)
Linear lasso	0.414	0.541	0.305
Random forest	0.417	0.563	0.363
Multilayer perceptron	0.464	0.573	0.443

to characterize their contribution to fusion. In the following formula, X is an acoustic objective parameter.

The linear model of the fusion of Chinese musical instruments is as follows: where $F_{Chinese}$ is the fusion degree of Chinese musical instruments.

$$\begin{aligned} F_{Chinese} &= 14.2X_{HarmErg} + 7.7X_{SpecCent} - 7.6X_{SpecFlat} \\ &- 6.9X_{NoiseErg} + 5.6X_{ZcrRate} + 4.8X_{SpecRolloff} \end{aligned} \quad (6)$$

The linear model of the fusion of Western musical instruments is as follows: where $F_{Western}$ is the fusion degree of Western musical instruments.

$$\begin{aligned} F_{Western} &= 18.7X_{HarmErg} - 11.6X_{NoiseErg} - 10.9X_{SpecFlat} \\ &- 10.3X_{SpecCrest} - 9.2X_{ZcrRate} - 8.9X_{RMSEEnv} \\ &+ 7.2X_{SpecSpread} \end{aligned} \quad (7)$$

The comprehensive linear model of Chinese and Western musical instruments is as follows: where F_{all} is the fusion of Chinese and Western musical instruments.

$$\begin{aligned} F_{all} &= 13.8X_{HarmErg} - 13.6X_{SpecKurt} + 13.3X_{SpecSkew} \\ &- 11.1X_{NoiseErg} + 10.5X_{SpecCent} - 10.3X_{SpecFlat} \\ &- 7.5X_{SpecCrest} - 5.7X_{ZcrRate} \end{aligned} \quad (8)$$

The parameters in the above equations are the temporal statistical mean of the parameters. The regressors in the equations can be divided into three categories: spectral centroid, spectral roll-off, and zero crossing rate are related to the perceptual brightness and can be classified as brightness factors. Harmonic energy, noisiness energy, and RMS relate to signal energy and can be classified as energy factors. Spectral flatness, Spectral crest, Spectral spread, Spectral skewness, and Spectral kurtosis are related to an Spectral envelope, which can be classified as an Spectral envelope factor.

For Chinese instruments (Equation 6), the spectral centroid, spectral roll-off, and zero crossing rate are positively correlated with the fusion, and these three parameters are brightness factors. The brighter timbre of the dyad, the better the degree of fusion is. This is opposite to the experimental results of Sandell (1995). Sandell's experimental stimulus was western instruments, and the result was that the higher the composite centroid of the spectrum, the worse the fusion. This result shows that the perceptual fusion degree of Chinese instruments is different from that of Western instruments. This may be related to the differences in timbre between Chinese and Western instruments or to the cultural background of the subjects. Previous studies have shown that cultural background is an important factor affecting the timbre with respect to an emotional perception (Wang et al., 2021).

For Western instruments (Equation 7), the zero crossing rate is negatively correlated with the fusion, and this parameter is a brightness factor. That is, the less bright timbre of the dyad, the better the perception of fusion. This is consistent with the experimental results of Sandell (1995). This proves that the higher the composite spectral centroid of western instruments, the worse the fusion.

From the perspective of the energy factor, harmonic energy is positively correlated with fusion, while noisiness energy is negatively correlated with fusion for both Chinese and Western instruments. This shows that the more prominent the musical characteristics of the dyad, the better the perceptual fusion. The more prominent the noise characteristic is, the worse the perception fusion is. This is also one of the main reasons for the worse perception fusion of Chinese-plucked instruments. Plucking instruments produce a large number of dissonant noise components at the moment when fingernails or picks touch the strings. As a result, the fusion of the whole strumming group is poor.

In addition, the RMS of Western instruments is negatively correlated with the fusion. Although RMS is an energy factor, there is a certain relationship between the energy of an instrument's sound (playing intensity) and timbre brightness. When an instrument is played with greater force, more high-frequency components are activated, resulting in a brighter tone. That is, the higher the RMS value, the brighter the tone, the worse the perceptual fusion.

DISCUSSION

Through the statistical processing of experimental data, the following analysis and discussion can be made:

Sustaining and non-sustaining temporal envelopes are important factors that affect the perception attributes of timbre. Moreover, for different timbre attributes, the temporal envelopes have different effects. For the fusion and segregation attributes, the temporal properties have an impact on both Chinese and Western instruments, although these impacts are more pronounced for Chinese instruments than Western instruments. Specifically, a timbre combination with the same temporal envelope has a higher degree of fusion and a higher degree of segregation. The values of fusion from high to low are S + S > N + N > S + N (W: 5.95 > 5.67 > 3.94; C: 5.99 > 5.31 > 4.04), while the values of the degree of segregation are opposite to those of fusion: S + N > N + N > S + S (W: 6.58 > 4.84 > 4.40; C: 6.47 > 5.23 > 4.32).

For the roughness attributes, the temporal properties have an impact on roughness for both Chinese and Western instruments, although the impact is more pronounced for Western instruments than Chinese instruments. Specifically, timbre combinations with more sustaining instruments have higher roughness. The values of roughness from high to low are S + S > S + N > N + N (W: 5.12 > 4.50 > 3.18; C: 4.85 > 4.31 > 3.98). This may be because sustaining instruments contain more beating, which is an important factor that causes roughness. Similarly, for the pleasantness attributes, the temporal properties have an impact on pleasantness for both Chinese and Western instruments. However, the values of pleasantness are exactly opposite to those of roughness: N + N > S + N >S + S (W: 6.54 > 5.25 > 4.65; C: 6.16 > 5.27 > 4.63).

Moreover, through the correlation analysis and multidimensional preference analysis for the four timbre attributes, it is found that the ranking of segregation is opposite to that of fusion. The ranking of roughness is also opposite to that of pleasantness. The results further confirm the above conclusions. These results further support the conclusion of Tardieu and McAdams (2012) and Lembke et al. (2019) that fusion is reduced in the presence of non-sustaining instruments with mixed timbre. Similarly, the results of our manuscript are consistent with the conclusions drawn by Fischer et al. (2021), namely, decreasing temporal differences reduce segregation ratings. In addition, the comparison of the results of these three papers shows that the higher the similarity of timbre, the higher the fusion, and the lower the segregation, and *vice versa*. This is consistent with the conclusion of the multidimensional preference analysis in our manuscript.

From the experimental results, it can be seen that, in most cases, the instrument type (i.e., Chinese instruments or Western instruments) has less influence on the four timbre attributes. However, when the temporal envelope is N + N, the roughness will be affected by the instrument type, and the roughness of Chinese instruments is greater than that of Western instruments.

According to the variance analysis for the four timbre attributes, there is no interaction between the instrument type and the temporal envelope. That is, the difference in timbre perception attributes caused by different temporal envelopes is, basically, the same between Chinese and Western instruments.

According to the experimental data of the four timbre perception attributes, the values of fusion and segregation vary more for different temporal envelopes of Chinese instruments. However, the value of roughness varies more for different temporal envelopes when using Western instruments. That is, fusion and segregation are important attributes to evaluate the timbre combination of Chinese instruments, while roughness is an important attribute to evaluate the timbre combination of Western instruments.

Comparing the models used for analysis, the random forest and multilayer perceptron models are more effective than the linear regression models. For the model of fusion, the best accuracy is 46.4% for Chinese and Western instruments, 57.3% for Chinese instruments, and 44.3% for Western instruments. It shows that these algorithms have some limitations, and the accuracy of the model can be greatly improved. Comparing the models used for analysis, the random forest and multilayer perceptron models are more effective than the linear regression models. These two machine learning algorithms non-linearly fit the data to achieve better performance. Comparing the models of the fusion of Chinese and Western musical instruments, it can be seen that the linear regression model fits Chinese musical instrument fusion better than Western musical instruments. The model could be even more accurate. This is partly because we have a limited amount of data. However, the algorithm that we used was not state of the art enough. All of these factors have some influence on the accuracy of the model. In future research, we will attempt to further increase the amount of data and adopt deep learning algorithms to improve the accuracy of the model.

Overall, the prediction effect of the integration model for Chinese musical instruments is better than that for Western musical instruments. This difference may be related to the distribution of musical instruments. Chinese instruments are more comprehensive and evenly distributed, so the model can learn more stably and achieve effective predictions, while, in the Western fusion dataset, audio data with a high fusion degree have a larger proportion, and the model-learned features are insufficient. This result is also consistent with the distribution characteristics of the perception results from the auditory perception experiment.

Comparing the coefficients of the linear regression models of Chinese musical instruments and Western musical instruments, we can see the contributions of various objective acoustic parameters to timbre fusion. For Chinese musical instruments, the important parameters that affect fusion are harmonic energy, spectrum centroid, spectrum flatness, noise energy, zero crossing rate, and spectrum roll-off. The fusion of Western musical instruments is mainly affected by objective acoustic parameters, such as harmonic energy, noise energy, spectral flatness, spectral crest factor, zero-crossing rate, root mean square energy, and spectrum expansion.

The important objective acoustic parameters of integrated models of Chinese and Western musical instruments are harmonic energy, spectral kurtosis, spectral skewness, noise energy, spectral centroid, spectral flatness, spectral crest factor, and zero crossing rate. These parameters combine the objective acoustic parameters that have made outstanding contributions to the fusion of Chinese musical instruments and Western musical instruments. It also proves the rationality and effectiveness of the model. Comparing the models of Chinese musical instruments and Western musical instruments, their common parameters are harmonic energy, noise energy, and spectral flatness. These parameters are all related to the perceptual consonance of timbre (Wang and Meng, 2013). Therefore, we believe that objective acoustic parameters related to perceptual harmony are important factors that affect timbre fusion.

CONCLUSION

In this paper, the characteristics of the timbre fusion of Chinese and Western instruments were explored, and a subjective evaluation experiment of a timbre perception based on the serial category method was designed and implemented. The effects of time domain characteristics and instrument types on fusion, segregation, roughness, and pleasantness were studied by statistical processing, which included variance analysis, multidimensional preference analysis, correlation analysis, and machine learning algorithms. The differences in the four timbre perception attributes between Chinese and Western instruments were compared. Through carrying out relevant subjective and objective experiments, the following conclusions were obtained.

Sustaining and non-sustaining time domain characteristics are important factors affecting the perception attributes of timbre. Moreover, for different timbre attributes, the time domain characteristics have different effects. According to the experimental data of the four timbre perception attributes, fusion and segregation are important attributes for evaluating the timbre combination of Chinese instruments, while roughness is an important attribute for evaluating the timbre combination of Western instruments. This conclusion further explains why the acoustic theory of symphonic orchestration is mostly based on roughness. For the study of the orchestration theory of Chinese instruments, it is necessary to explore the general rules of timbre fusion for Chinese instruments.

Multiple linear regression, random forest, and multilayer perceptron were used in this paper to construct a set of timbre fusion models for Chinese and Western instruments. The results showed that these models can better predict the timbre fusion attributes. From this research, it was also found that there are some differences between the timbre fusion models for Chinese and Western instruments, which is consistent with the analysis results of subjective experimental data. In addition, the spectrum centroid and spectrum roll-off were found to have an important influence on both the fusion model of Chinese and Western musical instruments. These parameters are all related to the brightness of the tone. Therefore, we can consider the parameter related to timbre brightness as important factors that affect the fusion of Chinese and Western instruments, although the impact is more pronounced in Chinese instruments than Western instruments. The contribution of the above parameters, especially the important parameters of the spectral centroid, was basically consistent with the results of Sandell (1991). However, there is no parameter, such as the attack time, in the regression model of the fusion degree, which was different from previous studies. This may be due to the melody content used in the fusion timbre database used in this experiment. Compared with monophonic audio data, the effect of vibration time on the entire time sequence was less obvious.

In this paper, the research on fusion is still in the exploratory stage, and this work needs to be further improved and supplemented. For example, the amount of data needed to build the database has yet to be expanded. Due to the limitation of data quantity, only a conventional algorithm was implemented in this paper to build the fusion degree model, and it is necessary to adopt deep learning to build the model on the large-scale dataset in later stages. In a follow-up study, we plan to make a special study on the timbre integration of Chinese instruments. In the aspect of database construction, a larger scale timbre fusion stimuli library should be built, and timbre fusion with different harmonies should be discussed. In addition to the timbre combinations of two instruments, the complexities of three or more timbre combinations should be considered. Future experiments should use real instrument sampling so that the research results can be extended to the practice of orchestration of Chinese and Western orchestras. From the perspective of research methods and theory, objective parameters related to timbre fusion should be further explored and analyzed, and the mathematic model should be explained from the perspective of instrument acoustics. It is also necessary to study how to apply relevant models to the orchestration practice and instrument reform of Chinese orchestral music, e.g., the development and construction of computer-aided Chinese orchestration software.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Boards of the Communication University of China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JLi designed the study and conducted acoustic analyses. JLi, YX, SW, JJ, YJ, and JLa collected the data. JLi and YX conducted data analyses and drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Funds for National Natural Science Foundation of China (62101514), Key Laboratory of Ministry of Culture and Tourism (WLBSYS2005), and Key Laboratory of Audio and Video Repair and Evaluation Ministry of Culture and Tourism (2021KFKT004).

ACKNOWLEDGMENTS

The experimental stimuli were produced by Yongzhan Huang from Communication University of China. Can Ran, Huiling Xiang, Jiayu Lin, and Yuqu Chen from Communication University of China participated in the experimental design, and some students from Digital Media Technology and Performing Arts Engineering Technology of Communication University of China participated in the subjective evaluation experiment. We would like to express my sincere gratitude to the above students. The authors would like to express their sincere gratitude to the above students. The music excerpt quoted in the manuscript is from the Spring Festival prelude; the composer is Huanzhi Li. Here, we would like to express our thanks to the composer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.878581/full#supplementary-material

REFERENCES

Alías, F., Socoró, J. C., and Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Appl. Sci.* 6, 143. doi: 10.3390/app6050143

Apel, W. (2003). *The Harvard Dictionary of Music*. Cambridge: Harvard University Press.

Bechtel, G. G. (2019). *Multidimensional Preference Scaling*. Berlin: De Gruyter Mouton.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.

Bregman, A. S., and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* 89, 244.

Bregman, A. S., and Pinker, S. (1978). Auditory streaming and the building of timbre. *Can. J. Psychol.* 32, 19.

Brook, R. J., and Arnold, G. C. (2018). *Applied Regression Analysis and Experimental Design*. Boca Raton: CRC Press. doi: 10.1201/9781315137674

Carterette, E. C., and Kendall, R. A. (1989). Dynamics of musical expression. *J. Acoust. Soc. Am.* 85, S141.

Chon, S. H., and Huron, D. (2014). "Instrument identification in concurrent unison dyads: the effect of timbre saliency," in *Proceedings of the 13th International Conference of Music Perception and Cognition (ICMPC),* Seoul, 289–292.

Chon, S. H., and McAdams, S. (2012). Investigation of timbre saliency, the attention-capturing quality of timbre. *J. Acoust. Soc. Am.* 131, 3433–3433. doi: 10.1121/1.4708879

DeWitt, L. A., and Crowder, R. G. (1987). Tonal fusion of consonant musical intervals: the oomph in Stumpf. *Percept. Psychophys.* 41, 73–84.

EBU-TECH 3253 (2008). *Sound Quality Assessment Material Recordings for Subjective Tests. Users' Handbook for the EBU SQAM CD*. Geneva: EBU.

Fales, C., and McAdams, S. (1994). The fusion and layering of noise and tone: implications for timbre in african instruments. *Leonardo Music J.* 4, 69–77.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2019). "Auto-sklearn: efficient and robust automated machine learning," in *Automated Machine Learning*, eds F. Hutter, L. Kotthoff and J. Vanschoren (Cham: Springer), 113–134. doi: 10.1007/978-3-030-053 18-5_6

Fischer, M., Soden, K., Thoret, E., Montrey, M., and McAdams, S. (2021). Instrument timbre enhances perceptual segregation in orchestral music. *Music Percept. Interdiscip. J.* 38, 473–498. doi: 10.1525/mp.2021.38.5.473

Fonti, V., and Belitser, E. (2017). Feature selection using lasso. *VU Amst. Res. Paper Bus. Anal.* 30, 1–25.

Goodwin, A. W. (1980). An acoustical study of individual voices in choral blend. *J. Res. Music Educ.* 28, 119–128.

Grey, J. (1975). *An Exploration of Musical Timbre*. Doctoral Dissertation published as report STAN-M2. Stanford, CA: University of Stanford.

Grzywczak, D., and Gwardys, G. (2014). "Audio features in music information retrieval," in *International Conference on Active Media Technology*, eds D. Ślęzak, G. Schaefer, S. T. Vuong and Y.-S. Kim (Cham: Springer), 187–199. doi: 10.1007/978-3-319-09912-5_16

ITU-RBS 1770-4 (2015). *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*. Geneva: International Telecommunications Union.

Jiang, W., Liu, J., Zhang, X., Wang, S., and Jiang, Y. (2020). Analysis and modeling of timbre perception features in musical sounds. *Appl. Sci.* 10, 789. doi: 10.3390/app10030789

Jingyu, L. (2013). Subjective preference study on timbre combination of chinese plucked instruments. *Sci. Technol. Perform. Arts* 5, 48–51.

Kendall, R. A., and Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Percept.* 8, 369–404.

Kendall, R. A., and Carterette, E. C. (1993). Identification and blend of timbres as a basis for orchestration. *Contemp. Music Rev.* 9, 51–67.

Kim, N. E. (2018). *The Effects of Timbre on Harmonic Interval Tuning and Perception*. Montreal: McGill University.

Lartillot, O., and Toiviainen, P. (2007). "A Matlab toolbox for musical feature extraction from audio," *in International Conference on Digital Audio Effects, Vol*, Bordeaux, *237*, 244.

Lembke, S. A., and McAdams, S. (2012). "A spectral-envelope synthesis model to study perceptual blend between wind instruments," *Proceedings of the Acoustics 2012 Nantes Conference* (Nantes).

Lembke, S. A., and McAdams, S. (2015). The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica United with Acust.* 101, 1039–1051. doi: 10.3813/AAA.918898

Lembke, S. A., Narmour, E., and McAdams, S. (2013). "Predicting blend between orchestral timbres using generalized spectral-envelope descriptions," in *Proceedings of Meetings on Acoustics ICA2013, Vol. 19* (Montreal: Acoustical Society of America). doi: 10.1121/1.4800054

Lembke, S. A., Parker, K., Narmour, E., and Mcadams, S. (2019). Acoustical correlates of perceptual blend in timbre dyads and triads. *Musicae Sci. 23,* 102986491773180. doi: 10.1177/1029864917731806

Li, X. P. (2020). Reproduction and reconstruction: the concept, method and significance of national orchestral music recording. *Entertain. Technol.* 2020, 15–20. doi: 10.3969/j.issn.1674-8239.2013.05.011

Marchetto, E., and Peeters, G. (2015). "A set of audio features for the morphological description of vocal imitations," in *Proceedings of the 18th International Conference on Digital Audio Effects*, Trondheim.

Martin, F. N., and Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *J. Am. Acad. Audiol.* 11, 64–66. doi: 10.1055/s-0042-1748011

McAdams, S. (1982). "Spectral fusion and the creation of auditory images," in *Music, Mind, and Brain*, ed M. Clynes (*Boston, MA: Springer),* 279–298.

McAdams, S. (1984). *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*. Standford, CA: Stanford university.

McAdams, S. (2019). "Timbre as a structuring force in music," in *Timbre: Acoustics, Perception, and Cognition*, eds K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper and R. R. Fay (Cham: Springer), 211–243. doi: 10.1007/978-3-030-14832-4_8

Melara, R. D., and Marks, L. E. (1990). Interaction among auditory dimensions: timbre, pitch, and loudness. *Percept. Psychophys.* 48, 169–178.

Olive, D. J. (2017). "Multiple linear regression," in *Linear Regression* (Cham: Springer), 17–83. doi: 10.1007/978-3-319-55252-1_2

Pal, M. (2005). Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* 26, 217–222. doi: 10.1080/01431160412331269698

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.* 130, 2902–2916. doi: 10.1121/1.3642604

Ramchoun, H., Idrissi, M. A. J., Ghanou, Y., and Ettaouil, M. (2016). Multilayer perceptron: architecture optimization and training. *Int. J. Interact. Multim. Artif. Intell.* 4, 26–30. doi: 10.9781/ijimai.2016.415

Reuter, C. (2003). "Stream segregation and formant areas," in *Proceedings of the European Society for the Cognitive Sciences of Music Conference (ESCOM)*, Hanover, 329–331.

Rossetti, D. (2016). "The qualities of the perceived sound forms: a morphological approach to timbre composition, in *International Symposium on Computer Music Multidisciplinary Research* (Cham: Springer), 259–283. doi: 10.1007/978-3-319-67738-5_16

Sandell, G. (1989a). "Perception of concurrent timbres and implications for orchestration," in *Proceedings, International Computer Music Conference* (Ann Arbor, MI), 268–272.

Sandell, G. (1989b). "Effect of spectrum and attack properties on the evaluation of concurrently sounding timbres," in *Program of the 118th Meeting of the Acoustical Society of America*, St. Louis, S59.

Sandell, G. J. (1991). *Concurrent Timbres in Orchestration: A Perceptual Study of Factors Determining "Blend"*. (Doctoral dissertation), Evanston, IL: Northwestern University.

Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. *Music Percept.* 13, 209–246.

Scheirer, E., and Slaney, M. (1997). "Construction and evaluation of a robust multifeature speech/music discriminator," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2* (Munich: IEEE), 1331–1334.

Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Appl. Acoust.* 158, 107020. doi: 10.1016/j.apacoust.2019.107020

Shields, R., and Roger, K. (2004). "The relation of timbre to dissonance and spectral fusion," in *Proceedings of the 8th International Conference on Music Perception and Cognition* (Evanston, IL), 596–9.

Siedenburg, K., Saitis, C., McAdams, S., Popper, A. N., and Fay, R. R. (Eds.). (2019). Timbre: Acoustics, Perception, and Cognition, Vol. 69. Russia: Springer. doi: 10.1007/978-3-030-14832-4

Tardieu, D., and McAdams, S. (2012). Perception of dyads of impulsive and sustained instrument sounds. *Music Percept.* 30, 117–128. doi: 10.1525/mp.2012.30.2.117

Wang, X., and Meng, Z. (2013). The evaluation method of sound concord of chinese national plucked stringed instruments. *Acta Acoust. 38*, 486–492.

Wang, X., Wei, Y., Heng, L., and McAdams, S. (2021). A cross-cultural analysis of the influence of timbre on affect perception in western classical music and chinese music traditions. *Front. Psychol.* 12, 732865. doi: 10.3389/fpsyg.2021.732865

Wang, X., Wu, F., and Li, Y. (2016). *Listening Training and Subjective Evaluation of Sound Quality*. Beijing: Communication University of China Press.

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V., and Zhu, Z. (2020). "On the noisy gradient descent that generalizes as SGD," in *International Conference on Machine Learning* (Toronto: PMLR), 10367–10376.

Zhu, J., Liu, J., and Li, Z. (2018). "Research on loudness balance of Chinese national orchestra instrumental sound," in *Proceedings of the 2018 National Acoustical Congress of Physiological Acoustics, Psychoacoustics, Music Acoustics* (Beijing), 34–35.

Zihou, M. (2008). *Experimental Psychological Method for Subjective Evaluation of Sound Quality*. Beijing: National Defence of Industry Press.

Conflict of Interest: JJ was employed by China Digital Culture Group Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Digital Health

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fdgth.2022.789980

Evaluating the COVID-19 Identification ResNet (CIdeR) on the INTERSPEECH COVID-19 From Audio Challenges

Alican Akman[1*†], Harry Coppock[1†], Alexander Gaskell[1], Panagiotis Tzirakis[1], Lyn Jones[2] and Björn W. Schuller[1,3]

[1] GLAM–Group on Language, Audio, and Music, Imperial College London, London, United Kingdom, [2] Department of Radiology, North Bristol NHS Trust, Bristol, United Kingdom, [3] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany

OPEN ACCESS

Edited by:
Jayaraman J. Thiagarajan,
Lawrence Livermore National Laboratory (DOE), United States

Reviewed by:
Shekh Md Mahmudul Islam,
University of Dhaka, Bangladesh
Riaz Ullah Khan,
University of Electronic Science and Technology of China, China

***Correspondence:**
Alican Akman
a.akman21@imperial.ac.uk

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Health Informatics,
a section of the journal Frontiers in Digital Health

Received: *05 October 2021*
Accepted: *31 May 2022*
Published: *07 July 2022*

Citation:
Akman A, Coppock H, Gaskell A, Tzirakis P, Jones L and Schuller BW (2022) Evaluating the COVID-19 Identification ResNet (CIdeR) on the INTERSPEECH COVID-19 From Audio Challenges.
Front. Digit. Health 4:789980.
doi: 10.3389/fdgth.2022.789980

Several machine learning-based COVID-19 classifiers exploiting vocal biomarkers of COVID-19 has been proposed recently as digital mass testing methods. Although these classifiers have shown strong performances on the datasets on which they are trained, their methodological adaptation to new datasets with different modalities has not been explored. We report on cross-running the modified version of recent COVID-19 Identification ResNet (CIdeR) on the two Interspeech 2021 COVID-19 diagnosis from cough and speech audio challenges: ComParE and DiCOVA. CIdeR is an end-to-end deep learning neural network originally designed to classify whether an individual is COVID-19-positive or COVID-19-negative based on coughing and breathing audio recordings from a published crowdsourced dataset. In the current study, we demonstrate the potential of CIdeR at binary COVID-19 diagnosis from both the COVID-19 Cough and Speech Sub-Challenges of INTERSPEECH 2021, ComParE and DiCOVA. CIdeR achieves significant improvements over several baselines. We also present the results of the cross dataset experiments with CIdeR that show the limitations of using the current COVID-19 datasets jointly to build a collective COVID-19 classifier.

Keywords: COVID-19, computer audition, digital health, deep learning, audio

1. INTRODUCTION

The current coronavirus pandemic (COVID-19), caused by the severe-acute-respiratory-syndrome-coronavirus 2 (SARS-CoV-2), has infected a confirmed 126 million people and resulted in 2,776,175 deaths (WHO)[1]. Mass testing schemes offer the option to monitor and implement a selective isolation policy to control the pandemic without the need for regional or national lockdown (1). However, physical mass testing methods, such as the Lateral Flow Test (LFT) have come under criticism since the tests divert limited resources from more critical services (2, 3) and due to suboptimal diagnostic accuracy. Sensitivities of 58 % have been reported for self-administered LFTs (4), unacceptably low when used to detect active virus, a context where high sensitivity is essential to prevent the reintegration into society

[1] As of 29th March 2021 https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

of falsely reassured infected test recipients (5). In addition to mass testing, radar remote life sensing technology offers non-contact applications to combat COVID-19 including heart rate tracking, identity authentication, indoor monitoring and gesture recognition (6).

Investigating the potential for digital mass testing methods is an alternative approach, based on findings that suggest a biological basis for identifiable vocal biomarkers caused by SARS-CoV-2's effects on the lower respiratory track (7). This has recently been backed up by empirical evidence (8). Efforts have been made to collect and classify a range of different modality audio recordings of COVID-19-positive and COVID-19-negative individuals and several datasets have been released that use applications to collect the breath and cough of volunteer individuals. Examples include the "COUGHVID" (9), "Breath for Science"[2], "Coswara" (10), COVID-19 sounds[3], and 'CoughAgainstCovid' (11). In addition, to focus the attention of the audio processing community onto the task of binary classification of COVID-19 from audio, two INTERSPEECH competitions: the INTERSPEECH 2021 Computational Paralinguists Challenge (ComParE)[4] (12) with its COVID-19 Cough and Speech Sub-Challenges, and Diagnosing COVID-19 using acoustics (DiCOVA)[5] (13) have been organized with this focus as their challenge.

Several studies have been published that propose machine learning-based COVID-19 classifiers exploiting distinctive sound properties between positive and negative cases to classify these datasets. Brown et al. (14) and Ritwik et al. (15) demonstrate that simple machine learning models perform well in these relatively small datasets. In addition, deep neural networks are exploited in Laguarta et al. (16), Pinkas et al. (17), Imran et al. (18), and Nessiem et al. (19) with proven performance at the COVID-19 detection task. Although there are works that try to combine different modalities computing the representations separately, Coppock et al. (20) (CIdeR) proposes an approach computing joint representation of a number of modalities. The adaptability of this approach to different types of datasets has not to our knowledge been explored or reported.

To this end, we propose a modified version of COVID-19 Identification ResNet (CIdeR), a recently developed end-to-end deep learning neural network optimized for binary COVID-19 diagnosis from cough and breath audio (20), which is applicable to common datasets with further modalities. We present the competitive results of CIdeR to the two COVID-19 cough and speech Challenges of INTERSPEECH 2021, ComParE and DiCOVA. We also investigate the behavior of a strong COVID-19 classifier across different datasets by running cross dataset experiments with CIdeR. We describe the limitations of current COVID-19 classifiers with these experiments regarding the ultimate goal of proposing a universal COVID-19 classifier.

2. METHODS

2.1. Model

CIdeR (20) is a 9 layer convolutional residual network. A schematic detailing of the model can be seen in **Figure 1**. Each layer or block consists of a stack of convolutional layers with Rectified Linear Units (ReLUs). Batch normalization (21) also features in the residual units, acting as a source of regularization and supporting training stability. A fully connected layer with sigmoid activation terminates the model yielding a single logit output which can be interpreted as an estimation of the probability of COVID-19. As detailed in **Figure 1** the network is modified to be compatible with a varying number of modalities, for example, if a participant has provided cough, deep breathing, and sustained vowel phonation audio recordings, they can be stacked in a depth wise manner and passed through the network as a single instance. We use PyTorch library in python to implement CIdeR and baseline models.

2.2. Pre-processing

At training time, a window of s-seconds, which was fixed at 6 s for these challenges, is sampled from the audio recording randomly. If the audio recording is less than s-seconds long, the sample is padded with repeated versions of itself. The sampled audio is then converted into Mel-Frequency Cepstral Coefficients (MFCCs) resulting in an image of width s * the sample rate and height equal to the number of MFCCs. Three data augmentation steps are then applied to the sample. First, the pitch of the recording is randomly shifted, secondly, bands of the Mel spectrogram are masked in the time and Mel coefficient axes and finally, Gaussian noise is added. At test time, the sampled audio recording is chunked into a set of s-second clips and processed in parallel. The mean of the set of logits is then returned as the final prediction.

2.3. Baselines

The DiCOVA team ran baseline experiments for the track 1 (coughing) sub-challenge; only the best performing (MLP) model's score was reported. For the track 2 (deep breathing/vowel phonation/counting) sub-challenge, however, baseline results were not provided. Baseline results were provided for the ComParE challenge but only Unweighted Average Recall (UAR) was reported rather than Area Under Curve of the Receiver Operating Characteristics curve (ROC-(AUC)). To allow comparison across challenges, we created new baseline results for the ComParE sub-challenges and the DiCOVA Track 2 sub-challenge, using the same baseline methods described for the DiCOVA Track 1 sub-challenge. The three baseline models applied to all four sub-challenge datasets were Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forrest (RF), where the same hyperparameter configurations that were specified in the DiCOVA baseline algorithm was used (13).

To provide a baseline comparison for the CIdeR track 2 results, we built a multimodal baseline model. We followed a similar strategy with the provided DiCOVA baseline algorithm, while extracting the features for each modality. Rather than individual training for different models, we developed an algorithm that concatenates input features from separate

[2] https://www.breatheforscience.com
[3] https://www.covid-19-sounds.org/en/
[4] http://www.compare.openaudio.eu/
[5] https://dicova2021.github.io/

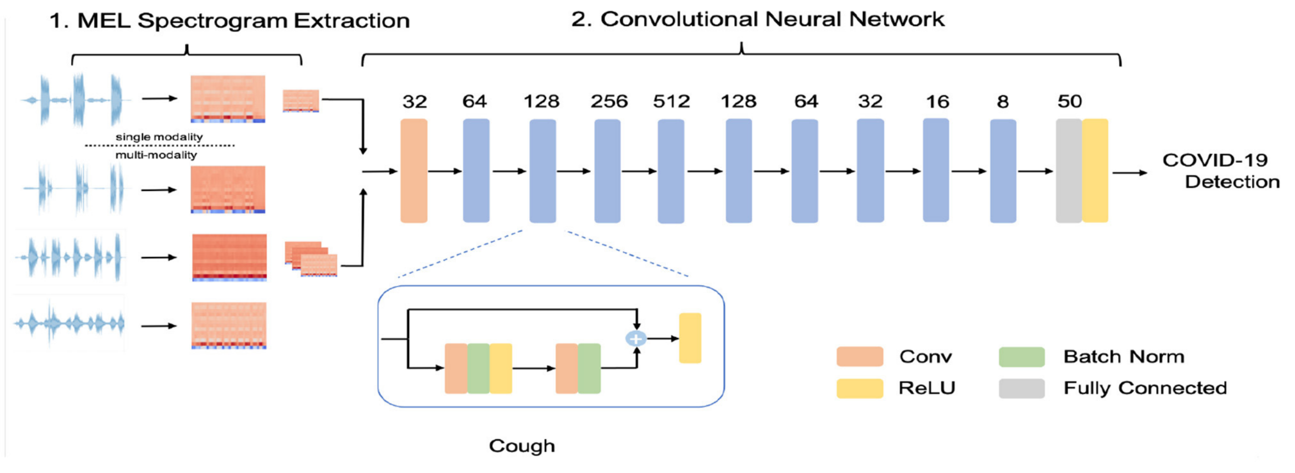


FIGURE 1 | A schematic of the COVID-19 Identification ResNet, (CIdeR). The figure shows a blow-up of a residual block, consisting of convolutional, batch normalization, and Rectified Linear Unit (ReLU) layers.

TABLE 1 | ComParE sub-challenge dataset splits.

	CCS			CSS		
#	Train	Val	Test	Train	Val	Test
COVID-19-postive	71	48	39	72	142	94
COVID-19-negative	215	183	169	243	153	189
Total	286	231	208	315	295	283

Values specify the number of audio recordings, not the number of participants.

TABLE 2 | DiCOVA sub-challenge dataset splits.

	Track-1		Track-2	
#	Train + Val	Test	Train + val	Test
COVID-19-postive	75	blind	60	21
COVID-19-negative	965	blind	930	188
Total	1,040	234	990	209

The test set labels were withheld by the DiCOVA team, contestants had to submit predictions for each test case, on which a final AUC was returned.

modalities. Then, this combined feature set was fed to the baseline models: LR, MLP, and RF.

We used 39 dimensional MFCCs as our feature type to represent the input sounds. For LR, we used Least Square Error (L2) as a penalty term. For MLP, we used a single hidden layer of size 25 with a Tanh activation layer and L2 regularization. The Adam optimiser and a learning rate of 0.0001 was used. For RF, we built the model with 50 trees and split based on the gini impurity criterion.

3. DATASETS

3.1. ComParE

ComParE hosted two COVID-19 related sub-challenges, the COVID-19 Cough Sub-Challenge (CCS) and the COVID-19 Speech Sub-Challenge (CSS). Both CCS and CSS are subsets of the crowd sourced Cambridge COVID-19 sound database (14, 22). CCS consists of 926 cough recordings from 397 participants. Participants provided 1–3 forced coughs resulting in a total of 1.63 h of recording. CSS is made up of 893 recordings from 366 participants totalling 3.24 h of recording. Participants were asked to recite the phrase *"I hope my data can help manage the virus pandemic"* in their native language 1–3 times. The train-test splits for both sub-challenges are detailed in **Table 1**.

3.2. DiCOVA

Once again, DiCOVA hosted two COVID-19 audio diagnostic sub-challenges. Both sub-challenge datasets were subsets of the crowd sourced Coswara dataset (10). The first sub-challenge, named Track-1, comprised of a set of 1,274 forced cough audio recordings from 1,274 individuals totalling 1.66 h. The second, Track-2, was a multi-modality challenge, where 1,199 individuals provided three separate audio recordings; deep breathing, sustained vowel phonation, and counting from 1 to 20. This dataset represented a total of 14.9 h of recording. The train-test splits are detailed in **Table 2**.

4. RESULTS AND DISCUSSION

The results from the array of experiments with CIdeR and the 3 baseline models are detailed in **Table 3**. CIdeR performed strongly across all four sub-challenges, achieving AUCs of 0.799 and 0.787 in the DiCOVA Track 1 and 2 sub-challenges, respectively, and 0.732 and 0.787 in the ComParE CCS and CSS sub-challenges. In the DiCOVA cough sub-challenge, CIdeR significantly outperformed all three baseline models based on 95 % confidence intervals calculated following (23), and in the DiCOVA breathing and speech sub-challenge it achieved a

TABLE 3 | Results for CIdeR and a range of baseline models for 4 sub-challenges across the DiCOVA and ComParE challenges.

	Sub-challenge*	CIdeR	Baseline		
			LR	MLP	RF
DiCOVA	Track 1**	**0.799** ± 0.058	-	0.699 ± 0.068	-
	Track 2	0.786 ± 0.057	0.647 ± 0.014	0.684 ± 0.072	0.776 ± 0.063
ComParE	CCS	0.732 ± 0.068	0.722 ± 0.069	0.765 ± 0.065	0.753 ± 0.066
	CSS	**0.787** ± 0.060	0.583 ± 0.072	0.656 ± 0.070	0.628 ± 0.070

Testing is performed on the held-out test fold once final model decisions have been made on the validation sets. The Area Under Curve of the Receiver Operating Characteristics curve (AUC(-ROC)) is displayed. A 95% confidence interval is also shown following (23). CIdeR scores which are statistically higher than the best baseline results with a 95% confidence are in bold. The three baseline models are Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forrest (RF). All baseline models were trained on MFCC features.

**Track 1: coughing, Track 2: deep breathing + vowel phonation + counting, CCS: coughing, CSS: speech—" hope my data can help managethe virus pandemic". ** As the demographics were not provided for the Track 1 test set, when calculating the AUC confidence intervals, it was assumed that there was an equal number of COVID-19-positive and COVID-19-negative recordings.*

higher AUC although the improvement over the baselines was not significant. Conversely, while CIdeR performed significantly better than all three baseline models in the ComParE speech sub-challenge based on 95 % confidence intervals calculated following (23), it performed no better than baseline in the ComParE cough sub-challenge. One can speculate that this may have resulted from the small dataset sizes favoring the more classical machine learning approaches which do not need as much training data.

4.1. Limitations

A key limitation with both the ComParE and DiCOVA COVID-19 challenges is the size of the datasets. Both datasets contain very few COVID-19-positive participants. Therefore, the certainty in results is limited and this is reflected in the large 95 % confidence intervals detailed in **Table 3**. This issue is compounded by the demographics of the datasets. As detailed in Brown et al. (14) and Muguli et al. (13) for the ComParE datasets and the DiCOVA datasets, respectively, not all demographics from society are represented evenly—most notably, there is poor coverage of age and ethnicity and both datasets are skewed toward the male gender.

In addition, the crowd-sourced nature of the datasets introduces some confounding variables. Audio is a tricky sense to control. It contains a lot of information about the surrounding environment. As both datasets were crowd-sourced, there could have been correlations between ambient sounds and COVID-19 status, for example, sounds characteristic of hospitals or intensive care units being more often present for COVID-19-positive recordings compared to COVID-19-negative recordings. As the ground truth labels for both datasets were self reported, presumably the participants knew at the time of recording whether they had COVID-19 or not. One could postulate that the individuals who knew they were COVID-19-positive might have been more fearful than COVID-19-negative participants at the time of recording, an audio characteristic known to be identifiable by machine learning models (24). Therefore, the audio features which have been identified by the model may not be specific audio biomarkers for the disease.

We note that both the DiCOVA Track 1 and ComParE CCS sub-challenges were cough recordings. Therefore, there was an opportunity to utilize both training sets. Despite having access to both the DiCOVA and ComParE datasets, training on the two datasets together did not yield a better performance on either of the challenges' test sets. Additionally, a model which performed well on one of the challenges test sets would see a marked drop in performance on the other challenge's test set. We run cross dataset experiments to analyse this effect further. For these experiments, we also included the COUGHVID dataset (9) in which COVID-19 labels were assigned by experts and not as a results of clinically validated test. The results in **Table 4** show that the trained models for each dataset do not generalize well and perform poorly on excluded datasets. This is a worrying find, as it suggests that audio markers which are useful in COVID-19 classification in one dataset are not useful or present in the other dataset. This agrees with the concerns presented in Coppock et al. (25) that current COVID-19 audio datasets are plagued with bias, allowing for machine learning models to infer COVID-19 status, not by audio biomarkers uniquely produced by COVID-19, but by other correlations in the dataset such as nationality, comorbidity and background noise.

TABLE 4 | The results for cross dataset experiments.

	Test set		
Train set	DiCOVA	ComParE	COUGHVID
DiCOVA	0.799	0.554	0.464
ComParE	0.512	0.732	0.552
COUGHVID	0.395	0.518	0.566
All	0.673	0.717	0.531

5. FUTURE WORK

One of the most important next steps is to collect and evaluate machine learning COVID-19 classification on a larger dataset that is more representative of the population. To achieve optimal ground truth, audio recordings should be collected at the time that the Polymerase Chain Reaction (PCR) test is taken, before the result is known. This would ensure full blinding of the participant to their COVID-19 status and exclude any environmental audio biasing in the dataset. The Cycle Threshold (CT) of the PCR test should also be recorded, CT correlates with viral load (26) and therefore would enable researchers to determine the model's classification performance to the disease at varying viral loads. This relationship is critical in assessing the usefulness of any model in the context of a mass testing scheme, since the ideal model would detect a viral load lower than the level that confers infectiousness[6]. Finally, studies similar to Bartl-Pokorny et al. (8), directly comparing acoustic features of COVID-19-positive and COVID-19-negative participants should be conducted on all publicly available datasets.

6. CONCLUSION

Cross-running CIdeR on the two 2021 Interspeech COVID-19 diagnosis from cough and speech audio challenges has demonstrated the model's adaptability across multiple modalities. With little modification, CIdeR achieves competitive results in all challenges, advocating the use of end-2-end deep learning models for audio processing thanks to their flexibility and strong performance. Cross dataset experiments with CIdeR have revealed the technical limitations of the datasets for joint usage that prevent from creating a common COVID-19 classifier.

[6] Seventy-third SAGE meeting on COVID-19, 17th December 2020.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AA and HC designing alternative methods, literature analysis, performing and analyzing experiments, manuscript preparation, editing, and drafting manuscript. AG and PT designing and performing experiments. BS and LJ drafting manuscript and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

FUNDING

The support of the EPSRC Center for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged along with the UKRI CDT in Safe & Trusted AI. The authors further acknowledge funding from the DFG (German Research Foundation) Reinhart Koselleck-Project AUDI0NOMOUS (grant agreement No. 442218748) and the Imperial College London Teaching Scholarship.

REFERENCES

1. Peto J, Carpenter J, Smith GD, Duffy S, Houlston R, Hunter DJ, et al. Weekly COVID-19 testing with household quarantine and contact tracing is feasible and would probably end the epidemic. *R Soc Open sci*. (2020) 8:201546. doi: 10.1098/rsos.200915
2. Wise J. Covid-19: concerns persist about purpose, ethics, and effect of rapid testing in Liverpool. *BMJ*. (2020) 371:m4690. doi: 10.1136/bmj.m4690
3. Holt E. Newsdesk COVID-19 testing in Slovakia. *Lancet Infect Dis*. (2021) 21:32. doi: 10.1016/S1473-3099(20)30948-8
4. Mahase E. Covid-19: innova lateral flow test is not fit for test and release strategy, say experts. *BMJ*. (2020) 371:m4469. doi: 10.1136/bmj.m4469
5. Moyse KA. Lateral flow tests need low false positives for antibodies and low false negatives for virus. *BMJ*. (2021) 372:n90. doi: 10.1136/bmj.n90
6. Islam SMM, Fioranelli F, Lubecke VM. Can radar remote life sensing technology help combat COVID-19? *Front Commun Netw*. (2021) 2:648181. doi: 10.3389/frcmn.2021.648181
7. Quatieri T, Talkar T, Palmer J. A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open J Eng Med Biol*. (2020) 1:203–6. doi: 10.1109/OJEMB.2020.2998051
8. Bartl-Pokorny KD, Pokorny FB, Batliner A, Amiriparian S, Semertzidou A, Eyben F, et al. The voice of COVID-19: acoustic correlates of infection in sustained vowels. *J Acoust Soc Am*. (2020) 149:4377. doi: 10.1121/10.0005194
9. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset: a corpus for the study of large-scale cough analysis algorithms. *Sci Data*. (2020) 8:156. doi: 10.1038/s41597-021-00937-4
10. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, R N, et al. Coswara-a database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: *Proceedings Interspeech*. Shanghai (2020). p. 4811–5.
11. Bagad P, Dalmia A, Doshi J, Nagrani A, Bhamare P, Mahale A, et al. Cough against COVID: evidence of COVID-19 signature in cough sounds. *arXiv*. (2020). doi: 10.48550/arXiv.2009.08790
12. Schuller BW, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, et al. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. *arXiv:2102.13468*. (2021) doi: 10.21437/Interspeech.2021-19
13. Muguli A, Pinto L, Nirmala R, Sharma N, Krishnan P, Ghosh PK, et al. DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. *arXiv [Preprint]*. (2021). arXiv: 2103.09148. doi: 10.48550/ARXIV.2103.09148
14. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: *Proceedings of Knowledge Discovery and Data Mining*. (2020). p. 3474–84.
15. Ritwik KVS, Kalluri SB, Vijayasenan D. COVID-19 patient detection from telephone quality speech data. *arXiv*. (2020). doi: 10.48550/arXiv.2011.04299
16. Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J Eng Med Biol*. (2020) 1:275–81. doi: 10.1109/OJEMB.2020.3026928
17. Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 detection from voice. *IEEE Open J Eng Med Biol*. (2020) 1:268–74. doi: 10.1109/OJEMB.2020.3026468
18. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz S, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv*. (2020). doi: 10.1016/j.imu.2020.100378
19. Nessiem MA, Mohamed MM, Coppock H, Gaskell A, Schuller BW. Detecting COVID-19 from breathing and coughing sounds using deep neural networks. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. Aveiro: IEEE (2021). p. 183–8.
20. Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller BW. End-2-End COVID-19 detection from breath & cough audio. *BMJ Innovations*. (2021). doi: 10.1136/bmjinnov-2021-000668
21. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of International Conference on Machine Learning. vol. 37*. Lille (2015). p. 448–56.
22. Han J, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, et al. Exploring automatic COVID-19 diagnosis via voice and symptoms from

crowdsourced data. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON: IEEE (2021).

23. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. (1982) 143:29–36. doi: 10.1148/radiology.143.1.7063747
24. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller BW, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE (2016). p. 5200–4.
25. Coppock H, Jones L, Kiskin I, Schuller BW. COVID-19 detection from audio: seven grains of salt. *Lancet Digit Health*. (2021) 3:e537–8. doi: 10.1016/S2589-7500(21)00141-2
26. Singanayagam A, Patel M, Charlett A, Bernal JL, Saliba V, Ellis J, et al. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. *Eurosurveillance*. (2020) 25:2001483. doi: 10.2807/1560-7917.ES.2020.25.32.2001483

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 17 October 2022
DOI 10.3389/fpsyg.2022.1026116

Cochlear-implant Mandarin tone recognition with a disyllabic word corpus

Xiaoya Wang[1,2†], Yefei Mo[3†], Fanhui Kong[4], Weiyan Guo[5], Huali Zhou[4], Nengheng Zheng[4], Jan W. H. Schnupp[6], Yiqing Zheng[1,5,7]* and Qinglin Meng[3]*

[1]The First Clinical Medical College of Jinan University, Guangzhou, China, [2]Department of Otolaryngology, Guangzhou Women and Children's Medical Center, Guangzhou, China, [3]Acoustics Laboratory, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China, [4]The Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, [5]Department of Hearing and Speech Science, Xin Hua College of Sun Yat-sen University, Guangzhou, China, [6]Department of Biomedical Sciences and Department of Neuroscience, City University of Hong Kong, Hong Kong, Hong Kong SAR, China, [7]Department of Otolaryngology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

EDITED BY
Shengchen Li,
Xi'an Jiaotong-Liverpool University,
China

REVIEWED BY
Yue Zhang,
Oticon Medical, France
Jinqiu Sang,
Institute of Acoustics (CAS), China
Xi Shao,
Nanjing University of Posts and
Telecommunications, China

*CORRESPONDENCE
Qinglin Meng
mengqinglin@scut.edu.cn
Yiqing Zheng
zhengyiq@mail.sysu.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

RECEIVED 23 August 2022
ACCEPTED 28 September 2022
PUBLISHED 17 October 2022

CITATION
Wang X, Mo Y, Kong F, Guo W, Zhou H, Zheng N, Schnupp JWH, Zheng Y and Meng Q (2022) Cochlear-implant Mandarin tone recognition with a disyllabic word corpus. *Front. Psychol.* 13:1026116. doi: 10.3389/fpsyg.2022.1026116

Despite pitch being considered the primary cue for discriminating lexical tones, there are secondary cues such as loudness contour and duration, which may allow some cochlear implant (CI) tone discrimination even with severely degraded pitch cues. To isolate pitch cues from other cues, we developed a new disyllabic word stimulus set (Di) whose primary (pitch) and secondary (loudness) cue varied independently. This Di set consists of 270 disyllabic words, each having a distinct meaning depending on the perceived tone. Thus, listeners who hear the primary pitch cue clearly may hear a different meaning from listeners who struggle with the pitch cue and must rely on the secondary loudness contour. A lexical tone recognition experiment was conducted, which compared Di with a monosyllabic set of natural recordings. Seventeen CI users and eight normal-hearing (NH) listeners took part in the experiment. Results showed that CI users had poorer pitch cues encoding and their tone recognition performance was significantly influenced by the "missing" or "confusing" secondary cues with the Di corpus. The pitch-contour-based tone recognition is still far from satisfactory for CI users compared to NH listeners, even if some appear to integrate multiple cues to achieve high scores. This disyllabic corpus could be used to examine the performance of pitch recognition of CI users and the effectiveness of pitch cue enhancement based Mandarin tone enhancement strategies. The Di corpus is freely available online: https://github.com/BetterCI/DiTone.

KEYWORDS

cochlear implants, Mandarin tone, pitch contour, loudness contour, lexical tone

1. Introduction

Linguists define "lexical tone" as the phenomenon when two syllables which differ in their pitch contour but are otherwise identical can have different meanings. Mandarin Chinese is a tonal language in which each syllable has four typical tones, each has a characteristic pitch contour. By convention, Tone 1 has a high-flat pitch, Tone 2 a rising pitch, Tone 3 is falling-then-rising in a relatively low pitch range, and Tone 4 has a falling pitch. Linguistic meaning can be distinguished by these four tones. The register and range of pitch contours vary among utterances and persons. Psychoacoustical studies have shown that pitch-related acoustic cues are complex and manifest within multiple features in both temporal and spectral domains of sounds (Schnupp et al., 2011; Oxenham, 2018). Normal hearing (NH) listeners of tonal languages can use pitch cues to distinguish lexical tones robustly even when acoustic signals are degraded by environmental noise, low-fidelity playback, human speech production variability, etc. In contrast, for most cochlear implant (CI) recipients, lexical tone perception is still challenging (Lu et al., 2022), and performance varies significantly across recipients and in environments (Chang et al., 2016; Liu et al., 2017; Mao and Xu, 2017; Li et al., 2018; Tang et al., 2019). This is perhaps unsurprising given CI recipients' weaker and more variable abilities to extract pitch cues from acoustic signals (Tao et al., 2015; Mok et al., 2017; Vandali et al., 2017). Limitations in pitch extraction can occur on multiple stages of the CI supplied auditory system, from the device's signal processing strategy through peripheral auditory neural processing, all the way to auditory cortical processing and cognition (Zhang, 2019; Zhou et al., 2022).

However, speech researchers have long recognized that pitch cues are not the only acoustic cues that could be used for lexical tone discrimination. Secondary cues, such as amplitude contour (Whalen and Xu, 1992; Kuo et al., 2008), duration (Fu and Zeng, 2000; Xu et al., 2002; Yang et al., 2017), and spectral (timbre) contour (Liang, 1963), tend to covary with the pitch cues and may be useful when pitch cues are significantly degraded. Thus, loudness and timbre can occasionally serve as alternative cues in tasks which are classically thought of as pitch-dependent, including lexical tone and musical melody perception, and this has been observed in both NH and, more strongly, in CI listeners (McDermott et al., 2008; Cousineau et al., 2010; Luo et al., 2014, 2019). Manipulating the timbre contour for tone enhancement in speech is problematic since changing the spectral shape would likely affect the formant structure of the manipulated speech. In contrast, the amplitude contour could be manipulated to co-vary more strongly with the fundamental frequency (F0) contour to facilitate Mandarin tone perception with CIs (Luo and Fu, 2004; Kim et al., 2021), and some studies indicated that these kinds of strategies can be effective in actual CI users (Ping et al., 2017; Meng et al., 2018).

The confounds created by co-varying pitch and non-pitch cues to the Mandarin tone also imply that previous Mandarin tone recognition experiments with CI participants, which simply used naturally recorded speech stimuli, will have measured the ability to utilize some combination of several types of acoustic cues. These experiments therefore cannot give an independent estimate of the CI user's ability to use specifically pitch cues to discriminate lexical tones. Indeed, secondary cues can be quite reliable and could be strong enough to lead to ceiling effects in tone identification. This could perhaps explain why some previous tests of lexical tone enhancement strategies found no or only little improvement (Han et al., 2009; Vandali et al., 2017).

Pitch and duration cues for lexical tone perception have been studied by Peng et al. (2009, 2017). They orthogonally manipulated F0 (pitch) contour, intensity (loudness) contour, and duration, to study how the interaction between these cues influence the perception of English intonation (Peng et al., 2009) or Chinese lexical tone (Peng et al., 2017). Covarying cues generally caused better results than conflicting cues for CI listeners, but no significant difference was found for NH listeners. In the tone perception study by Peng et al. (2017), the pitch contour and duration of the second syllable /jing/ in the disyllabic word /yǎn jing/ were manipulated to generate two alternative meanings: /yǎn jīng/ (Tone 1) means eyes, and /yǎn jìng/ (Tone 4) means eyeglasses. Using disyllables rather than monosyllables for tests of this nature is preferable because Chinese monosyllables tend to have many homophonic meanings, while the meaning of disyllables tends to be much more unambiguously determined by the tone, creating less uncertainty in the participants' mind. While Peng et al. (2017) did study pitch and duration cues for lexical tone, they did not investigate the role of the amplitude contour, even though this is a powerful secondary cue.

In order to dissociate the contributions of pitch and non-pitch cues to tone recognition, we developed a set of Mandarin syllables where the pitch cues of target tones vary independently of secondary loudness and duration cues. This was inspired by Peng et al. (2009, 2017). In our preliminary study (Meng et al., 2018), we manipulated the pitch contour and the loudness contour of the second syllable /shi/ of a disyllable /lǎo shi/ independently to generate speech sounds that could be interpreted to convey one of three possible word meanings: /lǎo shī/ (Tone 1) means "teacher", /lǎo shí/ (Tone 2) means "well-behaved", and /lǎo shì/ (Tone 4) means "always". Different weighting strategies were found in four CI participants, in that two participants relied more on loudness cues, and the other two participants relied more on pitch cues. The influence of loudness (or amplitude) contour on CI tone recognition has been demonstrated in several studies (Luo and Fu, 2004; Meng et al., 2016, 2018; Ping et al., 2017; Kim et al., 2021).

In this study, a much larger CI participant cohort was used to expand the findings, and more disyllables were carefully selected to generate an expanded speech corpus. The disyllable corpus includes five disyllabic words, which were decomposed and resynthesized into words whose primary (pitch) and secondary (loudness) cues varied independently. The syllables with flat tone were resynthesized to have either a high-flat, a rising, or a falling pitch contour. The pitch-manipulated monosyllables were then amplitude-modulated by three loudness gain functions, which are flat, rising, or falling. These resynthesized syllables formed a stimulus set of 270 disyllabic words (denoted by "Di"), each having a distinct meaning depending on the perceived tone. Thus, listeners who hear the primary pitch cue clearly will often hear a different meaning from listeners who are insensitive to the pitch cue and must rely on the secondary cue given by the loudness contour. The new stimulus sets thus make it possible to evaluate the contribution of pitch contour cues to lexical tone perception in CIs in isolation.

A tone recognition experiment was carried out with the new disyllabic set Di as well as with a set of natural monosyllabic recordings ("Mono") (Wei et al., 2004) so that responses could be directly compared. The Mono stimuli consist of monosyllabic words with four tones which were recorded naturally from a female speaker. As noted before, natural Mandarin recordings contain pitch cues as well as co-varying secondary cues that can both help listeners identify lexical tones. In contrast, while Di includes only three tones (Tone 1, 2, and 4), its pitch contours and loudness contours were manipulated to vary independently, so that secondary loudness cues were no longer reliable, and pitch cue performance can therefore be assessed in isolation. In order to train the participants to use pitch contour as much as possible, participants were given trial-by-trial feedback of whether their answers were correct according to the pitch contour. Since pitch contour is the primary cue on which NH Mandarin speakers overwhelmingly rely for tone recognition, we scored a word as "correctly identified" when the listener reported the meaning of the word that corresponds to the lexical tone given by the pitch contour, irrespective of (secondary) loudness contour cues values.

2. Materials and methods

2.1. Participants

In total, seventeen CI recipients and eight NH listeners participated in this study. The CI recipients were recruited in Guangdong Province, and the NH listeners (age 18–32) were college students from two universities (South China University of Technology and Sun Yat-Sen University) in Guangdong Province. Further details about the CI recipients are shown in Table 1. The selection criteria for these CI participants were: (1) severe-to-profound sensorineural hearing loss in both ears, (2) more than 1-year CI use experience, (3) self-reported efficient speech communication ability without the use of visual cues, and (4) capable of cooperating to complete the experiment. Note that most of the participants were from Southern China, and some of them may use a Southern Chinese dialect to complete their day-to-day conversation with family members, such as Cantonese, so Mandarin may not have been their "mother tongue". Participation was compensated and all participants gave informed consent in accordance with the Shenzhen University's ethical review board.

2.2. Stimuli

The new disyllables corpus consists of five main disyllabic words (i.e., /Lǎo Shī/, /Róng Huā/, /Shè Jī/, /Píng Fāng /, and /Huā Xiāng/), each recorded from 2 speakers (1 male and 1 female) in a studio at a sampling rate of 22,050 Hz and resampled using MATLAB resample.m to a sampling rate of 16,000 Hz. The STRAIGHT toolbox (17/09/2005) (Kawahara et al., 2004) was used to manipulate the pitch and loudness contours of the recorded signals. Firstly, the recorded words were decomposed according to a source-filter model to extract the excitation and spectral envelope related information. Then all the syllables with Tone 1 (i.e., the flat tone) were transformed to have 9 different F0 contours (changing linearly with time) including 3 flat contours, 3 rising contours, and 3 falling contours. Specific settings are shown in the Figure 1. For the female speaker, the 3 flat contours are 300, 250, and 200 Hz, respectively; the 3 rising contours are 150 to 300, 250, and 200 Hz, respectively; and the 3 falling contours are 300 to 220, 170, and 120 Hz, respectively. For the male speaker, the 3 flat contours are 200, 170, and 130 Hz, respectively; the 3 rising contours are 100 to 220, 180, and 150 Hz, respectively; and the 3 falling contours are 180 to 140, 110, and 80 Hz, respectively. These F0 values and frequency steps were selected with reference to the range of naturally recorded Chinese lexical tone frequency variations (Traunmuller and Eriksson, 1993; Moore and Jongman, 1997). The transformation was done by changing the F0 of the excitation signal accordingly and keeping the spectral envelope parts unchanged. This kept almost all information other than the pitch contour unchanged in the resynthesized signals. Finally, the amplitude of the voiced portion of each pitch-modified monosyllable was multiplied by three gain functions (i.e., 0 dB flat, −10 to +10 dB rising, and +10 to −10 dB falling) to generate different loudness contours. Figures 1A,B shows some examples of how the new disyllables were generated from the original recordings.

Permuting the 9 pitch contours with the 3 loudness contours, we generated 27 stimuli from each of the ten original disyllabic words (five for each gender), all having the same duration but differing independently in pitch and loudness contours. Thus, we obtained 270 stimuli (5 original disyllabic words × 2 speakers × 9 pitch contours × 3 loudness contours) in total. These 270 disyllabic stimuli formed our new Mandarin tone perception test stimulus set. Among the 270 disyllabic tokens, 90 tokens have the same pitch contours and loudness

TABLE 1 Participant demographic and device information.

Participant	Gender	Age range (yr)	CI experience (yr)	CI processor (R: Right; L: Left)	Etiology
C21	F	31–35	7	R: Cochlear CP900	Drug-induced
C28	F	36–40	11	R: Cochlear N6	Ototoxicity
C30	F	21–25	1	R: Cochlear Freedom	Unknown
C34	M	11–15	13	R: Med-El OPUS2	Genetic
C36	M	16-20	L:4	L: Med-El OPUS2	Virus infection
			R:14	R: Med-El OPUS2	*Virus infection*
C37	M	11–15	10	L: Cochlear N6	Jaundice
C38	F	6–10	5	L: Med-El OPUS1	Pregnancy infection
C39	M	6–10	7	R: Cochlear N5	Unknown
C40	F	11–15	8	R: Cochlear Freedom	Unknown
C41	F	11–15	9	R: Cochlear CP900	Unknown
C42	M	21–25	18	R: Cochlear Freedom	Gentamicin allergy
C43	M	11–15	8	R: Med-El OPUS2	Ototoxicity
C44	F	31–35	8	R: Nurotron NSP560b	Progressive hearing loss
C45	M	11–15	10	R: Med-El OPUS2	Genetic
C46	F	21–25	1	L: Med-El OPUS2	Unknown
C47	F	16–20	10	L: Med-El OPUS2	Ototoxicity
C48	F	6–10	6	R: Med-El OPUS2	Ototoxicity

contours (both contours are high-flat, rising or falling, denoted by “Cov”), whereas the rest 180 have different pitch contour and loudness contours (denoted by “Conf”).

The synthesized syllables could be identified as one of the 15 disyllabic words shown in Figure 1C. It organizes them according to whether the second syllable has Tone 1, Tone 2, or Tone 4. Note that all the words created in this manner are common, easily understood, and easily distinguished words of Mandarin. Their English meanings are also shown in Figure 1C.

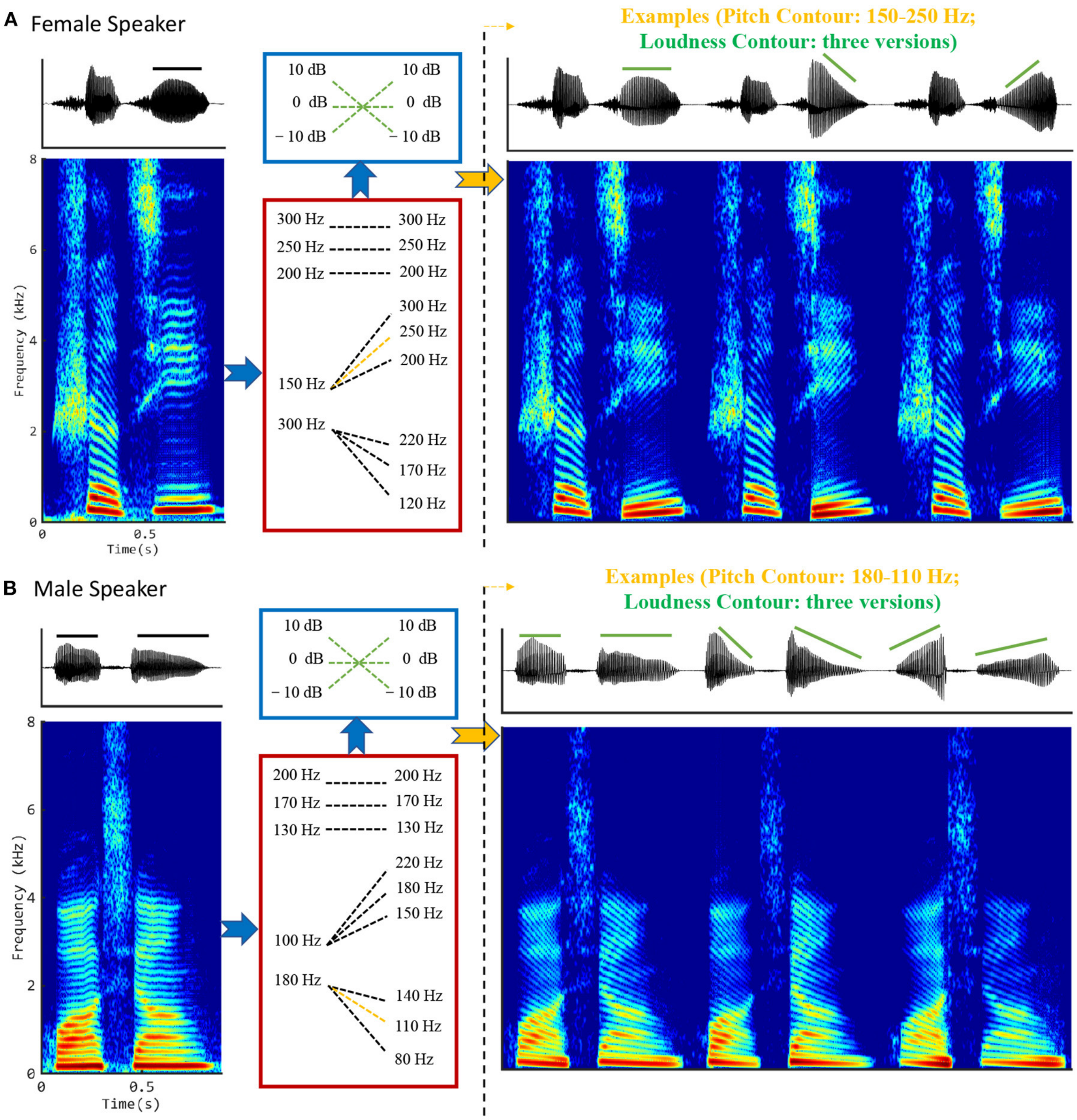
An existing stimulus set of naturally produced monosyllables (Wei et al., 2004) was used for comparison. It includes 100 tokens (25 monosyllabic words, each having four tone patterns) pronounced by a female speaker. For convenience, the disyllabic stimulus set generated in this study is noted as “Di” and the monosyllable set by Wei et al. (2004) is noted as “Mono”. Note that the Mono stimulus set consists entirely of unaltered recordings of naturally spoken Mandarin, and pitch and non-pitch cues to lexical tone will therefore naturally co-vary in the Mono stimulus set. In contrast, the Di stimuli are resynthesized so that pitch and loudness cues to lexical tone vary independently by design.

2.3. Procedure

For each participant, the 270 Di stimuli were divided in a random order into 6 sessions, each with 45 stimuli. Between the third and fourth Di sessions, a test session with the 100 monosyllables from Mono was conducted, with all stimuli in a random order. The session order is shown in Figure 2A. The sound levels of all stimuli were normalized to have the same root-mean-square amplitude. Each stimulus was presented in one trial through an audio interface (Focusrite Scarlett 2i4) and a loudspeaker (Yamaha HS5I) at a sound pressure level of about 70 dBA in a sound-proof room. For the Di trials, a three-alternative-forced-choice (3AFC; T1, 2, or 4) task was used; for the Mono trials, a 4AFC was used (T1, 2, 3, or 4). In each trial, three or four buttons including the Chinese characters and Mandarin Pinyin were shown in a graphic user interface for the subjects to select using a mouse, and the correctness of the subject’s choice (according to the pitch-tone) was shown as green (correct) and red (incorrect) colors in another user interface element.

2.4. Statistical analysis methods

A Wilcoxon signed rank test was used to compare within-subject conditions; a Wilcoxon rank sum test was used to compare between-subject conditions; a Holm-Bonferroni correction was used for multi-pair comparison; and a Spearman’s rank correlation analysis was used to quantify correlations between performance and CI hearing experience. In the result figures, the raw percentage correct scores are



C Words in the new disyllabic stimulus set “Di”

Chinese	Pinyin (T1)	English	Chinese	Pinyin (T2)	English	Chinese	Pinyin (T4)	English
老师	Lǎo Shī	teacher	老实	Lǎo Shí	well-behaved	老是	Lǎo Shì	always
绒花	Róng Huā	velvet flower	荣华	Róng Huá	glory	融化	Róng Huà	melt
射击	Shè Jī	shooting	涉及	Shè Jí	involve	设计	Shè Jì	design
平方	Píng Fāng	square	平房	Píng Fáng	bungalow	平放	Píng Fàng	lie
花香	Huā Xiāng	flower fragrance	滑翔	Huá Xiáng	gliding	画像	Huà Xiàng	portrait

FIGURE 1
Illustrations of disyllabic stimulus generation. **(A)** The word /Shè Jī/ spoken by a female; **(B)** /Huā Xiāng/ spoken by a male. The left column shows the original waveforms and spectrograms. The middle column shows the strategies for manipulating the pitch and loudness contours. As shown, although the loudness-manipulation strategies are consistent for both male and female, the pitch-manipulation parameters are different between genders, reflecting the fact that females generally have higher pitched voices than males. The right column shows waveforms and spectrograms for /Shè Ji/ with a rising pitch contour (150–250 Hz) and three different loudness contours and /Huà Xiàng/ with falling pitch contours (180–110 Hz) and three different loudness contours. **(C)** Words in the new disyllabic stimulus set “Di”.

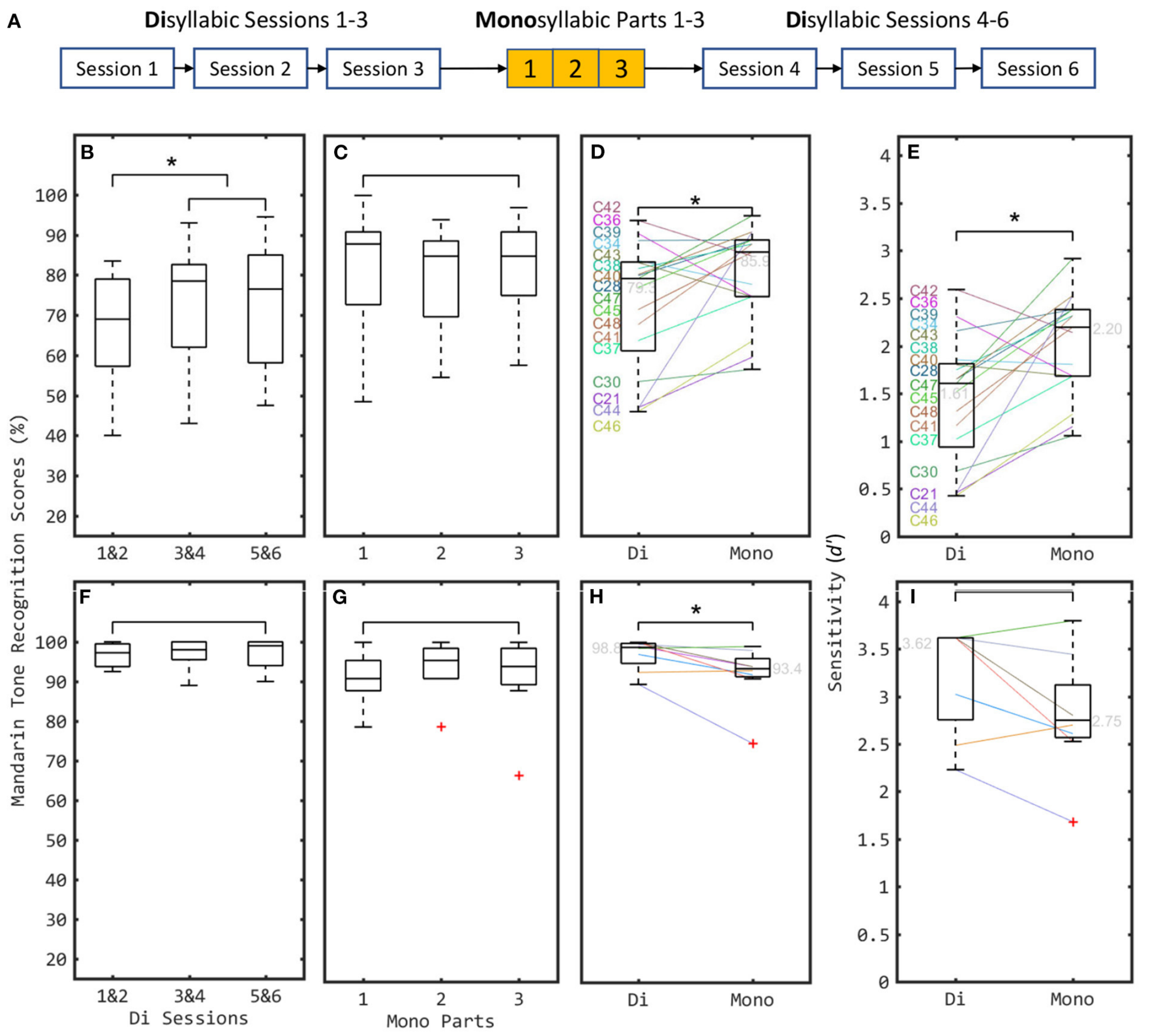


FIGURE 2
Mandarin tone recognition the new disyllabic stimulus set (denoted by "Di") and the old monosyllabic set (denoted by "Mono") for 17 CI **(B–E)** and 8 NH **(F–I)** participants. **(A)** Session procedure of experiment. Both Di sessions and Mono session were equally grouped or divided into three groups for analyzing the training effects of pitch-contour based correctness feedback. **(B,F)** Comparing results with three subgroups of Di. **(C,G)** Comparing results with three parts of Mono. **(D,H)** Individual results (shown by colored lines) and boxplots of the scores with Di and Mono stimulus sets, respectively. **(E,I)** Sensitivity index transferred from D&H. Significant differences between different conditions are illustrated by asterisks. Red plus signs indicate outliers, i.e., data exceeding 1.5 times the interquartile range. They are included in the formal analysis.

shown for simplicity, but to make the results from Di 3AFC and those from Mono 4AFC tests quantitatively comparable, irrespective of their differing chance % correct levels, sensitivity index (d-prime, d') values were computed and statistical tests were carried out on the d' values. The dprime.mAFC function from the psyphy library of the R programming language was used for this conversion. The mapping between percentage scores and d' can also be found in Hacker and Ratcliff (1979).

3. Results

3.1. Training effects

Feedback was given in each trial based on whether the response was correct according to the pitch cue of the stimulus. This encouraged the subjects to use pitch-contour information to do the task. For the CI subjects, the median scores pooled over Di Sessions 1 & 2 were significantly lower than those for

Sessions 3 & 4 ($Z = -2.771$, $p < 0.01$, $n = 17$, Wilcoxon signed rank test) and 5 & 6 ($Z = -2.699$, $p < 0.01$, $n = 17$). No significant difference was found when comparing the pooled median scores from Di Sessions 3 & 4 against 5 & 6 ($Z = -0.466$, $p = 0.641$). Also, no significant difference was found between the median scores obtained with three parts of Mono ($Z = -1.434$, -0.035 and -1.846, respectively, $p > 0.05$ for all comparisons, $n = 17$) (see Figures 2B,C). For the NH subjects, the median scores between three subgroups of Di and between three parts of Mono showed no significant difference ($Z = -2.521$, -0.542, 0.000, -1.511, 0.000, and -1.121, respectively, $p > 0.05$ for all comparisons, $n = 8$) (see Figures 2F,G). Therefore, a significant training effect was found over the first two sessions of Di with CIs. The performance reaches a ceiling from session 3 onwards. Consequently, the results from Di Sessions 3, 4, 5, and 6 were pooled to compute the performance scores for both CI and NH cohorts in the Di task.

3.2. CI vs. NH

The Mandarin tone recognition scores for both Di and Mono stimulus sets are summarized in Figures 2D,H. The median scores of the CI participants (79.3% for Di and 85.9% for Mono) were significantly lower than those of the NH participants (98.8% for Di and 93.4% for Mono) [$Z = -2.521$ (Di) and -2.240 (Mono), $p < 0.05$ for two comparisons, $n = 25$, Wilcoxon rank sum test, Holm-Bonferroni corrected]. NH listeners recognized the words from both stimulus sets with general good scores (see Figures 2H, 3A). The only difficulty for NH with Mono is they sometimes (26.0%) identified the Tone 3 as Tone 2. For Di, Tone 3 was not included, so this confusion was not examined. What's more, in the Di stimulus set, where pitch and loudness cues often diverged, the primary cue (pitch) clearly dominated for NH listeners, as NH listeners were hardly ever misled by conflicting loudness cues. In contrast, CI users scored more poorly, particularly in the tests involving the Di speech material, where accurate pitch coding is particularly important.

3.3. Di vs. Mono

Indeed, for the CI cohort, the median performance with Di (79.3%, $d' = 1.61$) was significantly lower than with Mono (85.9 %, $d' = 2.02$) ($Z = -2.911$, $p = 0.004$, $n = 17$, Wilcoxon signed rank test on d' values, see Figure 2E). In contrast, for the NH cohort, the median score with Di (98.3%) and the scores of most (6/8) participants was higher than those with Mono (see Figure 2H), even though this median d' difference was not statistically significant (3.62 with Di, and 2.75 with Mono) ($Z = -1.823$, $p > 0.05$, $n = 8$, see Figure 2I. Thus, Di was more difficult than Mono for CI users, as expected given the at times conflicting secondary cues.

3.4. Dominant cues for CIs

In Figures 3A,B, we also show the results of CI listeners using the disyllabic stimuli subdivided according to whether the pitch and loudness cues were "Covarying" or "Conflicting". CI users performed significantly better in covarying conditions than in conflicting conditions (see Figure 3A). The median score in the covarying condition was significantly higher than that in the conflicting condition ($Z = -2.215$, $p < 0.05$, $n = 17$, Wilcoxon signed rank test) (Figure 3B). When comparing Mono with the subgroups of Di, the median score with Mono was significantly higher than the score for the Conflicting ($Z = -3.006$, $p < 0.05$, $n = 17$, Wilcoxon signed rank test, Holm-Bonferroni corrected) but did not differ significantly from the Covarying stimulus trial results ($Z = -1.728$, $p > 0.05$). These results indicate that secondary cues were used by most CI users for tone recognition.

3.5. Correlations with CI listening experience

Figure 3C shows the correlations between the tone recognition scores with two stimulus sets and the CI subjects' listening experience. No significant correlation was found between the tone recognition scores obtained with the two stimulus sets ($r = 0.160, p = 0.539$, Spearman's rank correlation analysis). With the Mono stimulus set, no significant correlation was found between tone recognition results and CI listening experience ($r = 0.179, p = 0.492$). With Di stimulus set, however, a highly significant correlation was found between tone recognition results and the amount of CI listening experience ($r = 0.601, p = 0.011$). In the CI cohort, subjects with longer experience generally also have an earlier implantation age. A significant (albeit somewhat weaker) correlation was also found between tone recognition results with Di and their implantation ages ($r = -0.537, p = 0.026$).

4. Discussion

Many studies have shown that Chinese CI users have reasonably good Mandarin tone recognition in quiet environments, usually higher than 60% on average, and higher than 90% for some star participants (Wang et al., 2011, 2012; Tao et al., 2015; Gu et al., 2017; Mao and Xu, 2017; Vandali et al., 2017; Li et al., 2018). However, all these experiments used stimulus sets of naturally produced sound recordings, in which secondary cues, such as loudness contour and syllable duration, can also contribute to a CI user's tone recognition, and pitch contours are not the only cues. Therefore, it is hard to attribute a CI participant's performance in Mandarin tone recognition specifically to the strengths or weaknesses of their pitch encoding, even if pitch cues are generally acknowledged to

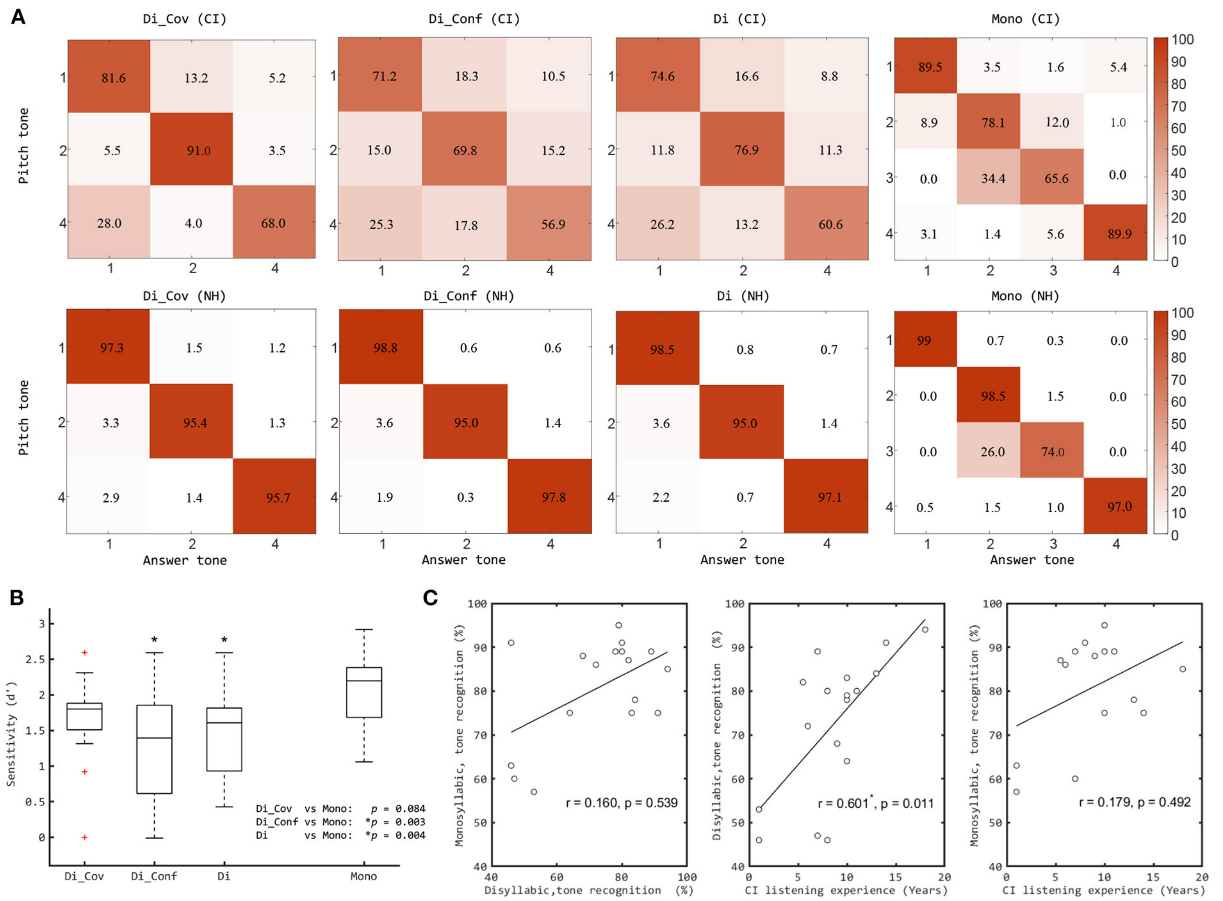


FIGURE 3
(A) Confusion matrices for pitch tone in CI and NH listeners using different corpus (the number represents the tone recognition percentage scores). **(B)** Boxplots of Mandarin tone recognition percentage scores of CI listeners based sensitivity index with sub-conditions (Cov and Conf) of "Di" conditions, compared with "Mono" conditions. The significant differences between Di conditions and Mono are illustrated by the asterisks. **(C)** Correlations between Mandarin tone recognition scores with "Mono" and "Di" and the CI participants' listening experience. The significant correlation is marked by an asterisk. Di, the new disyllabic stimulus set; Mono, the old monosyllabic stimulus set; Cov, covarying pitch and loudness contours; Conf, conflicting pitch and loudness contours. Red plus signs indicate outliers (like that in Figure 2).

dominate tone perception in NH listeners (a fact also confirmed in this study). Furthermore, multiple cues may lead to ceiling effects in performance, which make it difficult to evaluate the effectiveness of pitch-based tone enhancement strategies (Vandali et al., 2017).

Our new disyllabic stimulus set isolates pitch cues from secondary cues by eliminating duration cues and varying amplitude contour cues orthogonally to pitch cues. Results from CI users revealed a substantial drop in median tone recognition scores when they were tested with our new stimuli in comparison to the existing monosyllabic stimulus set in which multiple cues covaried (see Figure 2D). The tone recognition scores with both Di and Mono were much higher for NH listeners than for CI users. This indicates that considerable shortcomings remain in the encoding of pitch cues for tone recognition through CIs. Furthermore, the tone recognition performance of CI users was better when secondary cues co-varied with the pitch cues compared to when these varied independently. This discrepancy was not found in NH listeners (see Figure 3A). These observations can be explained if we assume that the pitch and amplitude cues to Mandarin tone are weighted differently in NH and CI listeners. While NH listeners appear to rely on pitch cues almost exclusively, some CI users who have difficulty using pitch cues (i.e., poor tone recognition in Conf conditions) may learn to rely more on secondary cues instead. The fact that in the DI corpus pitch and secondary cues vary independently makes it possible to determine the extent to which individual CI users are able to rely on primary pitch or secondary loudness contour cues respectively when attempting to identify lexical tones.

Furthermore, the scores with the new stimulus set correlated strongly and significantly with the CI participants' implantation ages and listening experience, in contrast with the scores obtained with the older stimulus set which conflates multiple cues, and which therefore cannot accurately assess the users' ability to discriminate pitch cues for tone recognition. Thus, the ability to utilize pitch cues for tone recognition tasks improves both with earlier implantation and longer hearing experience with CIs (see Figure 3C). However, NH listeners recognized the new disyllabic tones more accurately than the monosyllabic tones, which might benefit from the context of pitch between the two syllables, and the removal of the Tone 3 (falling-then-rising), which is easily confused with Tone 2 (rising) (Figure 3A). In addition, the naturalness of the stimuli could perhaps be somewhat compromised by the fact that the tones of the disyllabic words are synthetic. However, the STRAIGHT method used is generally capable of synthesizing quite naturally sounding speech samples. Interested readers familiar with the sound of Mandarin can of course download the Di speech samples from the github repository and judge for themselves how natural they sound. In any event, the fact that NH cohorts were able to score very highly with the Di corpus, and no worse than with the Mono corpus which consisted of natural recordings (Figures 2F,G), suggests that the naturalness of the Di stimuli is at least adequate to facilitate highly accurate word recognition among native Mandarin speakers, giving confidence that the stimuli are adequate for the intended purpose in audiological assessment.

An important application of the new Di stimulus sets is to reduce the confounds and ceiling effects that can be caused by the secondary cues, and which can plague the evaluation of some tone enhancement strategies (Vandali et al., 2017). In the light of our findings, it seems likely that CI users with poorer pitch coding may compensate by weighting loudness and duration cues more heavily, which would mask the true extent of their pitch coding deficits. Some authors have sought to reduce the ceiling effects by adding noise (Wei et al., 2004; Gu et al., 2017; Mao and Xu, 2017; Vandali et al., 2017). Understanding speech in noise is a challenge that both NH and CI listeners often have to contend with. However, the addition of noise may mask both pitch and loudness contour cues in complex ways that will depend on the type of background noise and may be hard to predict. It would therefore be very useful to conduct speech-in-noise recognition experiments with stimulus sets like the one developed here, which make it possible to study the relative effects of noise on pitch and loudness cue processing for lexical tone independently.

5. Conclusion

A new Mandarin tone corpus consisting of five main disyllabic words from two speakers was developed in this study. In this corpus, there is no reliable secondary cue that could be used by listeners to facilitate the pitch-contour based tone recognition (i.e., loudness contours change independently of pitch contours). When compared to NH listeners, CI users had poorer pitch cue encoding, and their tone recognition performance was significantly influenced by the "missing" or "confusing" secondary cues with this corpus. The corpus could be used to examine the performance of pitch recognition of CI users and the effectiveness of pitch cue enhancement based Mandarin tone enhancement strategies. Listeners with longer CI listening experiences tend to get higher scores of tone recognition with this corpus.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Shenzhen University's Ethical Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

QM and YZ contributed to conception and design of the study. XW, FK, and WG carried out the experiment and organized the database. YM and HZ performed the statistical analysis. XW and YM wrote the first draft of the manuscript. NZ and JS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the Guangdong Basic and Applied Basic Research Foundation Grant (2020A1515010386 and 2022A1515011361) and Science and Technology Program of Guangzhou (202102020944).

Acknowledgments

We would like to thank all the research volunteers that generously donated their time to participate in this study.

Thanks to Chuanyi Chen and his colleagues for help with the disyllabic speech recording, to Fan-Gang Zeng for providing their monosyllabic stimulus set.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Chang, Y.-P., Chang, R. Y., Lin, C.-Y., and Luo, X. (2016). Mandarin tone and vowel recognition in cochlear implant users: effects of talker variability and bimodal hearing. *Ear Hear.* 37, 271. doi: 10.1097/AUD.0000000000000265

Cousineau, M., Demany, L., Meyer, B., and Pressnitzer, D. (2010). What breaks a melody: perceiving F0 and intensity sequences with a cochlear implant. *Hear. Res.* 269, 34–41. doi: 10.1016/j.heares.2010.07.007

Fu, Q.-J., and Zeng, F.-G. (2000). Identification of temporal envelope cues in chinese tone recognition. *Asia Pac. J. Speech Lang. Hear.* 5, 45–57. doi: 10.1179/136132800807547582

Gu, X., Liu, B., Liu, Z., Qi, B., Wang, S., Dong, R., et al. (2017). A follow-up study on music and lexical tone perception in adult Mandarin-speaking cochlear implant users. *Otol. Neurotol.* 38, e421–e428. doi: 10.1097/MAO.0000000000001580

Hacker, M. J., and Ratcliff, R. (1979). A revisted table of d' for m-alternative forced choice. *Percept. Psychophys.* 26, 168–170. doi: 10.3758/BF03208311

Han, D., Liu, B., Zhou, N., Chen, X., Kong, Y., Liu, H., et al. (2009). Lexical tone perception with hiresolution and hiresolution 120 sound-processing strategies in pediatric Mandarin-speaking cochlear implant users. *Ear Hear.* 30, 169. doi: 10.1097/AUD.0b013e31819342cf

Kawahara, H., Banno, H., Irino, T., and Zolfaghari, P. (2004). "Algorithm amalgam: morphing waveform based methods, sinusoidal models and straight," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Montreal: IEEE), 1–13. doi: 10.1109/ICASSP.2004.1325910

Kim, S., Chou, H.-H., and Luo, X. (2021). Mandarin tone recognition training with cochlear implant simulation: amplitude envelope enhancement and cue weighting. *J. Acoust. Soc. Am.* 150, 1218–1230. doi: 10.1121/10.0005878

Kuo, Y.-C., Rosen, S., and Faulkner, A. (2008). Acoustic cues to tonal contrasts in Mandarin: implications for cochlear implants. *J. Acoust. Soc. Am.* 123, 2815–2824. doi: 10.1121/1.2896755

Li, Y.-L., Lin, Y.-H., Yang, H.-M., Chen, Y.-J., and Wu, J.-L. (2018). Tone production and perception and intelligibility of produced speech in Mandarin-speaking cochlear implanted children. *Int. J. Audiol.* 57, 135–142. doi: 10.1080/14992027.2017.1374566

Liang, Z. (1963). The auditory basis of tone recognition in standard chinese. *Acta Physiol. Sin.* 26, 85–92.

Liu, H., Peng, X., Zhao, Y., and Ni, X. (2017). The effectiveness of sound-processing strategies on tonal language cochlear implant users: a systematic review. *Pediatr. Investig.* 1, 32–39. doi: 10.1002/ped4.12011

Lu, H.-P., Lin, C.-S., Wu, C.-M., Peng, S.-C., Feng, I. J., and Lin, Y.-S. (2022). The effect of lexical tone experience on english intonation perception in Mandarin-speaking cochlear-implanted children. *Medicine* 101, e29567. doi: 10.1097/MD.0000000000029567

Luo, X., and Fu, Q.-J. (2004). Enhancing chinese tone recognition by manipulating amplitude envelope: implications for cochlear implants. *J. Acoust. Soc. Am.* 116, 3659–3667. doi: 10.1121/1.1783352

Luo, X., Masterson, M. E., and Wu, C.-C. (2014). Contour identification with pitch and loudness cues using cochlear implants. *J. Acoust. Soc. Am.* 135, EL8–EL14. doi: 10.1121/1.4832915

Luo, X., Soslowsky, S., and Pulling, K. R. (2019). Interaction between pitch and timbre perception in normal-hearing listeners and cochlear implant users. *J. Assoc. Res. Otolaryngol.* 20, 57–72. doi: 10.1007/s10162-018-00701-3

Mao, Y., and Xu, L. (2017). Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants. *Int. J. Audiol.* 56(Suppl 2), S23–S30. doi: 10.1080/14992027.2016.1219073

McDermott, J. H., Lehr, A. J., and Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychol. Sci.* 19, 1263–1271. doi: 10.1111/j.1467-9280.2008.02235.x

Meng, Q., Zheng, N., and Li, X. (2016). Loudness contour can influence Mandarin tone recognition: vocoder simulation and cochlear implants. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 641–649. doi: 10.1109/TNSRE.2016.2593489

Meng, Q., Zheng, N., Mishra, A. P., Luo, J. D., and Schnupp, J. W. (2018). "Weighting pitch contour and loudness contour in Mandarin tone perception in cochlear implant listeners," in *Interspeech* (Hyderabad), 3768–3771. doi: 10.21437/Interspeech.2018-1245

Mok, M., Holt, C. M., Lee, K., Dowell, R. C., and Vogel, A. P. (2017). Cantonese tone perception for children who use a hearing aid and a cochlear implant in opposite ears. *Ear Hear.* 38, e359–e368. doi: 10.1097/AUD.0000000000000453

Moore, C. B., and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *J. Acoust. Soc. Am.* 102, 1864–1877. doi: 10.1121/1.420092

Oxenham, A. J. (2018). How we hear: the perception and neural coding of sound. *Annu. Rev. Psychol.* 69, 27–50. doi: 10.1146/annurev-psych-122216-011635

Peng, S.-C., Lu, H.-P., Lu, N., Lin, Y.-S., Deroche, M. L., and Chatterjee, M. (2017). Processing of acoustic cues in lexical-tone identification by pediatric cochlear-implant recipients. *J. Speech, Lang. Hear. Res.* 60, 1223–1235. doi: 10.1044/2016_JSLHR-S-16-0048

Peng, S.-C., Lu, N., and Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiol. Neurotol.* 14, 327–337. doi: 10.1159/000212112

Ping, L., Wang, N., Tang, G., Lu, T., Yin, L., Tu, W., et al. (2017). Implementation and preliminary evaluation of "c-tone": a novel algorithm to improve lexical tone recognition in Mandarin-speaking cochlear implant users. *Cochlear Implants Int.* 18, 240–249. doi: 10.1080/14670100.2017.1339492

Schnupp, J., Nelken, I., and King, A. (2011). *Auditory Neuroscience: Making Sense of Sound*. Cambridge: MIT Press. doi: 10.7551/mitpress/7942.001.0001

Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., and Demuth, K. (2019). The acquisition of Mandarin tonal processes by children with cochlear implants. *J. Speech Lang. Hear. Res.* 62, 1309–1325. doi: 10.1044/2018_JSLHR-S-18-0304

Tao, D., Deng, R., Jiang, Y., Galvin, J. J. III, Fu, Q.-J., et al. (2015). Melodic pitch perception and lexical tone perception in Mandarin-speaking cochlear implant users. *Ear Hear.* 36, 102. doi: 10.1097/AUD.0000000000000086

Traunmuller, H., and Eriksson, A. (1993). *The frequency range of the voice fundamental in the speech of male and female adults.* Stockholm: Institutionen lingvistik, Stockholms Univ.

Vandali, A. E., Dawson, P. W., and Arora, K. (2017). Results using the opal strategy in Mandarin speaking cochlear implant recipients. *Int. J. Audiol.* 56(Suppl 2), S74–S85. doi: 10.1080/14992027.2016.1190872

Wang, S., Liu, B., Dong, R., Zhou, Y., Li, J., Qi, B., et al. (2012). Music and lexical tone perception in chinese adult cochlear implant users. *Laryngoscope* 122, 1353–1360. doi: 10.1002/lary.23271

Wang, W., Zhou, N., and Xu, L. (2011). Musical pitch and lexical tone perception with cochlear implants. *Int. J. Audiol.* 50, 270–278. doi: 10.3109/14992027.2010.542490

Wei, C.-G., Cao, K., and Zeng, F.-G. (2004). Mandarin tone recognition in cochlear-implant subjects. *Hear. Res.* 197, 87–95. doi: 10.1016/j.heares.2004.06.002

Whalen, D. H., and Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49, 25–47. doi: 10.1159/000261901

Xu, L., Tsai, Y., and Pfingst, B. E. (2002). Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses. *J. Acoust. Soc. Am.* 112, 247–258. doi: 10.1121/1.1487843

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). "On the duration of Mandarin tones," in *Interspeech* (Stockholm), 1407–1411. doi: 10.21437/Interspeech.2017-29

Zhang, C. (2019). Brain plasticity under early auditory deprivation: evidence from congenital hearing-impaired people. *Adv. Psychol. Sci.* 27, 278. doi: 10.3724/SP.J.1042.2019.00278

Zhou, H., Kan, A., Yu, G., Guo, Z., Zheng, N., and Meng, Q. (2022). Pitch perception with the temporal limits encoder for cochlear implants. *IEEE Trans. Neural Syst. Rehabil. Eng.* 30, 2528–39. doi: 10.1109/TNSRE.2022.3203079

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 08 March 2023
DOI 10.3389/fdgth.2023.1058163

A summary of the ComParE COVID-19 challenges

Harry Coppock[1]*, Alican Akman[1], Christian Bergler[2], Maurice Gerczuk[3], Chloë Brown[4], Jagmohan Chauhan[5], Andreas Grammenos[4], Apinan Hasthanasombat[4], Dimitris Spathis[4], Tong Xia[4], Pietro Cicuta[4], Jing Han[4], Shahin Amiriparian[3], Alice Baird[3], Lukas Stappen[3], Sandra Ottl[3], Panagiotis Tzirakis[1], Anton Batliner[3], Cecilia Mascolo[4] and Björn W. Schuller[1,3]

[1]Department of Computing, Imperial College London, London, United Kingdom, [2]Department of Computing, FAU Erlangen-Nürnberg, Erlangen-Nürnberg, Germany, [3]Institute of Computer Science, Universität Augsburg, Augsburg, Germany, [4]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [5]Department of Computing, University of Southampton, Southampton, United Kingdom

EDITED BY
Lina F. Soualmia,
Université de Rouen, France

REVIEWED BY
Rüdiger Christoph Pryss,
Julius Maximilian University of Würzburg,
Germany,
Mikko Kurimo,
Aalto University, Finland

*CORRESPONDENCE
Harry Coppock
✉ harry.coppock@imperial.ac.uk

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 30 September 2022
ACCEPTED 09 January 2023
PUBLISHED 08 March 2023

CITATION
Coppock H, Akman A, Bergler C, Gerczuk M, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, Xia T, Cicuta P, Han J, Amiriparian S, Baird A, Stappen L, Ottl S, Tzirakis P, Batliner A, Mascolo C and Schuller BW (2023) A summary of the ComParE COVID-19 challenges.
Front. Digit. Health 5:1058163.
doi: 10.3389/fdgth.2023.1058163

The COVID-19 pandemic has caused massive humanitarian and economic damage. Teams of scientists from a broad range of disciplines have searched for methods to help governments and communities combat the disease. One avenue from the machine learning field which has been explored is the prospect of a digital mass test which can detect COVID-19 from infected individuals' respiratory sounds. We present a summary of the results from the INTERSPEECH 2021 Computational Paralinguistics Challenges: COVID-19 Cough, (CCS) and COVID-19 Speech, (CSS).

KEYWORDS
COVID-19, machine learning, Digital Health, computer audition, deep learning

Introduction

Significant work has been conducted exploring the possibility that COVID-19 yields unique audio biomarkers in infected individuals' respiratory signals (1–14). This has shown promising results although many still remain sceptical, suggesting that models could simply be relying on spurious bias signals in the datasets (15, 12). These worries have been supported by findings that when sources of bias are controlled, the performance of the classifiers decreases (16, 17). Along with this, cross dataset experiments have reported a marked drop in performance when models trained on one dataset are then evaluated on another dataset, suggesting dataset specific bias (18).

Last summer, the machine learning community were called upon to address some of these challenges, and help answer the question whether a digital mass test was possible, through the creation of two COVID-19 challenges within the Interspeech Computational Paralinguistics challengE (ComParE) series: COVID-19 Cough, (CCS) and COVID-19 Speech, (CSS) (19). Contestants were tasked to create the best performing COVID-19 classifier from user cough and speech recordings. We note that another COVID-19 detection from audio challenge was run at a similar time to ComParE, named DiCOVA (20), and point the inquisitive reader to their blog post[1] which details a summary of the results.

[1] https://dicova2021.github.io

Challenge methodology

Both COVID-19 cough and speech challenges were binary classification tasks. Given an audio signal of a user coughing or speaking, challenge participants were tasked with predicting whether the respiratory signal came from a COVID-19 positive or negative user. After signing up to the challenge, teams were sent the audio files along with the corresponding labels for both the training and development set. Teams were also sent the audio files from the test set without the corresponding labels. Teams were allowed to submit five predictions for the test set from which the best score was taken. The number of submissions was limited to avoid overfitting to the test set.

The datasets used in these challenges are two curated subsets of the crowd sourced Cambridge COVID-19 Sounds database (1, 21). COVID-19 status was self-reported and determined through either a PCR or rapid antigen test, the exact proportions of which are unknown. The number of samples of both positive and negative cases for these selected subsets are detailed in **Table 1**. The submission date for both COVID-19 positive and negative case recordings are detailed in **Figure 1A**. **Figure 1B** shows the age distribution for both CSS and CCS challenges.

TABLE 1 ComParE COVID-19 sub-challenges dataset splits. Values specify the number of audio recordings. We note that disjoint participant train, development, and test splits were ensured.

	CCS[a]			CSS[b]		
	Train	Dev	Test	Train	Dev	Test
COVID-19-positive	71	48	39	72	142	94
COVID-19-negative	215	183	169	243	153	189
Total	286	231	208	315	295	283

[a]CCS – COVID-19 Cough Sub-Challenge.
[b]CSS – COVID-19 Speech Sub-Challenge.

Overview of methodologies used in accepted papers at interspeech 2021

Last year, 44 teams registered in both the ComParE COVID-19 Cough Sub-Challenge (CCS) and the COVID-19 Speech Sub-Challenge (CSS) of which 19 submitted test set predictions. Five of the 19 teams submitted papers to INTERSPEECH which were then accepted. Results for both CCS and CSS were reported in two of these papers, while two papers reported results exclusively for CCS and one paper exclusively for CSS. In this section, we provide a brief overview of methodologies used in these accepted works which included data augmentation techniques, feature types, classifier types, and ensemble model strategies. Teams that did not have their work accepted at INTERSPEECH 2021 will be named NN_X to preserve anonymity. NN refers to *nomen nescio* and X is the order in which they appear in **Figure 2**. The performance measured in Unweighted Average Recall (UAR) achieved by these methodologies is summarised in **Table 2**; UAR has been used as a standard measure in the Computational Paralinguistics Challenges at Interspeech since 2009 (26). It is the mean of the diagonal of the confusion matrices in percent and by that, fair towards sparse classes. Note that UAR is sometimes called “macro-average,’ see (27).

Data augmentation

To combat the limited size and imbalance of the Cambridge COVID-19 Sounds database, the majority of the teams used data augmentation techniques in their implementation. Team Casanova et al. exploited a noise addition method and SpecAugment to augment the challenge dataset (23). Team Illium et al. targeted spectrogram-level augmentations with temporal shifting, noise addition, SpecAugment and loudness adjustment (25). Instead of using a data augmentation method to manipulate the challenge dataset, team Klumpp et al. used three auxiliary datasets in

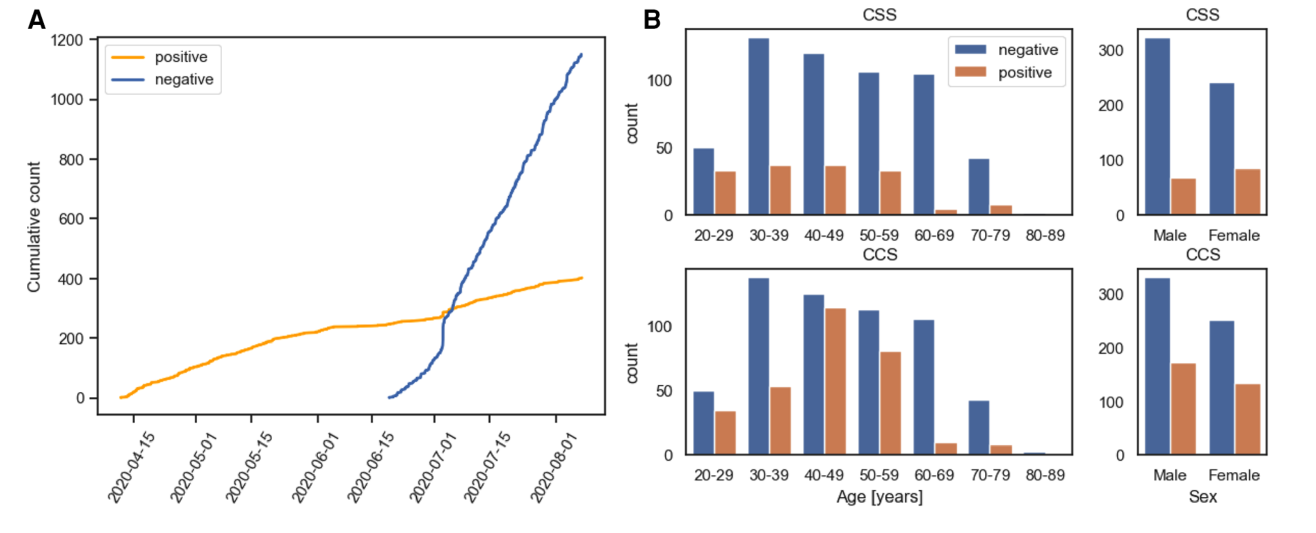


FIGURE 1
(A) Is a cumulative plot detailing when COVID-19 positive and negative submission to both the CCS and CSS were made. (B) Details the age and sex distribution of COVID-19 positive and negative participants for the CCS and CSS Sub-Challenges.

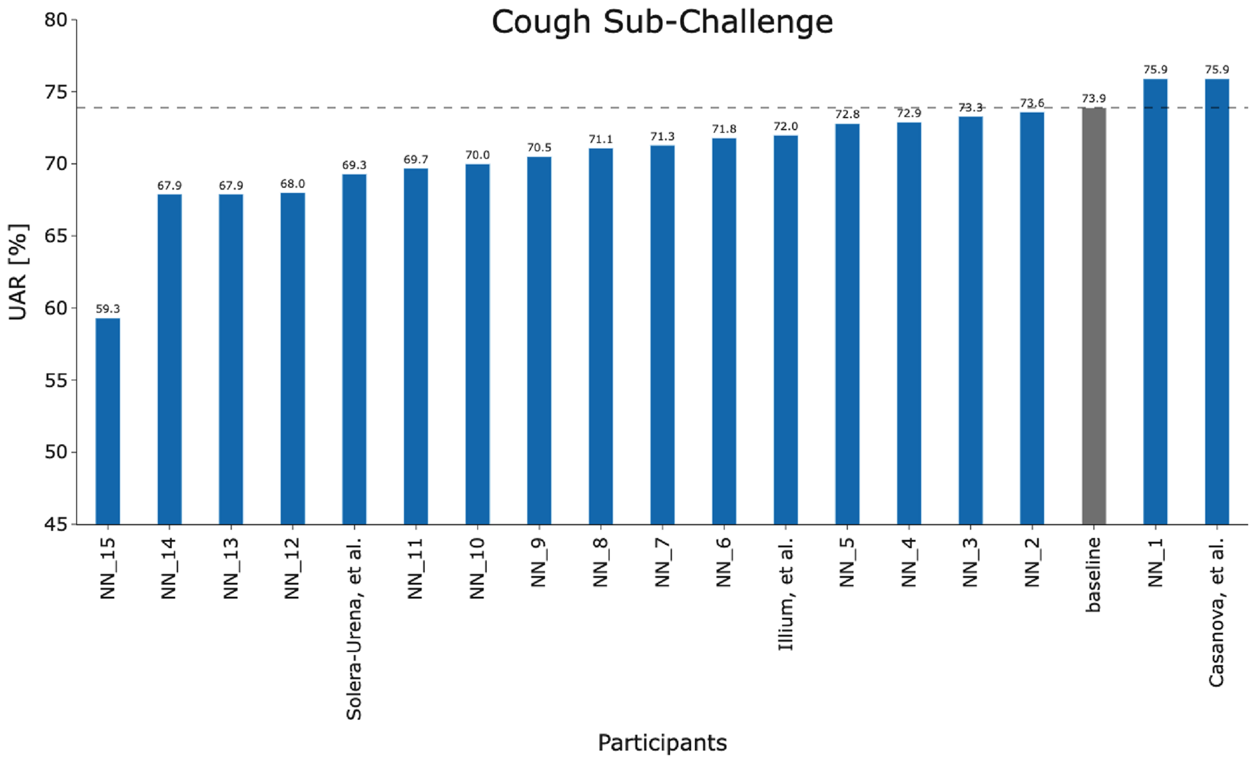


FIGURE 2
Team performance on the held out test set for the COVID-19 Cough Sub-Challenge.

TABLE 2 Summary of methodologies used in accepted papers at Interspeech 2021 along with their classification performance. Unweighted Average Recall (UAR) and Unweighted Average F1 (UF1) metrics are provided [%].

Team name	Data Aug.	Feature type	Classifiers	Ensemble	Cough		Speech	
					UAR	UF1	UAR	UF1
Solera-Urena et al. (22)	✗	TDNN-F, VGGish, PASE+	SVM	✓	69.3	65.2	–	–
Casanova et al. (23)	✓	MFCC, mel-spectrogram	SpiraNet, CNN14, ResNet-38, MobileNet	✓	**75.9**	69.6	70.3	71.0
Klumpp et al. (24)	✓	mel-spectrogram	CNN, LSTM, SVM, LR	✗	–	–	64.2	64.3
Illium et al. (25)	✓	mel-spectrogram	Vision transformer	✗	72.0	71.1	–	–
Baseline (19)	✗	openSMILE, openXBOX, DiFE, DeepSpectrum, auDeep	SVM, End2You	✓	73.9	–	**72.1**	–

different languages aiming their deep acoustic model to better learn the properties of healthy speech (24).

Feature type

The teams chiefly used spectrogram-level features including mel-frequency cepstral coefficients (MFCC) and mel-spectrograms. For higher-level features, the teams used the common feature extraction toolkits openSMILE (28), openXBOX (29), DeepSpectrum (30), and auDeep (31), where a simple support vector machine (SVM) model was built on top of these features. Team Solera-Urena et al. exploited transfer learning to extract feature embeddings by using pre-trained TDNN-F (32), VGGish (33), and PASE+ (34) models with appropriate fine-tuning on the challenge dataset. Team Klumpp et al. targeted to extract their own phonetic features by using an acoustic model consisting of convolutional neural network (CNN) and long short-term memory (LSTM) parts.

Classifier type

Team Solera-Urena et al. (22) and the challenge baseline (19) fitted a SVM model to high level audio embeddings extracted using TDNN-F (32), VGGish (33), and PASE+ (34) models, and the openSMILE framework (28), respectively. While the challenge baseline (19) searched for the complexity parameter of the SVM ranging from 10^{-5} to 1, team Solera-Urena et al. (22) explored different kernels (linear, RBF), data normalisations (zero mean and unit variance, [0,1] range) and class balancing methods (majority class downsampling, class weighting). In addition to the SVM model, the baseline explored using the multimodel profiling toolkit End2You (35) to train a recurrent neural network using Gated Recurrent Units (GRUs) with hidden units of 64. Team Casanova et al. (23) utilised the deep models: SpiraNet (36), CNN14 (37), ResNet-38 (37), and MobileNetv1 (37) where they explored kernel size, convolutional dilatation, dropout, number of fully connected layer neurons, learning rate, weight decay and optimizer. Team Klumpp et al. (24) trained SVM and logistic regression (LR) models to perform COVID-19 classification on top of phonetic features extracted by their deep acoustic model. They explored the complexity parameter of the SVM ranging from 10^{-4} to 1. Team Illium et al. (25) adapted a vision transformer (38) for mel-spectrogram representations of audio signals. Tree-structured Parzen Estimator-algorithm (TPE) (39) was exploited in (25) for hyperparameter search mainly exploring embedding size, learning rate, batch size, dropout, number of heads and head dimension. The teams Solera-Urena et al., Casanova et al., and the baseline also reported classification results by using the fusion of their best features and classifiers. To conclude, Casanova et al. performed best among the accepted papers with a consistent performance over both CCS and CSS. This showed the importance of using proper data augmentation techniques and exhaustive exploration of deep models and hyperparameters for a transfer learning approach.

Assessment of performance measures

Figure 4 visualises a two-sided significance test (based on a *Z*-test concerning two proportions, (40), section 5B) employing the CCS and CSS test sets and the corresponding baseline systems (19). Various levels of significance (α-values) were used for calculating an absolute deviation with respect to the test set, being considered as significantly better or worse than the baseline systems. Due to the fact that a two-sided test is employed, the α-values must be halved to derive the respective *Z*-score used to calculate the *p*-value of a model fulfilling statistical significance for both sides (40). Consequently, significantly outperforming the best CCS baseline system (73.9% and 208 test set samples) at a significance level of $\alpha = 0.01$ requires at least an absolute improvement of 6.7%; for CSS (best baseline system with 72.1% and 283 test set samples), the improvement required is 6.0%. Note that Null-Hypothesis-Testing with *p*-values as criterion has been criticised from its beginning; see the statement of the American Statistical Association in Wasserstein and Lazar (41) and Batliner et al. (42). Therefore, we provide this plot with *p*-values as a service for readers interested in this approach, not as a guideline for deciding between approaches.

Another way of assessing performance measures as for their "uncertainty" is computing confidence intervals (CIs). Schuller et al. (19) employed two different CIs: first, 1000× bootstrapping for test (random selection with replacement) and UARs based on the same model that was trained with Train and Dev; in the following, the CIs for these UARs are given first. Then, 100× bootstrapping for the corresponding combination of Train and Dev; the different models obtained from these combinations were employed to get UARs for test and subsequently, CIs; these results are given in second place. Note that for this type of CI, the test results are often above the CI, sometimes within and in a few cases below, as can be seen in (19); obviously, reducing the variability of the samples in the training phase with bootstrapping results on average in somehow lower performance. For CCS with a UAR of 73.9%, the first CI was 66.0%–82.6%; the second one could not be computed because this UAR is based on a fusion of different classifiers. For CSS with a UAR of 72.1%, the CIs were 66.0%–77.8% and 70.2%–71.1%, respectively. Both **Figure 4** and the spread of the CIs reported demonstrate the uncertainty of the results, caused by the relatively low number of data points in the test set.

Results and discussion

Figures 2 and **3** detail the rankings for the 19 teams which submitted predictions for the test set. We congratulate (23) for winning the COVID-19 Cough Sub-Challenge with an UAR of 75.9% on the held out test set.[2] We note that for the COVID-19 Speech Sub-Challenge, no team exceeded the performance of the baseline which scored 72.1% UAR on the held out test set. To significantly outperform the baseline system for the cough modality, with a significance level of $\alpha = 0.1$, as detailed in **Figure 4**, would require an improvement of 6.7%, an improvement which the winning submission fell short of by 4.7%.

For both Sub-Challenges, teams struggled to outperform the baseline. Postulating why this could be the case one could suggest one, or a combination, of the following: COVID-19 detection from audio is a particularly hard task, the baseline score—being already a fusion of several state-of-the-art systems for CCS—represents a performance ceiling and that higher classification scores are not possible for this dataset, or, as a result of the limited size of the dataset, the task lends itself to less data hungry algorithms, such as the openSMILE-SVM baseline models for CSS.

It is important to analyse the level of agreement of COVID-19 detection between participant submissions. This is shown schematically in **Figures 5** and **6**. From these figures, we can see that there are clearly COVID-19 positive cases which teams across the board are able to correctly predict, but there are also positive COVID-19 cases which all teams have missed. These findings are reflected in the minimal performance increase of 0.3% and 0.8% for cough and speech tasks, respectively, obtained when fusing *n* best submission predictions through majority voting schemes. The

[2] UAR is the established ComParE evaluation metric. UAR is equivalent to balanced accuracy. We note that if F1 had been the evaluation metric, Illium, et al. (25) would have infact won the cough sub challenge. This is thanks to their model's superior precision performance, i.e., what proportion of the model's positive predictions are correct.

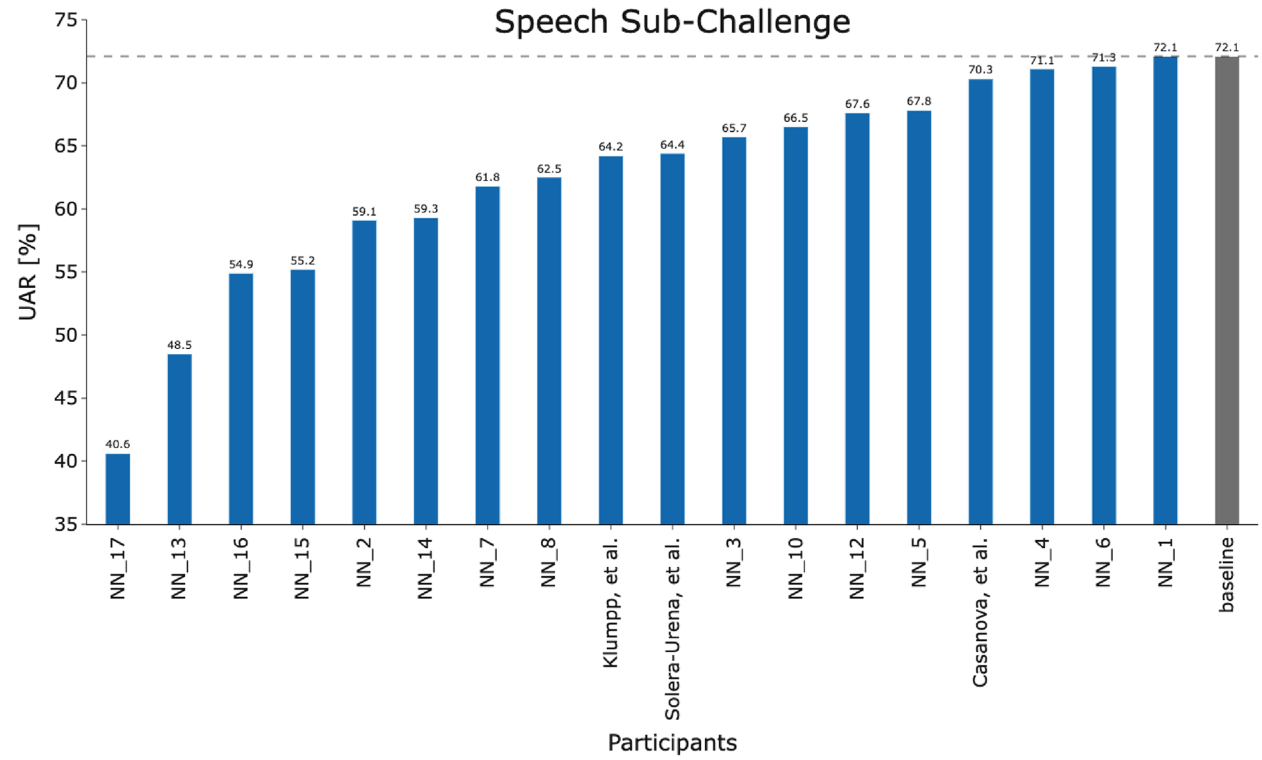


FIGURE 3
Team performance on the held out test set for the COVID-19 Speech Sub-Challenge.

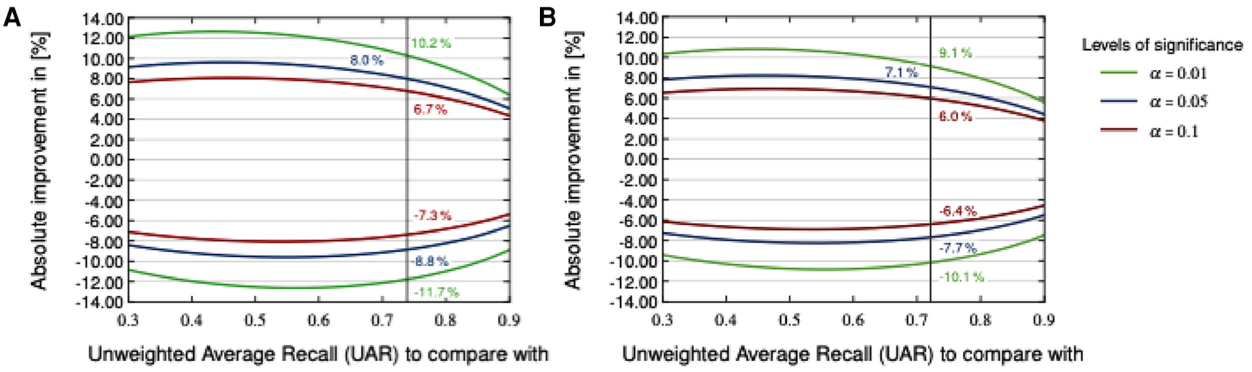


FIGURE 4
Two-sided significance test on the COVID-19 Cough (A) and Speech (B) test sets with various levels of significance according to a two-sided *Z*-test.

results from fusing *n* best models using majority voting are detailed in **Figures A2** and **A3** . This suggests that models from all teams are depending on similar audio features when predicting COVID-19 positive cases.

Figures 5 and **6B,C** detail the level of agreement across submissions for curated subset of the test set, where participants were selected if they were displaying at least one symptom (b) and when they were displaying no symptoms (c). These figures can be paired with **Figure A1** which details the recall scores for positive cases across these same curated test sets. From this analysis, it does not appear that there was a trend across teams to perform favourably on cases where symptoms were being displayed or vice versa. While this does not disprove worries that these algorithms are simply cough or symptom identifiers, it does not add evidence in support of this claim.

Limitations

While this challenge was an important step in exploring the possibilities of a digital mass test for COVID-19, it has a number of limitations. A clear limiting factor of the challenge

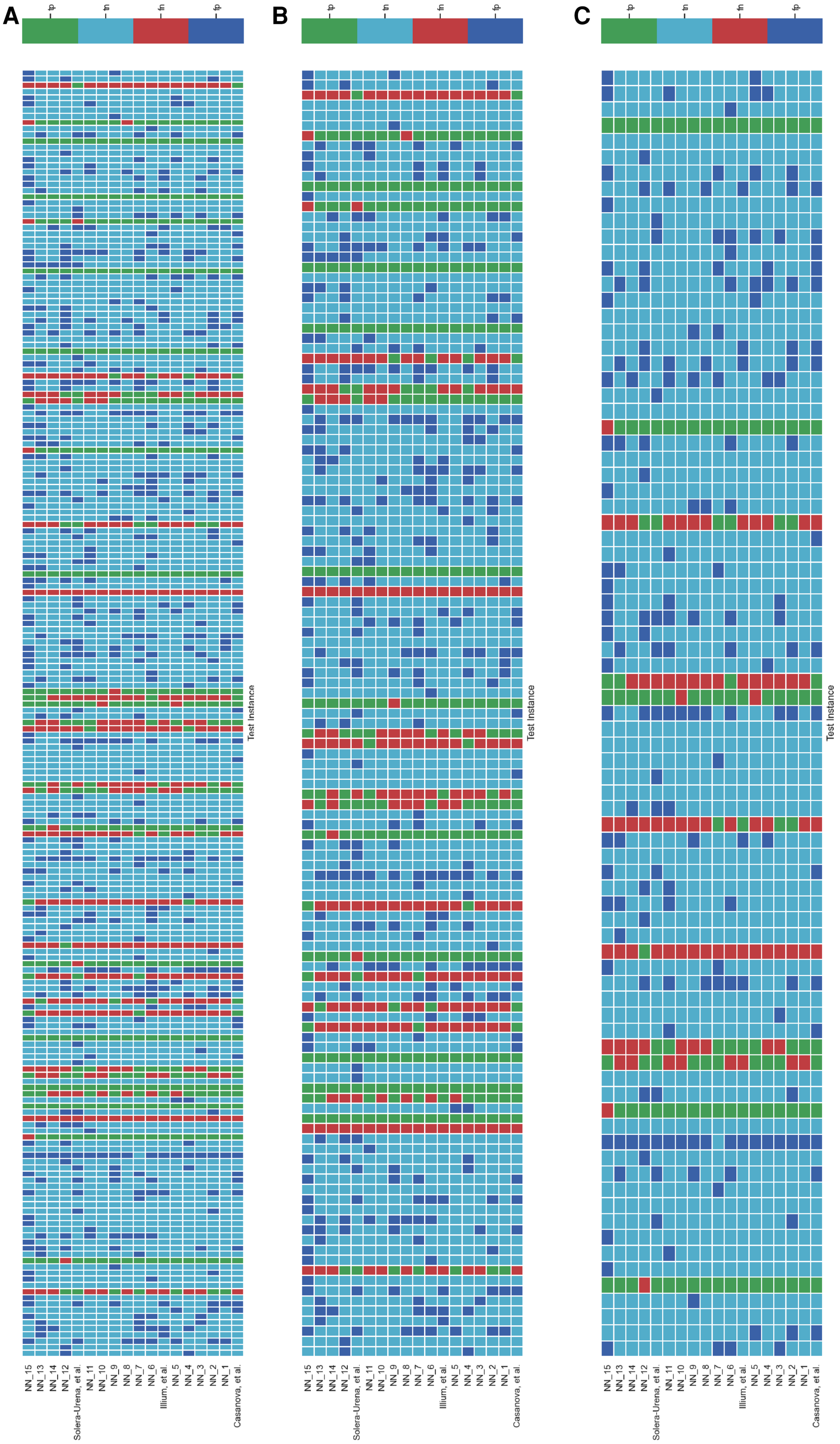


FIGURE 5
Schematic detailing the level of agreement between teams for each test instance for the **COVID-19 Cough Sub-Challenge**. Each row represents a team's submission results. The teams have been ordered by Unweighted Average Recall, from the bottom up (team Casanova et al.'s predictions represent the highest scoring submission). Each column represents all teams predictions, across the competition, for one test instance. The test instances appear in the order in which they are in the test set. (A) Details all the test instances, (B) details only the test instances which were experiencing symptoms at the time of recording, and (C) details only the test instances which were experiencing no symptoms at the time of recording.

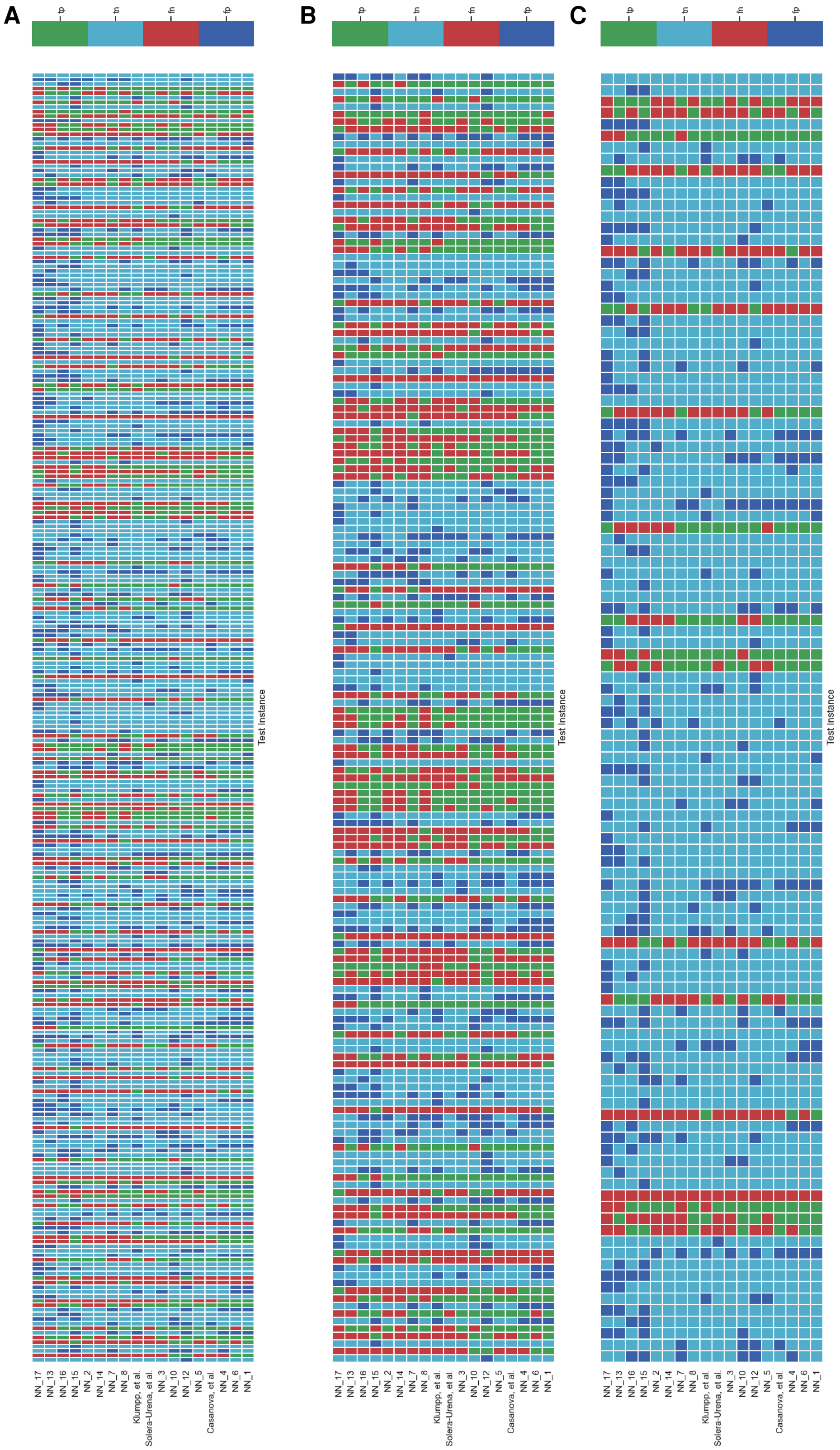


FIGURE 6
Schematic detailing the level of agreement between teams for each test instance for the **COVID-19 Speech Sub-Challenge**. Each row represents a team's submission results. The teams have been ordered by Unweighted Average Recall (UAR), from the bottom up (team *yoshiharuyamamoto*'s predictions represent the highest scoring submission). Each column represents all teams' predictions, across the competition, for one test instance. The test instances appear in the order which they are in the test set. note: There are more test cases in the COVID-19 Speech Sub-Challenge than in the COVID-19 Cough Sub-Challenge. (**A**) Details all the test instances, (**B**) details only the test instances which were experiencing symptoms at the time of recording, and (**C**) details only the test instances which were experiencing no symptoms at the time of recording.

was the small size of the dataset. While many participants addressed this through data augmentation and regularisation techniques, it restricted the extent to which conclusions could be taken from the results, particularly investigating teams' performance on carefully controlled subsets of the data. We look forward to the newly released COVID-19 sounds dataset (21) which represents a vastly greater source of COVID-19 samples.

A further limitation of this challenge is the unforeseen correlation between low sample rate recordings, below 12 kHz, and COVID-19 status. In fact all low sample rate recordings in the challenge for both CCS and CSS were COVID-19 positive. For CCS and CSS there were 30 and 37 low sample rate cases, respectively. The reason for this is that at the start of the study the label in the survey for COVID-19 negative was unclear, and could have been interpreted as either "not tested" or "tested negative." For this reason no negative samples from the time period were used. This can be seen in **Figure 1A**. This early version of data collection also correlated with the study allowing for lower sample rate recordings, a feature which later was changed to restrict submissions to higher sample rates. This resulted in all the low sample rate recordings being COVID-19 positive. As can be seen in **Figures A4**, **A5**, **A6** and **A7**, teams' trained models were able to pick up on the sample rate bias, with most teams correctly predicting all the low sample rate cases as COVID-19 positive. When this is controlled for and low sample rate recordings are removed from the test set, as shown in **Figures A6** and **A7**, teams' performances drop significantly. For the challenge baselines this too was the case, with the fusion of baseline models for CCS falling from 73.8% to 68.6% UAR and the opensmile-SVM baseline for CSS dropping from 72.1% to 70.9% UAR. This is a great example of the effect of overlooked bias which expresses itself as an identifiable audio feature, leading to inflated classification scores. We regret that this was not found earlier. Inspecting **Figure 1A**) further, one will also realise that all the COVID-19 negative individuals were collected in the summer of 2020, one could argue that this ascertainment bias injected further imbalance between COVID-19 negative and positive individuals. An example of this is that individuals are much less likely to have the flu in summer (43), resulting in respiratory symptoms having an inflated correlation with COVID-19 status in the collected dataset compared to the general population. This has been shown to artificially boost model performance at COVID-19 detection (44–46). In future more factors, which can be a source of bias, should be controlled for, namely in this case, age of participant, gender, symptoms, location of recording and date of recording. Matching on these attributes would yield more realistic performance metrics.

As with most machine learning methods, it still remains unclear how to interpret the decision making process at inference time. This results in it being tricky to determine which acoustic features the model is correlating with COVID-19. Whether that be true, acoustic features caused by the COVID-19 infection or other acoustic bias (15, 44). We also note that this is a binary classification task, in that models only had to decide between COVID-19 positive or negative. This "closed word fallacy" (42) leads to inflated performance as models are not tasked with discerning between confounding symptoms such as heavy cold or asthma. Tasking models to predict COVID-19 out of a wide range of possible conditions/symptoms would be a harder task. The test set provided saw a complete temporal overlap with the training set, in future it would be nice to experiment with time disjoint test sets, as in (44) to investigate whether the signal for COVID-19 changes over time. Collecting a dataset which yields a test set with a higher proportion of COVID positive individuals is also desirable.

In this challenge, participants were provided with the test set recordings (without the corresponding labels). In future challenges, test set instances should be kept private, requiring participants to submit trained models along with pipeline scripts for inference. Teams' test set predictions can then be run automatically by the challenge organisers. This will help in reducing the possibility of overfitting and foul play. We note that there was no evidence of foul play, e.g., training in an unsupervised manner on the test set, in this challenge.

Another limitation of this challenge was the lack of meta data that organisers could provide to participants. This tied teams' hands to some extent in evaluating for themselves the level of bias in the dataset and so their opportunity to implement methods to combat it. This was not a desired feature. However, we now point teams towards the newly open sourced COVID-19 Sounds database (21) which also provides collected meta data. It is this dataset from which a subset of samples was taken for this challenge.

Conclusion

This challenge demonstrated that there is a signal in crowdsourced COVID-19 respiratory sounds that allows for machine learning algorithms to fit a classifier which achieves moderate detection rates of COVID-19 in infected individuals' respiratory sounds. Exactly what this signal is, however, still remains unclear. Whether these signals are truly audio biomarkers in respiratory sounds of infected individuals uniquely caused by COVID-19 or rather identifiable bias in the datasets, such as confounding flu like symptoms, is still an open question to be answered next.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

Ethical review and approval was not required for this study in accordance with the local legislation and institutional requirements.

Author contributions

HC and AA summarising the challenge work, analysing the challenge results, manuscript preparation, editing, and drafting manuscript. CB, MG, CBr, JC, AG, AH, DS, TX, PC, JH, SA, AB, LS, SO, PT, ABat, CM and BS organizing the challenge work. BS drafting manuscript and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

Funding

We acknowledge funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS) and the ERC project No. 833296 (EAR).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; New York, NY, USA. Association for Computing Machinery (2020). p. 3474–84. Available from: https://doi.org/10.1145/3394486.3412865.

2. Xia T, Han J, Qendro L, Dang T, Mascolo C. Uncertainty-aware COVID-19 detection from imbalanced sound data. *arXiv*. (2021) [preprint]. doi: 10.48550/ARXIV.2104.02005

3. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform Med Unlocked* (2020) 20:100378. doi: 10.1016/j.imu.2020.100378

4. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, Ghosh NRPK, et al. Coswara—a database of breathing, cough,, voice sounds for COVID-19 diagnosis. In *Proceedings INTERSPEECH 2020*; Shanghai, China. ISCA (2020). p. 4811–15. Available from: https://doi.org/10.21437/Interspeech.2020-2768

5. Bagad P, Dalmia A, Doshi J, Nagrani A, Bhamare P, Mahale A, et al. Cough against COVID: evidence of COVID-19 signature in cough sounds. *arXiv* (2020) [preprint]. doi: 10.48550/arXiv.2009.08790

6. Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 detection from voice. *Open J Eng Med Biol* (2020) 1:268–74. doi: 10.1109/OJEMB.2020.3026468

7. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci Data* (2021) 8:156. doi: 10.1038/s41597-021-00937-4

8. Andreu-Perez J, Perez-Espinosa H, Timonet E, Kiani M, Giron-Perez MI, Benitez-Trinidad AB, et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test, severity levels. *IEEE Trans Serv Comput* (2021). doi: 10.1109/TSC.2021.3061402

9. Pizzo DT, Santiago Esteban S, de los Ángeles Scetta M. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples. *arXiv*. (2021) [preprint]. doi: 10.48550/arXiv.2104.13247

10. Qian K, Schmitt M, Zheng H, Koike T, Han J, Liu J, et al. Computer audition for fighting the SARS-CoV-2 corona crisis – introducing the multi-task speech corpus for COVID-19. *Internet Things J* (2021) 8(21):16035–46. doi: 10.1109/JIOT.2021.3067605

11. Bartl-Pokorny KD, Pokorny FB, Batliner A, Amiriparian S, Semertzidou A, Eyben F, et al. The voice of COVID-19: acoustic correlates of infection in sustained vowels. *J Acoust Soc Am* (2021) 149:4377–83. doi: 10.1121/10.0005194

12. Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller BW. End-to-end convolutional neural network enables COVID-19 detection from breath, cough audio: a pilot study. *BMJ Innov* (2021) 7:356–62. doi: 10.1136/bmjinnov-2021-000668

13. Nessiem MA, Mohamed MM, Coppock H, Gaskell A, Schuller BW. Detecting COVID-19 from breathing, coughing sounds using deep neural networks. In *International Symposium on Computer-Based Medical Systems (CBMS)*; Aveiro, Portugal. IEEE (2021). p. 183–8. Available from: https://doi.org/10.1109/CBMS52027.2021.00069.

14. Ponomarchuk A, Burenko I, Malkin E, Nazarov I, Kokh V, Avetisian M, et al. Project achoo: a practical model and application for COVID-19 detection from recordings of breath, voice, and cough. *IEEE J Sel Top Signal Process* (2022) 16(2):175–87. doi: 10.1109/JSTSP.2022.3142514

15. Coppock H, Jones L, Kiskin I, Schuller BW. COVID-19 detection from audio: seven grains of salt. *Lancet Digit Health* (2021) 3(9):e537–8. doi: 10.1016/S2589-7500(21)00141-2

16. Han J, Xia T, Spathis D, Bondareva E, Brown C, Chauhan J, et al. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *NPJ Digit Med*. (2022) 28:5(1). doi: 10.1038/s41746-021-00553-x

17. Coppock H, Jones L, Kiskin I, Schuller BW. Bias and privacy in AI's cough-based COVID-19 recognition – Authors' reply. *Lancet Digit Health* (2021) 3:e761. doi: 10.1016/S2589-7500(21)00233-8

18. Akman A, Coppock H, Gaskell A, Tzirakis P, Jones L, Schuller BW. Evaluating the COVID-19 identification ResNet (CideR) on the INTERSPEECH COVID-19 from audio challenges. *Front Digit Health*. (2022) 4:789980. doi: 10.3389/fdgth.2022.789980

19. Schuller BW, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, et al. The interspeech 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. In *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*; Brno, Czechia (2021).

20. Muguli A, Pinto L, Sharma NRN, Krishnan P, Ghosh PK, Kumar R, et al. DiCOVA challenge: dataset, task, and baseline system for COVID-19 diagnosis using acoustics. *arXiv*. (2021) doi: 10.48550/arXiv.2103.09148

21. Xia T, Spathis D, Brown C, Ch AGJ, Han J, Hasthanasombat A, et al. COVID-19 sounds: a large-scale audio dataset for digital COVID-19 detection. *NeurIPS 2021 Track Datasets and Benchmarks Round2 Submission*. OpenReview (2021). Available from: https://openreview.net/forum?id=9KArJb4r5ZQ.

22. Solera-Ureña R, Botelho C, Teixeira F, Rolland T, Abad A, Trancoso I. Transfer learning-based cough representations for automatic detection of COVID-19. In *Proceedings of Interspeech 2021* (2021). p. 436–40. Available from: https://doi.org/10.21437/Interspeech.2021-1702.

23. Casanova E, Candido Jr A, Fernandes Jr RC, Finger M, Gris LRS, Ponti MA, et al. Transfer learning and data augmentation techniques to the COVID-19 identification tasks in ComParE 2021. In *Proceedings of Interspeech 2021* (2021). p. 446–50. Available from: https://doi.org/10.21437/Interspeech.2021-1798.

24. Klumpp P, Bocklet T, Arias-Vergara T, Vásquez-Correa J, Pérez-Toro P, Bayerl S, et al. The phonetic footprint of COVID-19? In *Proceedings of Interspeech 2021* (2021). p. 441–5. Available from: https://doi.org/10.21437/Interspeech.2021-1488.

25. Illium S, Müller R, Sedlmeier A, Popien C-L. Visual transformers for primates classification and covid detection. In *Proceedings of Interspeech 2021* (2021). p. 451–455. Available from: https://doi.ord/10.21437/Interspeech.2021-273.

26. Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge. In *Proceedings of INTERSPEECH*; Brighton. (2009). p. 312–5.

27. Manning CD, Raghavan P, Schütze H, *An introduction to information retrieval*. Cambridge: Cambridge University Press (2009).

28. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13; New York, NY, USA. Association for Computing Machinery (2013). p. 835–8. Available from: https://doi.org/10.1145/2502081.2502224

29. Schmitt M, Schuller B. openXBOW – introducing the passau open-source crossmodal bag-of-words toolkit. *J Mach Learn Res*. (2017) 18:3370–4.

30. Amiriparian S, Gerczuk M, Ottl S, Cummins N, Pugachevskiy S, Schuller B. Bag-of-deep-features: noise-robust deep feature representations for audio analysis. In *2018 International Joint Conference on Neural Networks (IJCNN)* (2018). p. 1–7. Available from: https://doi.org/10.1109/IJCNN.2018.8489416.

31. Freitag M, Amiriparian S, Pugachevskiy S, Cummins N, Schuller B. auDeep: unsupervised learning of representations from audio with deep recurrent neural networks. *J Mach Learn Res.* (2018) 18:6340–44.

32. Villalba J, Chen N, Snyder D, Garcia-Romero D, McCree A, Sell G, et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Comput Speech Lang* (2020) 60:101026. doi: 10.1016/j.csl.2019.101026

33. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. Cnn architectures for large-scale audio classification. New Orleans, LA, United States: 017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017). p. 131-5. doi: 10.1109/ICASSP.2017.7952132

34. Ravanelli M, Zhong J, Pascual S, Swietojanski P, Monteiro J, Trmal J, et al. Multi-task self-supervised learning for robust speech recognition. Barcelona, Spain: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). p. 6989–93. doi: 10.1109/ICASSP40776.2020.9053569

35. Tzirakis P, Zafeiriou S, Schuller BW. End2you – the imperial toolkit for multimodal profiling by end-to-end learning. *arXiv* (2018) [preprint]. doi: 10.48550/arXiv.1802.01115

36. Casanova E, Gris L, Camargo A, da Silva D, Gazzola M, Sabino E, et al. Deep learning against COVID-19: respiratory insufficiency detection in Brazilian Portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics (2021). p. 625–33. Available from: https://doi.org/10.18653/v1/2021.findings-acl.55.

37. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28 (2020). p. 2880–94. doi: 10.1109/TASLP.2020.3030497

38. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv*. (2021) [preprint]. doi: 10.48550/arXiv.2010.11929

39. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. CoRR. (2019) [preprint]. doi: 10.48550/arXiv.1907.10902

40. Isaac E. Test of hypothesis - concise formula summary (2015).

41. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat* (2016) 70:129–33. doi: 10.1080/00031305.2016.1154108

42. Batliner A, Hantke S, Schuller B. Ethics and good practice in computational paralinguistics. *IEEE Trans Affect Comput*. (2020). 13(3):1236–53. doi: 10.1109/TAFFC.2020.3021015

43. Lowen AC, Steel J. Roles of humidity and temperature in shaping influenza seasonality. *J Virol* (2014) 88(14):7692–5. doi: 10.1128/JVI.03544-13

44. Coppock H, Nicholson G, Kiskin I, Koutra V, Baker K, Budd J, et al. Audio-based ai classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. *arXiv*. (2022) [preprint]. doi: 10.48550/arXiv.2212.08570

45. Pigoli D, Baker K, Budd J, Butler L, Coppock H, Egglestone S, et al. Statistical design and analysis for robust machine learning: a case study from COVID-19. *arXiv*. (2022) [preprint]. doi: 10.48550/arXiv.2212.08571

46. Budd J, Baker K, Karoune E, Coppock H, Patel S, Cañadas AT, et al. A large-scale and PCR-referenced vocal audio dataset for COVID-19. *arXiv*. (2022) [preprint]. doi: 10.48550/arXiv.2212.07738

Appendix

Here we present some results from ablation studies of teams' performances through evaluating performance on curated subsets of the test set. Figure A1 details the effect of controlling for symptom cofounders on teams' performance. Figures A6 and A7 repeats this analysis however controlling for sample rate. Figures A4 and A5 details the level of agreement between teams for the low Figures A4A, A5A and high A4B, A5B sample rate test cases. Figures A2 and A3 detail the classification performance of a fusion of teams' predictions on the test set.

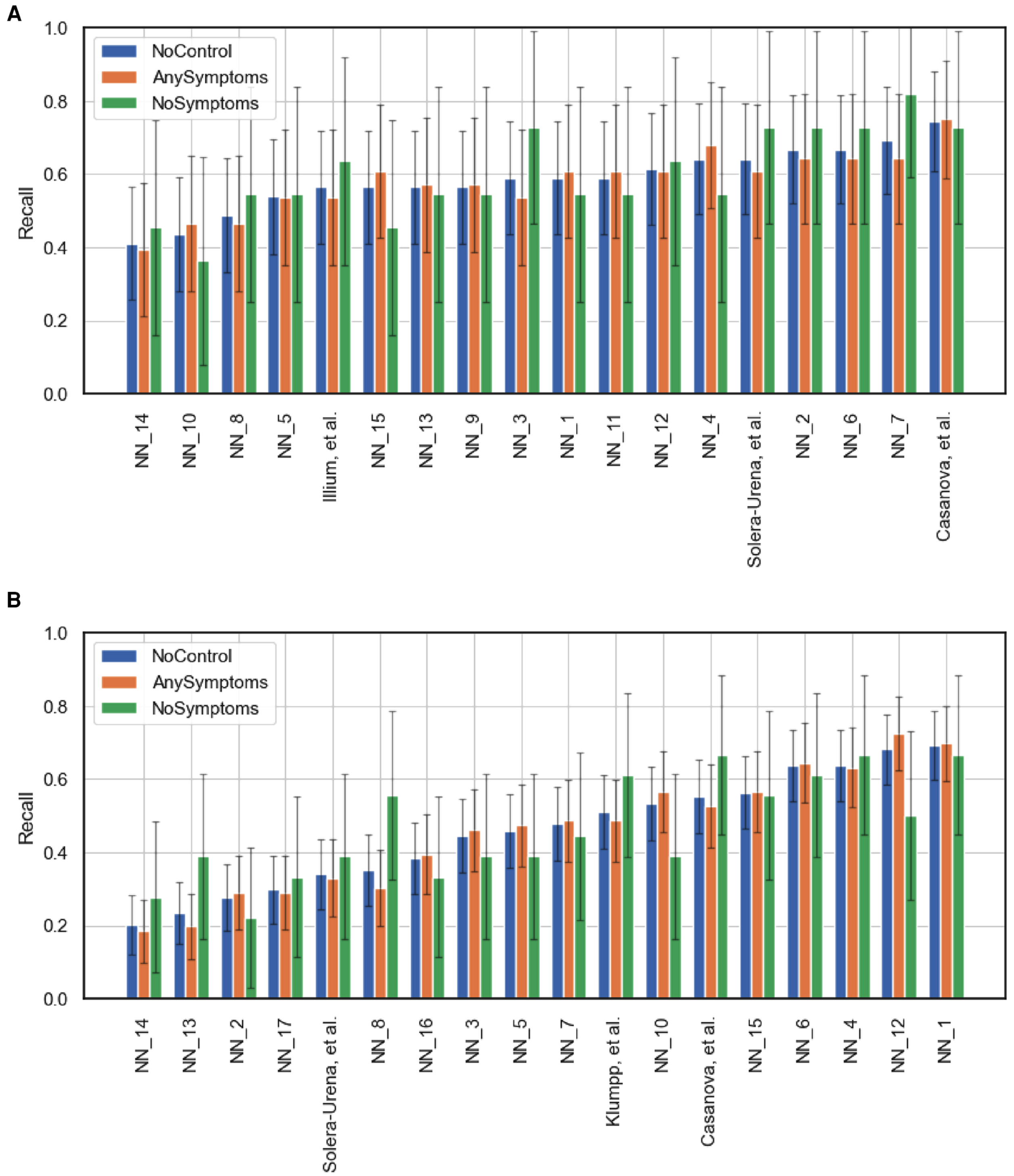


FIGURE A1
Team performance on the full test set (NoControl) and two curated test sets featuring only test instances where the participants either had at least one symptom (AnySymptoms) or were displaying no symptoms at all (NoSymptoms). The metric reported is recall for positive cases. 95% confidence intervals are shown, calculated via the normal approximation method. (A) Corresponds to the COVID-19 Cough Sub-Challenge, CCS, and (B) the COVID-19 Speech Sub-Challenge, CSS.

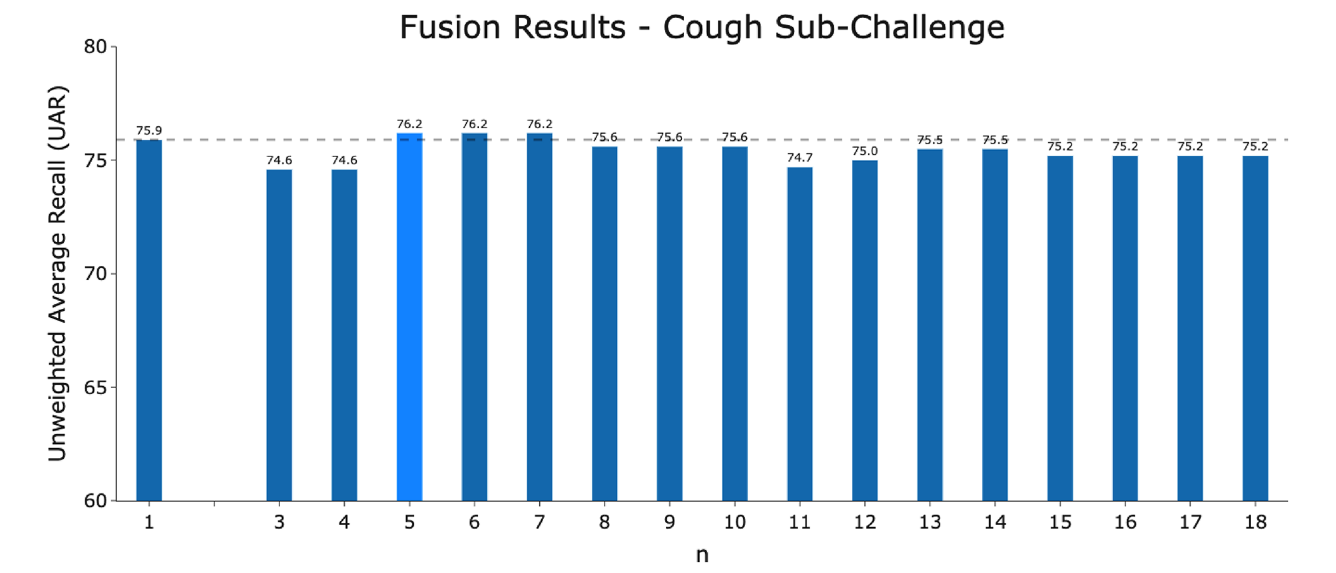


FIGURE A2
The performance of the fusion model of *n*-best models for the COVID-19 Cough Sub-Challenge using majority voting.

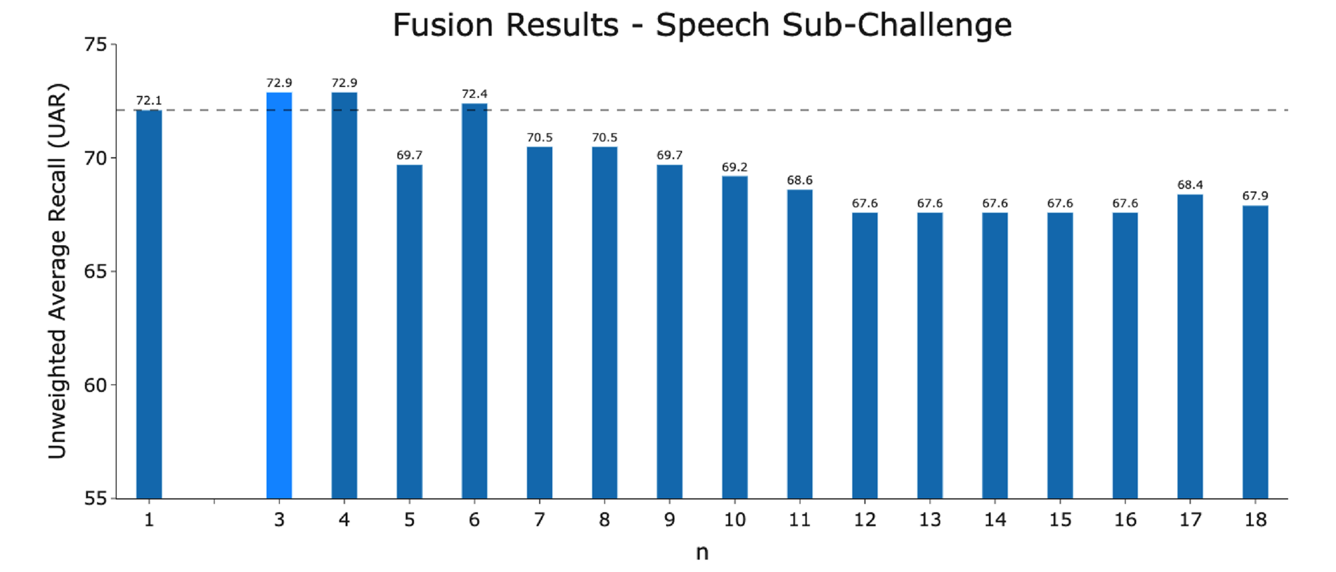


FIGURE A3
The performance of the fusion model of *n*-best models for the COVID-19 Speech Sub-Challenge using majority voting.

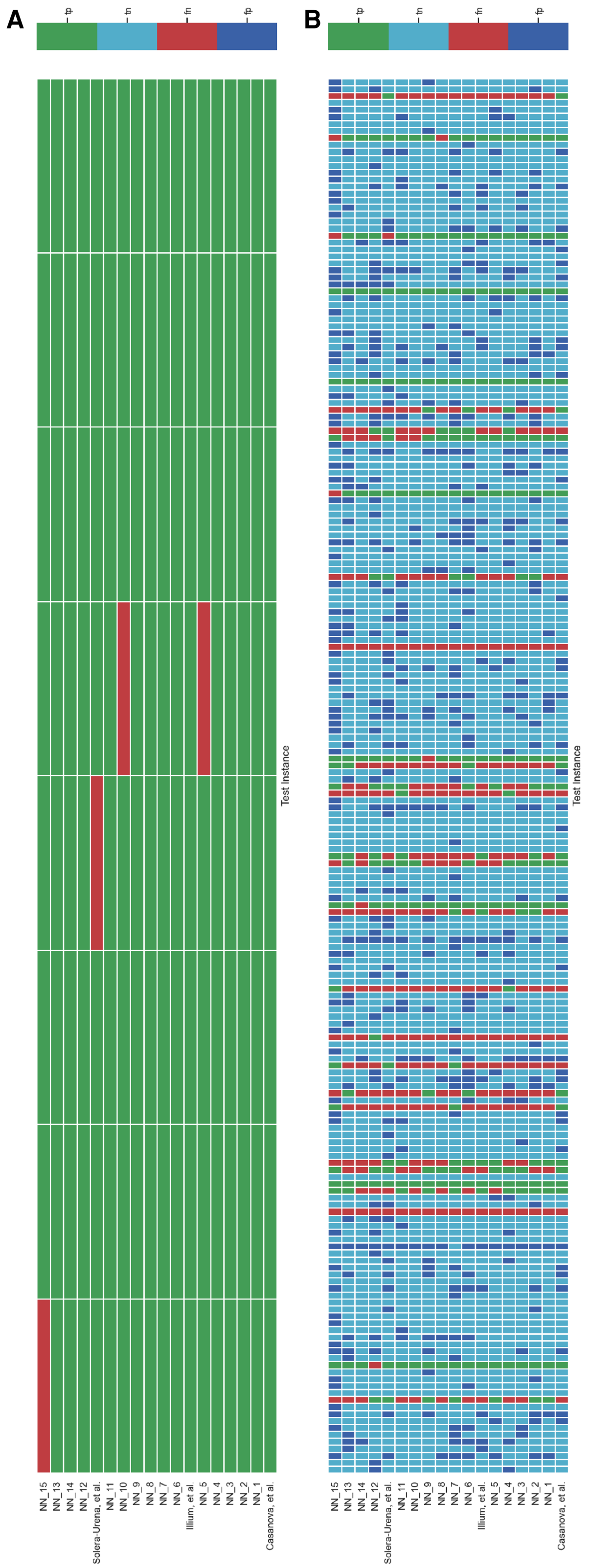


FIGURE A4
Schematic detailing the level of agreement as in Figure 5 with test instances with either a low sample rate (below 12 kHz) (A) or high sample rate (above 12 kHz) (B).

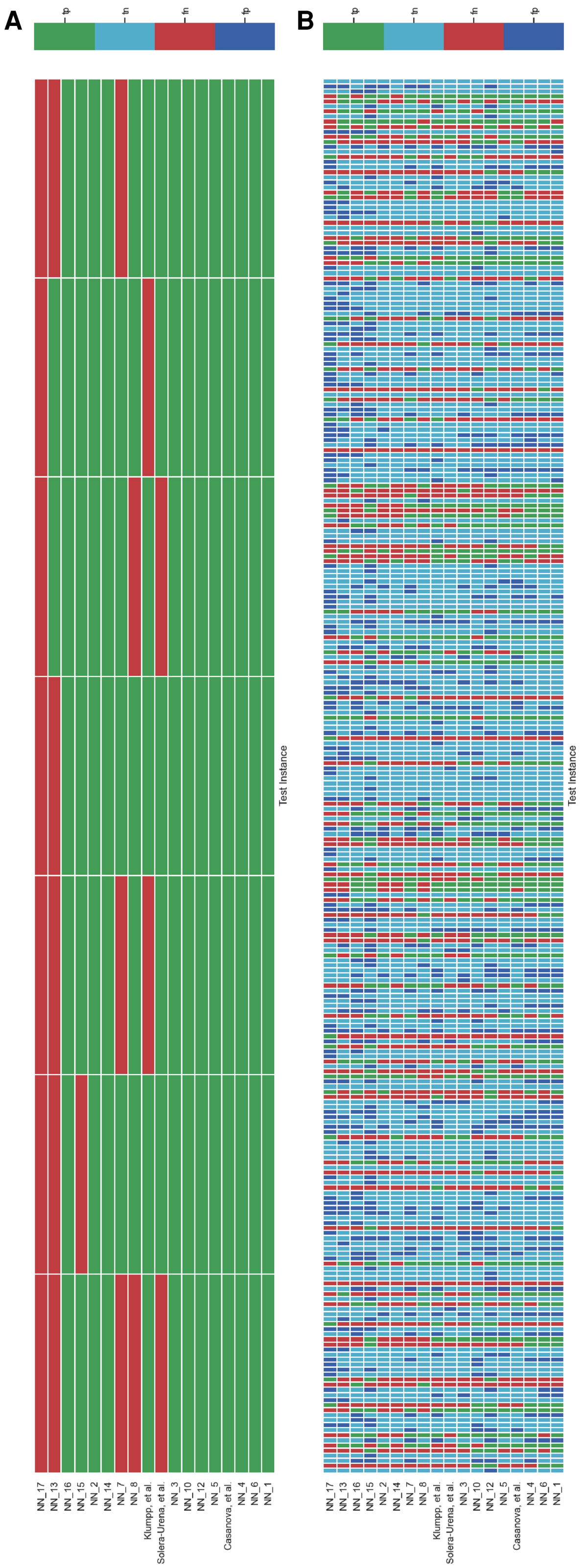


FIGURE A5
Schematic detailing the level of agreement as in Figure 6 with test instances with either a low sample rate (below 12 kHz) (A) or high sample rate (above 12 kHz) (B).

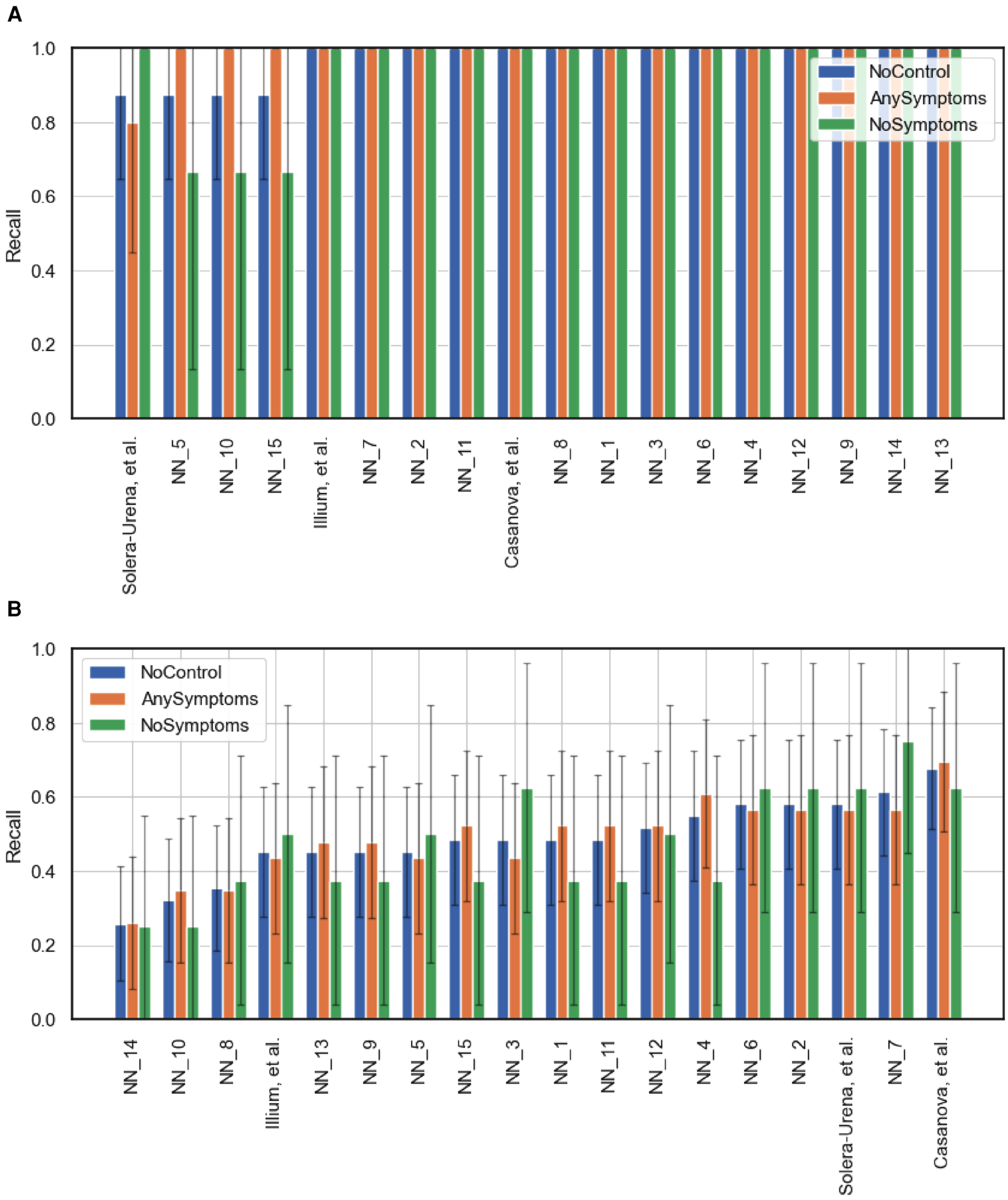


FIGURE A6
Team performance on two curated test sets from the COVID-19 Cough Sub-Challenge. (A) Controls for test samples with a sample rate of greater than 12 kHz and (B) controls for test samples with a sample rate of 12 kHz and below. The metric reported is recall for positive cases. 95% confidence intervals are shown, calculated via the normal approximation method.

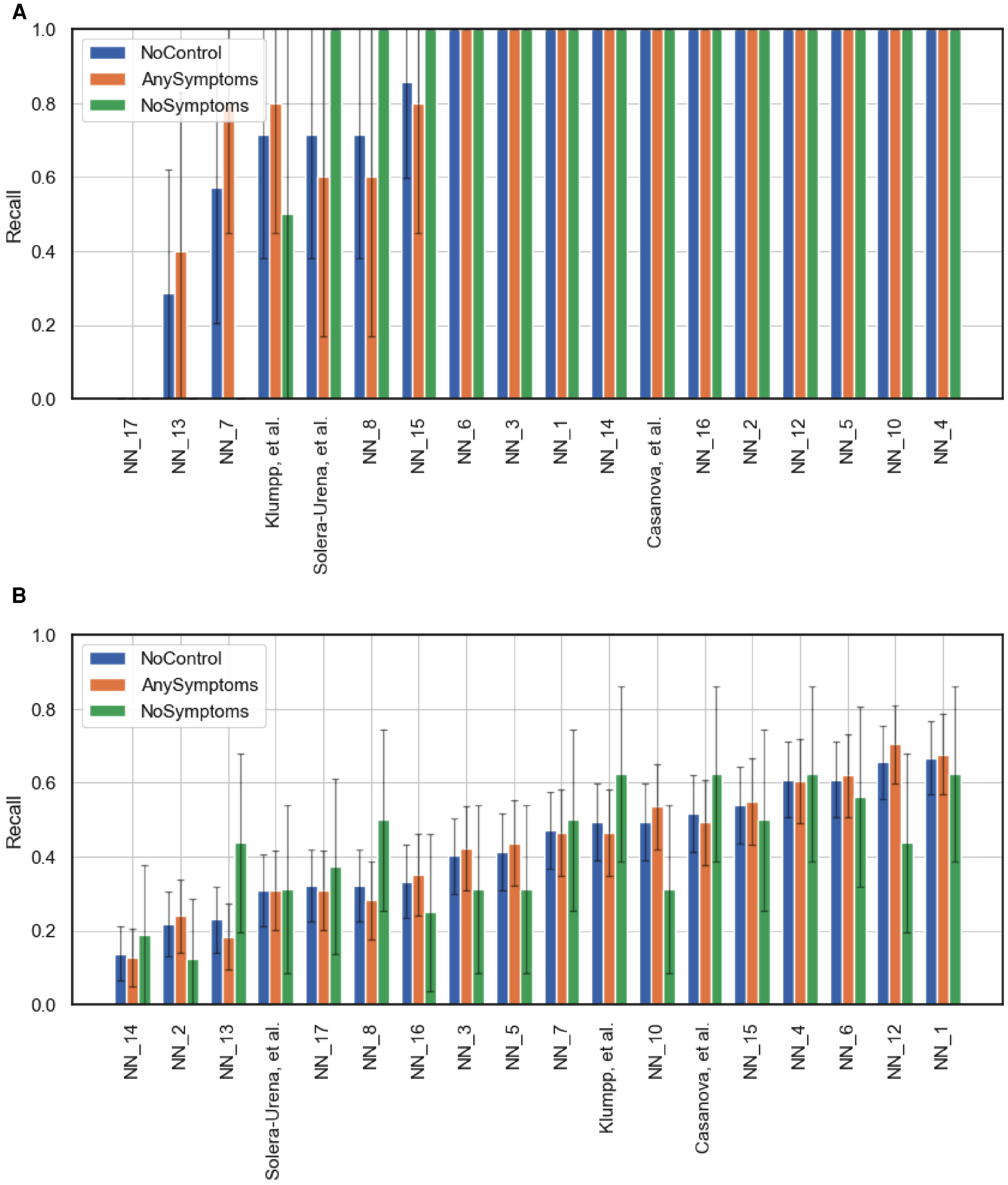


FIGURE A7
Team performance on two curated test sets from the COVID-19 Speech Sub-Challenge. (A) Controls for test samples with a sample rate of greater than 12 kHz and (B) controls for test samples with a sample rate of 12 kHz and below. The metric reported is recall for positive cases. 95% confidence intervals are shown, calculated via the normal approximation method.

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/fcomp.2023.1039261

EDITED BY
Kun Qian,
Beijing Institute of Technology, China

REVIEWED BY
Shahin Amiriparian,
University of Augsburg, Germany
Yu Tsao,
Academia Sinica, Taiwan

*CORRESPONDENCE
Shruti Kshirsagar
✉ shruti.kshirsagar@inrs.ca

SPECIALTY SECTION
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

RECEIVED 08 September 2022
ACCEPTED 13 March 2023
PUBLISHED 28 March 2023

CITATION
Kshirsagar S, Pendyala A and Falk TH (2023)
Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions.
Front. Comput. Sci. 5:1039261.
doi: 10.3389/fcomp.2023.1039261

Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions

Shruti Kshirsagar[1]*, Anurag Pendyala[2] and Tiago H. Falk[1]

[1]Institut National de la Recherche Scientifique, University of Quebec, Montréal, QC, Canada,
[2]International Institute of Information Technology, Bangalore, India

Automatic emotion recognition (AER) systems are burgeoning and systems based on either audio, video, text, or physiological signals have emerged. Multimodal systems, in turn, have shown to improve overall AER accuracy and to also provide some robustness against artifacts and missing data. Collecting multiple signal modalities, however, can be very intrusive, time consuming, and expensive. Recent advances in deep learning based speech-to-text and natural language processing systems, however, have enabled the development of reliable multimodal systems based on speech *and* text while only requiring the collection of audio data. Audio data, however, is extremely sensitive to environmental disturbances, such as additive noise, thus faces some challenges when deployed "in the wild." To overcome this issue, speech enhancement algorithms have been deployed *at the input signal level* to improve testing accuracy in noisy conditions. Speech enhancement algorithms can come in different flavors and can be optimized for different tasks (e.g., for human perception vs. machine performance). Data augmentation, in turn, has also been deployed *at the model level* during training time to improve accuracy in noisy testing conditions. In this paper, we explore the combination of task-specific speech enhancement and data augmentation as a strategy to improve overall multimodal emotion recognition in noisy conditions. We show that AER accuracy under noisy conditions can be improved to levels close to those seen in clean conditions. When compared against a system without speech enhancement or data augmentation, an increase in AER accuracy of 40% was seen in a cross-corpus test, thus showing promising results for "in the wild" AER.

KEYWORDS

multimodal emotion recognition, BERT based text features, modulation spectrum features, data augmentation, speech enhancement, context-awareness

1. Introduction

Affective human-machine interfaces are burgeoning as they provide more natural interactions between the human and the machine (Zeng, 2007). Automated emotion recognition (AER) systems have seen applications across numerous domains, from marketing, smart cities and vehicles, to call centers and patient monitoring, to name a few. In fact, the COVID-19 pandemic has resulted in a global mental health crisis that will have long-term consequences to society, economy, and healthcare systems (Xiong et al., 2020). Being able to detect changes in affective states in a timely and reliable manner can allow individuals and organizations to put in place interventions to prevent, for example, burnout and depression (Patrick and Lavery, 2007).

AER systems can rely on a wide range of modalities, including speech, text, gestures/posture, and physiological responses (e.g., *via* changes in heart/breathing rates). For so-called "in the wild" applications, multimodal systems are preferred in order to compensate for certain confounds and to improve overall AER accuracy by providing the system with some redundancy and complementary information not available with unimodal systems (Naumann et al., 2009; Parent et al., 2019). Multimodal systems, however, can be very time consuming to implement, costly to run, and potentially intrusive to the users (e.g., requiring on-body sensors with physiological data collection) and their privacy (Sebe et al., 2005). Notwithstanding, with audio inputs, one may be able to devise a multimodal speech-and-text system with the use of an advanced speech-to-text system, thus relying on a single input modality. As such, text and speech have emerged as two popular AER modalities.

Recent advances in deep learning architectures, such as transformers (Vaswani et al., 2017), have redefined the performance envelope of existing AER systems. In fact, most state-of-the-art systems today rely on deep neural network architectures in some way. For example, for text-based systems, self attention and dynamic max pooling has been proposed by Yang et al. (2019). The widely-used Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018), in turn, has been used to detect cyber abuse in English and Hindi texts (Malte and Ratadiya, 2019). The work by Lee and Tashev (2015) and Kratzwald et al. (2018), in turn, relies on recurrent neural networks (RNN) to better consider long-range contextual effects and to better model the uncertainty around emotional labels. For speech-based AER systems, in turn, mel-spectral features combined with a convolutional neural networks (CNNs) have been extensively explored, specially with self-attention mechanisms to extract emotionally-informative time segments (e.g., Chen et al., 2021). Long-short term memory networks (LSTM) have also been extremely popular (e.g., Haytham et al., 2017; Tripathi et al., 2018; Zhao et al., 2019) and end-to-end solutions have also been explored (Tzirakis et al., 2017).

As mentioned previously, one major advantage of the audio modality is that recent advances in automated speech-to-text conversion have allowed for multimodal speech-and-text-based systems to emerge while requiring the collection of just one signal modality (Chuang and Wu, 2004). Text and speech have been shown to be very useful modalities for multimodal AER systems (Patamia et al., 2021). In this regard, attention-based bidirectional LSTM models (Li et al., 2020), bi-directional RNNs (Poria et al., 2017), transformer-based models (Siriwardhana et al., 2020), multi-level multi-head fusion attention mechanisms (Ho et al., 2020), graph-based CNNs (Zhang et al., 2019), gated-recurrent units (Poria et al., 2019), early and late fusion strategies (Jin et al., 2015), and cross-modal attention (Sangwan et al., 2019) have been explored as strategies to optimally combine information from the two modalities.

One major disadvantage of speech-based systems (either uni- or multi-modal), however, is their sensitivity to environmental factors, such as additive and convolutional noise (e.g., room reverberation). These factors can be detrimental to AER systems (Patamia et al., 2021; Maithri et al., 2022). Commonly, speech enhancement algorithms are applied at the input level stage to minimize environmental factors for in-the-wild speech applications. Enhancement methods can range from more classical methods, such as spectral subtraction and Wiener filtering (Cauchi et al., 2015; Braun et al., 2016), to more recent deep neural network (DNN) based ones (e.g., Parveen and Green, 2004; Lu et al., 2013; Pascual et al., 2017; Zhao et al., 2018). The use of speech enhancement for AER in-the-wild has shown some benefits (e.g., Avila et al., 2021).

Speech enhancement methods can have two very different purposes. If aimed at improving intelligibility/ quality, for example, human perception becomes the main driving factor and quality/intelligibility improvements are typically used as a figure of merit (e.g., Fu et al., 2021). However, if enhancement is used to improve downstream speech recognition applications then other machine-driven outcome measures, such as word error rate improvements, are more appropriate. As such, depending on the final task, the enhancement procedure can be very different. The work by Bagchi et al. (2018), for example, showed that mimic loss-based enhancement was optimal for automatic speech recognition (ASR) downstream tasks. Having this said, it is hypothesized that for multimodal speech-and-text AER systems the use of two different enhancement procedures will be useful, with a quality-driven one used for the speech branch (mimicking how humans perceive emotions from speech) and a machine-driven one for the speech-to-text branch. We will test this hypothesis herein.

Lastly, with deep learning based approaches showing the latest state-of-the-art results, data augmentation has emerged as a useful technique to make systems more robust to in-the-wild distortions at the model training stage (e.g., Hannun et al., 2014). With data augmentation, the training set is increased multi-fold by applying certain transformations to the available training signals, including time-reversal, time-frequency masking, pitch alterations, background noise addition and reverberation corruption, to name a few. For AER specifically, the work by Etienne et al. (2018) showed that vocal track length perturbations served as a useful data augmentation strategy. In this paper, we further explore the advantages that data augmentation can provide, in addition to speech enhancement, for multimodal in-the-wild AER.

The remainder of this paper is organized as follows. Section 2 describes the proposed system. Section 3 describes the experimental setup. Experimental results and a discussion are presented in Section 4 and conclusions in Section 5.

2. Proposed method

Figure 1 depicts the block diagram of the proposed multimodal AER pipeline. In the case of interest here, speech $S(i)$ is assumed to be corrupted by additive background noise $N(i)$, resulting in noisy speech signal $Y(i) = S(i) + N(i)$. With the multimodal AER system, the top branch focuses on extracting emotion-relevant features directly from the speech component, whereas the bottom branch relies on a state-of-the-art automatic speech recognizer (ASR) to generate text from the noisy speech signal. Features are then extracted from the text transcripts. We concatenated Speech and text features, then these concatenated features are input to a deep neural network for final emotion classification. As noisy speech is known to corrupt AER/ASR performance, here we also include a speech enhancement step, one optimized for speech quality improvement (top branch) and another for ASR. Each sub-block is described in detail in the subsections to follow:

2.1. Speech enhancement

Enhancement and noise suppression has been widely used across many different speech-based applications. In human-to-human communications, the goal of enhancement is to improve the quality of the noisy signal, not only to increase intelligibility, but also to improve paralinguistic characterization that humans do so well, such as emotion recognition. In human-to-machine interaction (e.g., ASR), however, improving quality may not be the ultimate goal, and instead, improvement in downstream system accuracy could be regarded as a better optimization criterion. Here, we explore the use of a quality-optimized enhancement algorithm for the speech branch of the proposed method and an ASR-optimized algorithm for the text generation branch. The two algorithms used are described next:

2.1.1. MetricGAN+: A quality-optimized enhancement method

MetricGAN+ is a recent state-of-the-art deep neural network specifically optimized for quality enhancement of noisy speech and shown to outperform several other enhancement benchmarks (Fu et al., 2019, 2021). In particular, two networks are used. The discriminator's role is to minimize the difference between the predicted quality scores (given by the so-called PESQ, perceptual evaluation of speech quality, rating Rix et al., 2001) and actual PESQ quality scores. PESQ is a standardized International Telecommunications Union full-reference speech quality metric that maps a pair of speech files (a reference and the noisy counterpart) into a final quality rating between 1 (poor) and 5 (excellent). PESQ has been widely used and validated across numerous speech applications.

The generator's role, in turn, is to map a noisy speech signal into its enhanced counterpart. The discriminator and generator models are trained together to enhance the noisy signal in a manner that maximizes the PESQ score of the enhanced signal. MetricGAN+ builds on the original MetricGAN (Fu et al., 2019) *via* two improvements for the discriminator and one for the generator. More specifically, for the discriminator training, along with the enhanced and clean speech signals, the noisy speech was also used to minimize the distance between the discriminator and target objective metrics. The second improvement is that the speech generated from the previous epochs is reused to train the discriminator to avoid the catastrophic forgetting of the discriminator. For the generator, in turn, the learnable sigmoid function was used for mask estimation. The interested reader is referred to Fu et al. (2019, 2021) for more details on the MetricGAN and MetricGAN+ speech enhancement methods.

2.1.2. Mimic loss: An ASR-optimized enhancement method

Spectral mapping-based speech enhancement is an enhancement method specifically optimized for downstream ASR applications (Bagchi et al., 2018). We refer henceforth to this method as 'mimic loss based enhancement' as the model uses mimic loss instead of student-teacher learning, thus the speech enhancer is not jointly trained with a particular acoustic model. We use this enhancement model as it has been shown to be a useful pre-processing method for many ASR systems, thus offers some flexibility on the choice of ASR model to use (Bagchi et al., 2018). The overall system is comprised of two major components: a spectral mapper and a spectral classifier which are trained in three steps.

First, a spectral classifier is trained to predict senone labels from clean speech with a cross-entropy criterion, resulting in a classification loss L_C between predicted and actual senones. The weights of this spectral classifier are then frozen and used in the last step. Second, a spectral mapper is pre-trained to map noisy speech features to clean speech features using a mean squared error (MSE) criterion. This results in a fidelity loss L_F between the denoised features and features from the clean speech counterpart. Bagchi et al. (2018) relied on log-spectral magnitude components extracted over 25ms windows with a 10-ms shift as features and a deep feed-forward neural network for mapping.

Lastly, noisy speech is input to the pre-trained spectral mapper, resulting in a denoised version, which is input to the "frozen" spectral classifier, resulting in a predicted senone. In parallel, the clean speech counterpart is also input to the frozen spectral classifier, resulting in a soft senone label and a mimic loss L_M between the soft senone label and the predicted senone. The spectral mapper is then retrained using joint loss (L_F and L_M), thus allowing the enhancer to emulate the behavior of the classifier under clean conditions while keeping the projection of noisy signal closer to that of the clean signal counterpart. The same hyperparameters described by Bagchi et al. (2018) were used herein. The interested reader is referred to Bagchi et al. (2018) for more details on the mimic loss enhancement method.

2.2. Automatic speech recognition

In order to generate text from speech, a state-of-the-art automatic speech recognizer is needed. Here, wav2vec 2.0, an end-to-end speech recognition system is used (Baevski et al., 2020). A complete description of the method is beyond the scope of this paper, hence only an overview is provided; the interested reader can obtain more details from Baevski et al. (2020). Wav2vec 2.0 relies on the raw speech waveform as input. This 1-dimensional data then passes through a multi-layer 1-d CNN to generate speech representation vectors. Vector quantization is then used on these latent representations to match them to a codebook. Half of the available speech data is masked and the remaining quantized data is fed into a transformer network. By using contrastive loss, the model attempts to predict the masked vectors, thus allowing for

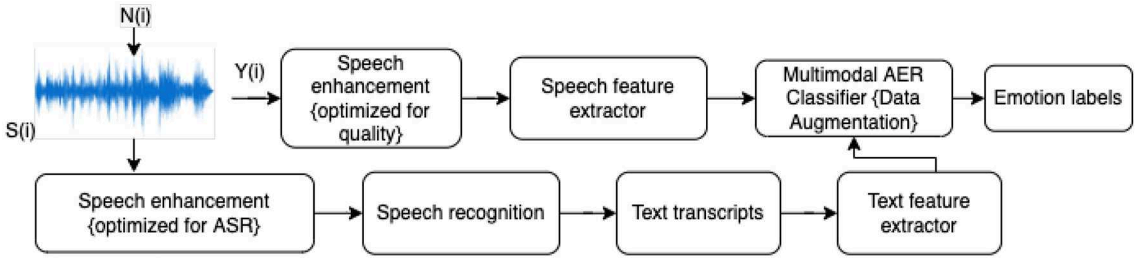

N(i)
S(i)
Y(i)
Speech enhancement {optimized for quality}
Speech feature extractor
Multimodal AER Classifier {Data Augmentation}
Emotion labels
Speech enhancement {optimized for ASR}
Speech recognition
Text transcripts
Text feature extractor

FIGURE 1
Experimental pipeline for AER using audio and text features.

pre-training on unlabeled speech data. The model is then fine-tuned on labeled data for the subsequent down-streaming ASR task.

2.3. Speech feature extractor

Several AER systems have been proposed recently, and they have relied on different speech feature representations. Here, we focus on the three most popular representations, namely: prosodic, eGeMAPS, and modulation spectral features. In particular, prosody features include fundamental frequency (F0), intensity measures, and voicing probabilities, as these have been widely linked to emotions (Banse and Scherer, 1996). Next, the so-called extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016), which has been widely used in many recent emotion recognition challenges (e.g., Valstar, 2016; Ringeval et al., 2019; Xue et al., 2019), is also explored and contains a set of 88 acoustic parameters relating to pitch, loudness, unvoiced segments, temporal dynamics, and cepstral features. Lastly, modulation spectral features are explored as they capture second-order periodicities in the speech signal and have been shown to convey emotional information (Wu et al., 2011; Avila et al., 2021). Modulation spectral features (termed MSFs) were extracted using a window size of 256 ms and a frame step of 40 ms. The interested reader is referred to Falk and Chan (2010b) and Avila et al. (2021) for complete details on the computation of this representation.

2.4. Text feature representations

Text has also been used to infer the emotional content of written material and several state-of-the-art methods and techniques exist. Here, we explore three recent methods, namely BERT (Bidirectional Encoder Representations from Transformers), TextCNN, and Bag-of-Words (BoW). A brief overview of each method is given below:

2.4.1. BERT-bidirectional encoder representations from transformers

BERT is based on a transformer network and attention mechanism (Devlin et al., 2018) that also learns contextual relations between words in the text (Tenney et al., 2019). BERT comes in two flavors: BERTBase and BERTLarge. The BERTBase model uses 12 layers of transformers block with a hidden dimension of 768 and 12 self-attention heads; overall, there are approximately 110 million trainable parameters. On the other hand, BERTLarge uses 24 layers of transformers block with a hidden size of 1024 and 16 self-attention heads, resulting in approximately 340 million trainable parameters. Here, we employ the BERTBase model for text feature extraction. The BERT hidden state vector is used as input to the AER system. The interested reader is referred to Devlin et al. (2018) for more details on BERT.

2.4.2. TextCNN

TextCNN is a deep learning model for short text classification tasks and has been used as a baseline model for text classification (Zhang et al., 2018). TextCNN transforms a word into a vector using word embeddings, which are then fed into a convolutional layer, followed by a max-pooling layer, and a fully connected output layer. In our experiment, TexCNN embeddings were extracted using the model described by Poria et al. (2018). We used three convolutional layers with 64 filter and kernel sizes of 3, 4, and 5 respectively in each layer, followed by max-pooling and finally 150 dense layers to extract the final text features. Specifically, with pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014), we first extracted the semantic vector space representation and then fed them to a 1-D-CNN to extract 100-dimensional text features vector.

2.4.3. Bag-of-Words

The bag-of-words (BOW) method is commonly employed in natural language processing (Alston, 1964). The approach is straightforward and flexible and can be used in many ways to extract features from documents. BOW represents the text by describing the occurrence of words within a document. It consists of two parts: a vocabulary of known words and a measure of the presence of these words. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not wherein the document. In this method, first, a word histogram is generated within a text document. Next, the frequencies of each word from a dictionary are computed, and finally the resultant vector is fused and used as the

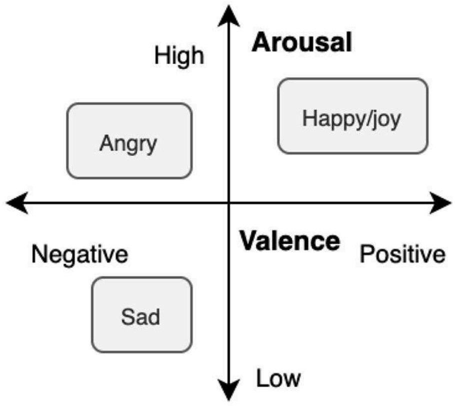


FIGURE 2
Valence-arousal emotional space with the three discrete emotions considered here.

text features. For our experiment, we used CountVectorizer from the sklearn library. A 652-dimensional feature vector was used for each utterance and the unigram model was used to generate the BOW representation.

2.5. Multimodal AER classifier

Here, we rely on a fully connected deep neural network for multimodal emotional recognition. Three dense layers (of dimensions 256, 128, 32) were used, plus a final classification layer. A dropout rate of 0.6 was used, batch normalization was performed after every layer, and class weights of [1, 1.8] were assigned during training. Grid search was performed on the validation set to obtain the optimal hyperparameters. Rmsprop, Adam, and SGD optimizers were explored, and learning rates of 0.01, 0.001, and 0.0001 were tested to find the optimal combination. Once the best parameters were found with the validation set, we reported the best performance on our test data. Experimentation codes are available on github[1]. The network is trained with and without data augmentation in order to explore its effect on in-the-wild AER performance.

3. Experimental setup

In this section, we present the setup used in our experiments.

3.1. Datasets used

The dataset used for experimentation is the Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018). It is a multimodal emotion classification dataset which has been created by extending the EmotionLines dataset (Chen et al., 2018). MELD contains approximately 13,000 utterances from 1,433 dialogues from the TV series 'Friends'. Each statement is annotated with emotion and sentiment labels and encompasses audio, visual, and textual modalities. The MELD dataset contains conversations, where each dialogue has utterances from multiple speakers. EmotionLines was created by crawling the discussions from each episode and then grouping them based on the number of statements in conversation into four groups of utterances. Finally, 250 dialogues were sampled randomly from each group, resulting in the final dataset of 1,000 dialogues. The utterances in each dialogue were annotated with the most appropriate emotion category.

For this purpose, the six universal emotions (joy, sadness, fear, anger, surprise, and disgust) were considered. This annotation list was extended with two additional emotion labels: neutral and non-neutral. Each utterance was annotated by five workers from the Amazon Mechanical Turk platform. A majority voting scheme was applied to select a final emotion label for each utterance. While the MELD dataset has labels for several emotions, here we focus on two specific binary tasks to gauge effects across the valence and arousal dimensions. More specifically, we first focus on two tasks. Task 1 comprises anger vs. sad classification to explore the benefits of the proposed tool for low/high arousal classification (Mower et al., 2010; Metallinou et al., 2012). Task 2, in turn, comprises joy vs. sad classification for positive-valence-high-arousal and negative-valence-low-arousal characterization (Park et al., 2013; Li et al., 2019). Figure 2 depicts the arousal-valence emotional space and the three discrete emotions considered. As such, the MELD dataset was split into three disjoint sets: training, test, and development. These were split as follows:

1. Training: angry (1,109 samples), joy (1,743 samples), and sad (682 samples);
2. Validation: angry (153 samples), joy (163 samples), and sad (111 samples);
3. Testing: angry (345 samples), joy (402 samples), and sad (208 samples).

To test the robustness of the proposed methods to in-the-wild conditions, the MELD dataset is corrupted by multi-talker babble noise, cafeteria noise, and noise recorded inside a commercial airplane at different SNR levels: −10,−20, 0, 5, 10, 15, and 20 dB. The AURORA (Hirsch and Pearce, 2000) and DEMAND noise datasets (Thiemann et al., 2013) are used for this purpose. Note that only a subset of these conditions are used during augmentation, including airport and babble noise and SNR levels of 0, 10, and 20 dB. The remainder are left as unseen conditions during testing.

Next, we utilized the IEMOCAP dataset to show the generalizability of the proposed model. The IEMOCAP dataset has 12 h of audio-visual data from 10 actors where the recordings follow the dialogue between a male and a female actor in both scripted or improvised topics. After the audio-video data was collected, it was divided into small utterances of length between 3 and 15 s, which were then labeled by evaluators. Each utterance was evaluated by 3–4 assessors. The evaluation form contained ten options (neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others). We consider only three: anger, sadness, and happy so as to remain consistent with the previous MELD data experiments and to be able to directly test the models

[1] https://github.com/shrutikshirsagar/Speech-enhancement-Audio-Text-ER

trained on the MELD dataset. To this end, the dataset was split into three disjoint sets: training (70%), development (10%) and test (20%). IEMOCAP contains utterances from the disjoined speakers in training and testing. More specifically, we used sessions 2,3,4, and 5 for training and session 1 for testing purposes. These were split as follows:

1. Training: angry (772 samples), happy (416 samples), and sad (758 samples);
2. Validation: angry (111 samples), happy (60 samples), and sad (110 samples);
3. Testing: angry (220 samples), happy (119 samples), and sad (216 samples).

Finally, we also utilized the spontaneous "in the wild" English-language Emoti-W database (Dhall et al., 2017). It was made available through the 2017 Emotion Recognition in the Wild Challenge. Some level of background noise was present in the recording as Emoti-w is "in the wild" dataset. The labels for the EMoti-W challenge dataset were created from the closed captions available in movies and TV series. Complete details about the Emoti-W dataset can be found in Dhall et al. (2017). The data is available in a sampling frequency of 48 kHz; videos are available in MPEG-2 format with 25 frames per second. Emotion labels are available for seven emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise were available in this dataset. We consider only three: anger, sadness, and happy/joy so as to remain consistent with the previous MELD data experiments and to be able to directly test the models trained on the MELD dataset as mentioned earlier. Again, we use only the labeled training and development subsets in our experiments. Training, testing and validation split of Emoti-W dataset are as follows:

1. Training: angry (110 samples), happy (120 samples), and sad (90 samples);
2. Validation: angry (11 samples), happy (24 samples), and sad (17 samples);
3. Testing: angry (64 samples), happy (60 samples), and sad (61 samples).

3.2. Benchmark systems

To gauge the benefits of the proposed method, two benchmark systems are used, namely BcLSTM and DialogueRNN and results are reported in Table 1 for task 1 and 2. BcLSTM is bi-directional RNN proposed by Poria et al. (2017). It is comprised of a two-step hierarchical training process. First, it extracts embeddings from each modality. For text, GloVe embeddings (Pennington et al., 2014) were used as input to a CNN-LSTM model to extract contextual representations for each utterance. For audio, Openmsile based features (Eyben, 2013) were input to an LSTM model to obtain audio representations for each audio utterance. Next, contextual representations from the audio and text modalities are fed to the BcLSTM model for emotion classification.

DialogueRNN, in turn, employs three stages of gated recurrent units (GRU) to model emotional context in conversations (Poria et al., 2019). The spoken utterances are fed into two GRUs: global and party GRU, to update the context and speaker state, respectively. In each turn, the party GRU updates its state based on i) the utterance spoken, ii) the speaker's previous state, and iii) the conversational context summarized by the global GRU through an attention mechanism. Finally, the updated speaker state is fed into the emotion GRU, which models the emotional information for classification. The attention mechanism is used on top of the emotion GRU to leverage contextual utterances by different speakers at various distances. Lastly, our proposed system comprises a feedforward DNN model and a 768- dimensional BERT(base) text feature vector fused (at the feature level) with a 311-dimensional vector comprised of eGEMAPs and MSF features.

TABLE 1 Benchmark system performance for the two AER tasks based on the MELD dataset.

	Task 1		Task 2	
Model	F1-score	BA	F1-score	BA
bcLSTM	0.70	0.72	0.82	0.83
DialogueRNN	0.72	0.72	0.84	0.85
Proposed system	0.74	0.73	0.87	0.87

3.3. Figures-of-merit

Balanced accuracy and F1-score are used as figures of merit to assess the performance of the proposed emotion classifier. In summary, precision shows us how many positive samples classified by the model are actually positive, i.e.,

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

Where TP corresponds to true positives and FP to false positives. Recall, in turn, calculates how many of the true positives are captured by the model. This is also called true positive rate or sensitivity and given by

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

Where FN corresponds to false negatives. Moreover, F1-score represents the harmonic mean of precision and recall and is useful in binary tasks where classes are unbalanced and is given by:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{3}$$

Lastly, balanced accuracy is given as the arithmetic mean of sensitivity (true positive rate or recall) and specificity (true negative rate) which, in turn, is given by:

$$Specificity = \frac{NP}{TN + FP}, \tag{4}$$

Where TN corresponds to true negatives. As such, balance accuracy (BA) is given as:

$$BA = \frac{Sensitivity + Specificity}{2}. \tag{5}$$

The interested reader is referred to Powers (2020) for more details on these classical performance metrics.

3.4. Quality scores

To gauge the improvements in quality and intelligibility of the enhancement algorithms, two objective speech quality measures are used, namely, PESQ and the short-term objective intelligibility (STOI) (Taal et al., 2011). While PESQ estimates the perceived speech quality on a 5-point mean opinion score scale ranging from bad to excellent, STOI measures the intelligibility of the signal on a 0–1 scale, with higher values suggesting greater intelligibility. Both methods are termed "intrusive" as they require access to the enhanced and a reference signal. More details on the PESQ and STOI measurement algorithms can be found in Rix et al. (2001).

4. Experimental results and discussion

In this section, we present and discuss the obtained experimental results.

4.1. Ablation study 1

In this first ablation experiment, we wish to explore the optimal set of text and speech features to include in the final system. We consider speech and text modalities separately in this study. We start with clean speech to find the best feature per modality and, subsequently, test the robustness of such set under unseen noisy conditions. In this study, babble and airport noises are considered. In both cases, the emotion classifier is trained on clean speech only. Table 2 shows the performance obtained for each modality individually for task 1. In the table, the feature termed 'fusion' corresponds to the fusion of MSF and eGeMAPS features.

As can be seen, for clean speech conditions and text-only AER, BERT-based text features resulted in the best performance across all metrics, hence corroborating previous reports (Yang et al., 2019; Stappen et al., 2021; Yang and Cui, 2021). As such, only BERT features are explored in the unseen noisy conditions. Babble noise is shown to degrade overall performance more severely than airport noise. Overall, BERT based features under 0 dB noise conditions are shown to achieve accuracy inline with that achieved by textCNN features under clean conditions, thus further suggesting improved robustness of the BERT text features. Given this finding, the final proposed system shown in Figure 1 will rely on BERT based text features.

As for speech features, under clean conditions eGeMAPS showed the highest overall performance of the three tested feature sets, thus corroborating findings by Eyben et al. (2013). Further gains could be seen with the fused feature set, however, thus suggesting the complementarity of spectral and modulation spectral features. As such, only the fused feature set is explored in the noisy mismatch condition. Moreover, similar to the text features, at low SNR levels, babble noise degraded performance more drastically compared to airport noise. Overall, the achieved performance with text-based features only was higher than what was achieved with audio features alone, thus corroborating the results reported by Patamia et al. (2021).

4.2. Ablation study 2

This second ablation study is an oracle experiment in which one modality in the multimodal system is kept clean and the other is corrupted by noise at varying levels and types. This study will allow us to gauge which modality is most sensitive to environmental factors and would benefit the most from speech enhancement. In all cases, the emotion classifier is trained on clean speech only. Table 3 show the performance obtained for Task 1 and Task 2, respectively.

As can be seen, the fusion of speech and text features in the clean condition (first row in the tables) showed improvements relative to each modality alone (i.e., Table 2) by as much as 2% for text and 7% for audio in terms of F1 score for Task 1. Furthermore, using noisy speech to generate "noisy" text resulted in more severe performance degradations for both Tasks, thus suggesting that more powerful machine-tuned enhancement algorithms may be useful for in-the-wild applications to assure the highest possible quality for text generation. Overall, on average, over the two types of noise, a drop of 32, 24, and 21% in F1 score was observed at 0, 10, 20 dB SNR levels relative to clean conditions, respectively, for Task 1. On the other hand, corrupting only the speech content had a less pronounced effect. Overall, on average, over the two types of noise, a drop of 16%, 13%, and 9% in F1 score was observed at 0, 10, 20 dB SNR levels over clean conditions for Task 1, respectively.

For Task 2, similar findings were observed. Overall, on average, over the two types of noise, a drop of 65, 33, and 28% in F1 score has been observed at 0, 10, and 20 dB SNR levels relative to clean conditions, respectively, when only text was corrupted. The drops in accuracy when the audio was corrupted were of 41, 27, and 25%, respectively. These findings corroborate those by Kessous et al. (2010) and Patamia et al. (2021) who showed that text modality achieved higher performance than audio in clean conditions. The drops in accuracy, however, under noisy conditions motivate the need for strategies to improve accuracy in the wild, as in the proposed system.

4.3. Ablation study 3

This third ablation study is an oracle experiment in which we wanted to test the hypothesis if we need two separate enhancement for improving ASR accuracy. As mentioned earlier, we used quality- (MetricGAN+) and ASR-optimized (mimic loss) enhancement algorithms for the speech and text branches shown in the proposed model in Figure 1. This study will allow us to gauge which combination of speech enhancement is better suited for this task. In all cases, the emotion classifier is trained on clean speech only. Table 4 show the performance obtained for Task 1 and Task 2.

As can be seen, for both Task 1 and Task 2, the best combination comprised the use of a quality-optimized enhancement algorithm for the top speech branch and an ASR-optimized (mimic loss) method for the bottom text branch. This combination resulted in the best accuracy for very extreme conditions (i.e., 0 dB SNR levels) and emphasizes the need for task-specific enhancement algorithms for AER.

TABLE 2 Ablation study 1: Performance comparison of different features for each individual modality.

Noise type	Feature	F1-score	BA	Feature	F1-score	BA
	Text			Audio		
Clean	BERT	0.72	0.76	Prosodic	0.62	0.61
Clean	TextCNN	0.56	0.54	eGEMAPS	0.69	0.67
Clean	BoW	0.62	0.59	MSF	0.66	0.67
Clean				Fusion	0.69	0.71
Airport (0 dB)	BERT	0.54	0.52	Fusion	0.51	0.51
Airport (10 dB)	BERT	0.60	0.57	Fusion	0.53	0.50
Airport (20 dB)	BERT	0.62	0.59	Fusion	0.55	0.52
Babble (0 dB)	BERT	0.58	0.56	Fusion	0.51	0.52
Babble (1 dB)	BERT	0.61	0.58	Fusion	0.52	0.51
Babble (20 dB)	BERT	0.61	0.58	Fusion	0.52	0.51

Feature termed "fusion" corresponds to the fusion of eGeMAPS and MSFs.

TABLE 3 Ablation study 2: Performance comparison of multimodal oracle system for Task 1 and Task 2.

		Task 1		Task 2	
Audio	Text	F1-score	BA	F1-score	BA
Clean	Clean	0.74	0.73	0.87	0.87
Clean	Airport (0 dB)	0.57	0.58	0.55	0.53
Clean	Airport (10 dB)	0.61	0.58	0.67	0.62
Clean	Airport (20 dB)	0.62	0.59	0.68	0.63
Clean	Babble (0 dB)	0.58	0.59	0.50	0.51
Clean	Babble (10 dB)	0.61	0.58	0.63	0.58
Clean	Babble (20 dB)	0.61	0.58	0.67	0.62
Airport (0 dB)	Clean	0.65	0.62	0.60	0.66
Airport (10 dB)	Clean	0.65	0.63	0.68	0.68
Airport (20 dB)	Clean	0.68	0.65	0.70	0.68
Babble (0 dB)	Clean	0.62	0.60	0.63	0.67
Babble (10 dB)	Clean	0.65	0.62	0.68	0.67
Babble (20 dB)	Clean	0.68	0.65	0.69	0.67

TABLE 4 Ablation study 3: Performance comparison of enhancement system for Task 1 and Task 2.

			Task 1		Task 2	
Noise	Enhancement-1	Enhancement-2	F1-score	BA	F1-score	BA
Airport (0 dB)	MetricGAN+	MetricGAN+	0.60	0.60	0.53	0.50
	MetricGAN+	Mimic-loss	0.65	0.64	0.56	0.51
	Mimic-loss	MetricGAN+	0.61	0.59	0.54	0.52
	Mimic-loss	Mimic-loss	0.61	0.60	0.55	0.52
Babble (0 dB)	MetricGAN+	MetricGAN+	0.59	0.59	0.56	0.51
	MetricGAN+	Mimic-loss	0.62	0.59	0.57	0.51
	Mimic-loss	MetricGAN+	0.60	0.60	0.56	0.51
	Mimic-loss	Mimic-loss	0.61	0.61	0.56	0.52

4.4. Overall system performance

This last study explores the performance of the proposed system described in Figure 1, combining speech enhancement optimized for each branch (speech and text), as well as data augmentation to provide robustness at the model training level. Data augmentation methods are useful to solve imbalanced data problems. It also helps the model to learn the complex distribution of the data and helps prevent overfitting. The work by Hu et al. (2018) showed that adding noisy versions of the clean speech data to the training set improved speech recognition accuracy in mismatched noisy conditions. Therefore, in this work, we utilized the same strategy. Table 5 show the obtained results in rows labeled 'Data augmentation only' for Task 1 and Task 2. As can be seen, data augmentation alone already improved AER results, thus corroborating findings by Trinh et al. (2021); Neumann and Vu (2021), and Kshirsagar and Falk (2022a,b).

Next, we gauge the benefits of using speech enhancement alone. As before, AER models are trained solely on clean speech. During run time, we pre-process the test data with the MetricGAN+ algorithm for the speech branch and the mimic loss enhancer for the text branch, as described in Section 2. Table 5, show the obtained results in rows labeled 'Enhancement only'. As can be seen, applying speech enhancement improves overall performance relative to the noisy conditions, but the final results are still below what was achieved in clean conditions, as well as what was achieved with data augmentation. The gains observed were typically more substantial at low SNR values, thus corroborating results by Triantafyllopoulos et al. (2019).

In an attempt to better understand the reason behind the poor AER performance with speech enhancement alone, Figure 3 depicts an average modulation spectrogram, from top to bottom, for clean, noisy (airport at 0 dB SNR), MetricGAN+, and mimic-loss enhanced speech for angry (left) and sad (right) emotions, respectively. Modulation spectrograms are a frequency-frequency representation where the y-axis depicts acoustic frequency and the x-axis modulation frequency. From the clean plot, we can see the typical speech modulation spectral representation with most modulation energy lying below 16 Hz (Falk and Chan, 2010a) and a slowing of the amplitude modulations with the sad emotion (Wu et al., 2011). Noise, in turn, is shown to affect the modulation spectrogram by smearing the energy across higher acoustic and modulation frequencies, as suggested by Falk et al. (2010). The enhancement algorithms, however, are not capable of completely removing these environmental artifacts and seem to be introducing other types of distortions that can make the AER task more challenging. Combined, these factors result in the reduced gains reported in the Tables. This was in fact confirmed by listening to the outputs of the MetricGAN+ enhancement algorithm. We have also presented the PESQ, and STOI scores in Table 6. This verifies the significance of having task-specific enhancement for improving the AER performance in noisy conditions.

TABLE 5 Performance comparison of the proposed method in different noisy test conditions for Task 1 and Task 2.

	Task 1		Task 2	
Signal	F1-score	BA	F1-score	BA
Clean	0.74	0.73	0.87	0.87
Noisy-Airport (−20 dB)	0.49	0.49	0.43	0.53
Data augmentation only	0.51	0.49	0.53	0.52
Enhancement only	0.56	0.52	0.51	0.48
Proposed	0.56	0.54	0.52	0.50
Noisy-Airport (−10 dB)	0.53	0.46	0.44	0.57
Data augmentation only	0.53	0.52	0.52	0.50
Enhancement only	0.57	0.52	0.54	0.51
Proposed	0.59	0.54	0.57	0.56
Noisy-Airport (0 dB)	0.57	0.55	0.50	0.50
Data augmentation only	0.67	0.68	0.61	0.61
Enhancement only	0.65	0.64	0.56	0.51
Proposed	0.65	0.63	0.62	0.59
Noisy-Airport (10 dB)	0.59	0.57	0.55	0.51
Data augmentation only	0.69	0.70	0.66	0.66
Enhancement only	0.68	0.65	0.61	0.55
Proposed	0.71	0.69	0.65	0.62
Noisy-Airport (20 dB)	0.60	0.58	0.60	0.55
Data augmentation only	0.69	0.68	0.67	0.66
Enhancement only	0.67	0.65	0.62	0.56
Proposed	0.71	0.69	0.67	0.65
Noisy-Babble(−20 dB)	0.52	0.49	0.49	0.49
Data augmentation only	0.52	0.49	0.54	0.54
Enhancement only	0.57	0.52	0.54	0.51
Proposed	0.58	0.58	0.56	0.51
Noisy-Babble (−10 dB)	0.54	0.51	0.52	0.51
Data augmentation only	0.56	0.51	0.55	0.52
Enhancement only	0.59	0.54	0.54	0.51
Proposed	0.56	0.52	0.59	0.54
Noisy-Babble (0 dB)	0.59	0.57	0.54	0.51
Data augmentation only	0.66	0.66	0.58	0.59
Enhancement only	0.62	0.59	0.57	0.51
Proposed	0.64	0.61	0.61	0.58
Noisy-Babble (10 dB)	0.60	0.58	0.58	0.54
Data augmentation only	0.72	0.71	0.63	0.62
Enhancement only	0.68	0.66	0.61	0.55
Proposed	0.70	0.68	0.66	0.64
Noisy-Babble (20 dB)	0.61	0.58	0.61	0.56
Data augmentation only	0.74	0.72	0.67	0.67
Enhancement only	0.70	0.67	0.66	0.60
Proposed	0.70	0.69	0.67	0.64

Finally, we test the combined effects of speech enhancement and data augmentation, as in the proposed system, to gauge the benefits of noise robustness applied at both the input and model levels, respectively. For Task 1, gains (relative to

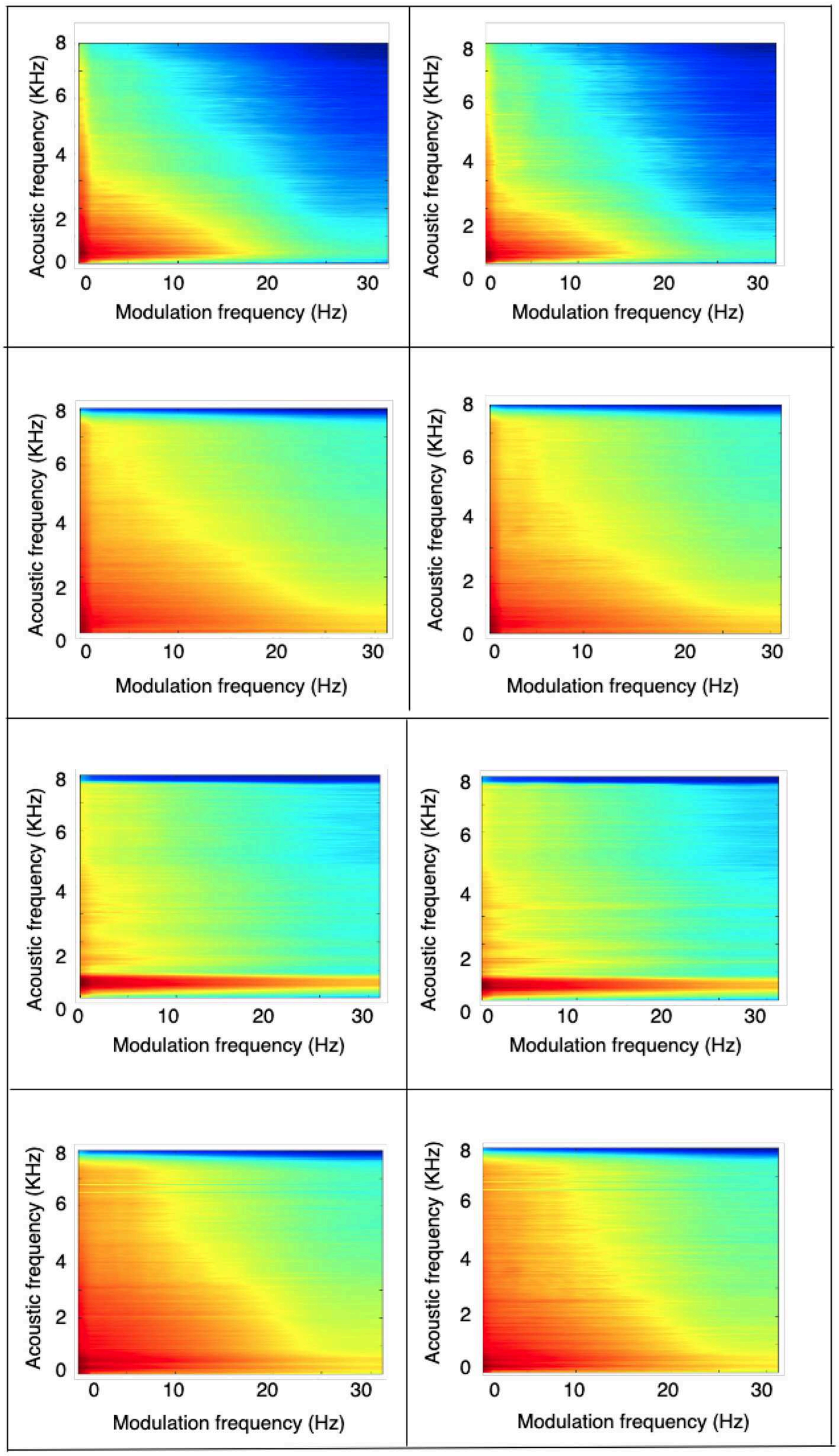


FIGURE 3
Modulation spectrogram for different conditions, from **(top–bottom)**: clean, (airport) noisy at 0 dB, MetriGAN+, and mimic-loss enhanced speech. **(Left)** plots correspond to angry and **(right)** plots to sad emotion.

TABLE 6 Performance comparison of PESQ and STOI score.

Signal	PESQ	STOI
Noisy-Airport (−20 dB)	1.107	0.219
MetricGAN+	1.130	0.312
MimicLoss	1.054	0.244
Noisy-Airport (−10 dB)	1.094	0.367
MetricGAN+	1.132	0.412
MimicLoss	1.112	0.368
Noisy-Airport (0 dB)	1.102	0.620
MetricGAN+	1.225	0.657
MimicLoss	1.112	0.627
airport (10 dB)	1.583	0.791
MetricGAN+	1.899	0.812
MimicLoss	1.622	0.800
Noisy-Airport (20 dB)	2.800	0.885
MetricGAN+	2.979	0.895
MimicLoss	2.894	0.886
Noisy-abble (−20 dB)	1.100	0.188
MetricGAN+	1.154	0.254
MimicLoss	1.038	0.229
Noisy-Babble (−10 dB)	1.103	0.342
MetricGAN+	1.151	0.363
MimicLoss	1.138	0.356
Noisy-Babble (0 dB)	1.139	0.576
MetricGAN+	1.229	0.639
MimicLoss	1.180	0.591
Noisy-Babble (10 dB)	1.577	0.764
MetricGAN+	1.939	0.789
MimicLoss	1.605	0.768
Noisy-Babble (20 dB)	2.792	0.871
MetricGAN+	2.968	0.880
MimicLoss	2.799	0.871

TABLE 7 Performance comparison of the proposed method in unseen noise and SNR levels for Task 1.

	Task 1		Task 2	
Signal	F1-score	BA	F1-score	BA
Noisy - Cafeteria (5dB)	0.58	0.57	0.52	0.47
Data augmentation only	0.63	0.60	0.67	0.64
Enhancement only	0.66	0.64	0.64	0.59
Proposed	0.68	0.65	0.67	0.65
Noisy - Cafeteria (15dB)	0.61	0.58	0.54	0.48
Data augmentation only	0.65	0.62	0.68	0.63
Enhancement only	0.69	0.66	0.67	0.61
Proposed	0.70	0.69	0.71	0.70

TABLE 8 Cross-corpus performance on unseen IEMOCAP and Emoti-W datasets for Tasks 1 and 2.

		Task 1		Task 2	
Experiment	Dataset	F1-score	BA	F1-score	BA
1	IEMOCAP	0.94	0.94	0.85	0.85
2		0.49	0.55	0.50	0.60
3		0.64	0.69	0.72	0.70
1	Emoti-W	0.67	0.66	0.61	0.62
2		0.46	0.52	0.48	0.53
3		0.58	0.60	0.56	0.56

using each strategy individually) were seen for the airport noise condition at higher and lower SNR conditions. In fact, with data augmentation alone, accuracy inline with what was achieved with clean speech was obtained. For Task 2, in turn, the proposed model showed improvements over the other methods for almost all tested conditions in terms of F1 score, thus showing the importance of the proposed method to classify between opposing emotions in extremely noisy scenaerios; in the case here, joy vs. sad. Notwithstanding, for Task 2 a gap of 23% remained between the best achieved performance and the clean speech accuracy. Furthermore, we also tested the generalization ability of the proposed system using unseen Cafeteria noise type and unseen SNR levels such as 5 dB and 15 dB. As can be seen in Table 7 the model was able to generalize across mismatched noise types and noise levels with significant performance gain with the proposed methodology. For comparison purposes, the state-of-the-art DialogueRNN system achieved an F1 score of 0.59 and 0.55 for Task 1 and Task 2, respectively, when corrupted with airport noise at 0 dB. The proposed system, in turn, was able to outperform this benchmark by 10 and 12%, respectively. Overall, the obtained results suggest that data augmentation combined with speech enhancement can be a viable alternative for robust in-the-wild automatic multimodal emotion recognition while requiring access to only one signal modality: audio.

4.5. Generalizability of proposed method

To test the generalizability of the proposed method, six additional experiments have been conducted on IEMOCAP and Emoti-W datasets. First, we retrain the proposed AER model using the IEMOCAP training dataset partition and test it on the IEMOCAP test set to obtain an upper bound on what can be achieved on this particular dataset. Next, to gauge the advantages brought by the proposed system, we retrain the AER system shown in Figure 1 but without the enhancement and data augmentation steps. Training was done on the MELD dataset and the model was then tested on the unseen IEMOCAP test data and the unseen Emoti-W testset. This gives us an idea of how challenging the cross-corpus task is when the proposed innovations are not present and should give us a lower bound on what could be achieved cross-corpus. Finally, we tested the full proposed method trained on the MELD dataset and tested on the unseen IEMOCAP and Emoti-W test data. Experimental results are reported in Table 8. As can be seen, cross-corpus testing is an extremely challenging task where performance accuracy can drop to chance levels if strategies are not put in place. The proposed innovations, on the other hand, provides some robustness, and gains of 30% and 44% on IEMOCAP and 26% and 17% on Emoti-W could be seen with the proposed system for Tasks 1 and 2, respectively, over a system without task-specific speech enhancement and data augmentation. The gaps to the upper bound obtained with Experiment 1 suggest that there is still room for improvement and emotion-aware enhancement and/or alternate data augmentation strategies may still be needed.

5. Conclusions

This paper has explored the use of task-specific speech enhancement combined with data augmentation to provide robustness to unseen test conditions for multimodal emotion recognition systems. Experiments conducted on the MELD dataset show the importance of BERT for text feature extraction and a fused eGEMAPS-modulation spectral set for audio features. The importance of data augmentation at the training stage and of task-specific speech enhancement at the testing stage are shown on two binary speech emotion classification tasks. Lastly, cross-corpus experiments showed the proposed innovations resulting in 40% gains relative to an AER system without enhancement/augmentation. While the obtained results suggest that task-specific enhancement, combined with data augmentation are important steps toward reliable “in the wild” emotion recognition, speech enhancement algorithms may still be suboptimal and may be removing important emotion information. As such, future work should explore the development of emotion-aware enhancement algorithms that can trade-off noise suppression and emotion recognition accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SK and TF: conceptualization. TF: funding acquisition and writing-review and editing. SK: methodology, software, validation, visualization, and writing-original draft. Part of this initial work was done during AP’s MITAC Globalink internship, he worked on experimentation in Tables 1–3. Therefore, he is considered a second author in this paper. All authors contributed to the article and approved the submitted version.

Funding

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada and MITACS via their Globalink program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alston, W. P. (1964). Philosophy of Language. *J. Philos. Logic*. 2, 458–508.

Avila, A., Akhtar, Z., Santos, J., O’Shaughnessy, D., and Falk, T. (2021). Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Trans. Affect. Comput*. 12, 177–188. doi: 10.1109/TAFFC.2018.2858255

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. *arXiv preprint* arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477

Bagchi, D., Plantinga, P., Stiff, A., and Fosler-Lussier, E. (2018). “Spectral feature mapping with mimic loss for robust speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5609–5613.

Banse, R., and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol*. 70, 614. doi: 10.1037/0022-3514.70.3.614

Braun, S., Schwartz, B., Gannot, S., and Habets, E. A. (2016). “Late reverberation psd estimation for single-channel dereverberation using relative convolutive transfer functions,” in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)* (Xi’an: IEEE), 1–5.

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukić, A., Gerkmann, T., et al. (2015). Combination of mvdr beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP J. Adv. Signal Process*. 2015, 61. doi: 10.1186/s13634-015-0242-x

Chen, S., Zhang, M., Yang, X., Zhao, Z., Zou, T., and Sun, X. (2021). The impact of attention mechanisms on speech emotion recognition. *Sensors* 21, 7530. doi: 10.3390/s21227530

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: an emotion corpus of multi-party conversations. *arXiv preprint* arXiv:1802.08379. doi: 10.48550/arXiv.1802.08379

Chuang, Z.-J., and Wu, C.-H. (2004). “Multi-modal emotion recognition from speech and text,” in *International Journal of Computational Linguistics Chinese Language Processing, Volume* 9*, Number 2, August 2004: Special Issue on New Trends of*

Speech and Language Processing, 45–62. Available online at: https://aclanthology.org/O04-3004.pdf

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). "From individual to group-level emotion recognition: emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow).

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., and Schmauch, B. (2018). Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint* arXiv:1802.05630. doi: 10.21437/SMM.2018-5

Eyben, F. (2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia* (Barcelona).

Eyben, F., Scherer, K., Schuller, B., and Sundberg, J., André, E., Busso, C., et al. (2016). The geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 417. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., and Schuller, B. (2013). "Affect recognition in real-life acoustic conditions-a new perspective on feature selection," in *Proceedings 14th INTERSPEECH* (Lyon).

Falk, T., and Chan, W. (2010a). Modulation spectral features for robust far-field speaker identification. *IEEE Trans Audio Speech Lang Process.* 18, 90–100. doi: 10.1109/TASL.2009.2023679

Falk, T., and Chan, W. (2010b). Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Trans. Instrum Meas.* 59, 24697. doi: 10.1109/TIM.2009.2024697

Falk, T., Zheng, C., and Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. Audio Speech Lang. Process.* 18, 1766–1774. doi: 10.1109/TASL.2010.2052247

Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. (2019). "Metricgan: generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning (PMLR)* (Long Beach, CA), 2031–2041.

Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., et al. (2021). Metricgan+: an improved version of metricgan for speech enhancement. *arXiv preprint* arXiv:2104.03538. doi: 10.21437/Interspeech.2021-599

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567.* doi: 10.48550/arXiv.1412.5567

Haytham, F., Lech, M., and Lawrence, C. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Netw.* 92, 60–68. doi: 10.1016/j.neunet.2017.02.013

Hirsch, H., and Pearce, D. (2000). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)* (Paris).

Ho, N.-H., Yang, H.-J., Kim, S.-H., and Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* 8, 61672–61686. doi: 10.1109/ACCESS.2020.2984368

Hu, H., Tan, T., and Qian, Y. (2018). "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5044–5048.

Jin, Q., Li, C., Chen, S., and Wu, H. (2015). "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD: IEEE), 4749–4753.

Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Interfaces* 3, 33–48. doi: 10.1007/s12193-009-0025-5

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., and Prendinger, H. (2018). Deep learning for affective computing: text-based emotion recognition in decision support. *Decis. Support. Syst.* 115, 24–35. doi: 10.1016/j.dss.2018.09.002

Kshirsagar, S., and Falk, T. H. (2022a). Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *Sensors* 22, 6445. doi: 10.3390/s22176445

Kshirsagar, S. R., and Falk, T. H. (2022b). Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Tran. Affect. Comput.* 13, 1892–1905. doi: 10.1109/TAFFC.2022.3188223

Lee, J., and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Interspeech* 2015, 336. doi: 10.21437/Interspeech.2015-336

Li, C., Bao, Z., Li, L., and Zhao, Z. (2020). Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Inf. Process. Manag.* 57, 102185. doi: 10.1016/j.ipm.2019.102185

Li, T.-M., Chao, H.-C., and Zhang, J. (2019). Emotion classification based on brain wave: a survey. *Hum. Centric Comput. Inf. Sci.* 9, 1–17. doi: 10.1186/s13673-019-0201-x

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). "Speech enhancement based on deep denoising autoencoder," in *Interspeech* (Lyon), 436–440.

Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., et al. (2022). Automated emotion recognition: current trends and future perspectives. *Comput. Methods Programs Biomed.* 2022, 106646. doi: 10.1016/j.cmpb.2022.106646

Malte, A., and Ratadiya, P. (2019). "Multilingual cyber abuse detection using advanced transformer architecture," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (Kochi: IEEE), 784–789.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affect. Comput.* 3, 184–198. doi: 10.1109/T-AFFC.2011.40

Mower, E., Matarić, M. J., and Narayanan, S. (2010). A framework for automatic human emotion classification using emotion profiles. *IEEE Trans. Audio Speech Lang. Process.* 19, 1057–1070. doi: 10.1109/TASL.2010.2076804

Naumann, A. B., Wechsung, I., and Hurtienne, J. (2009). "Multimodal interaction: Intuitive, robust, and preferred?" in *IFIP Conference on Human-Computer Interaction* (Uppsala: Springer), 93–96.

Neumann, M., and Vu, N. T. (2021). "Investigations on audiovisual emotion recognition in noisy conditions," in *2021 IEEE Spoken Language Technology Workshop (SLT)* (Shenzhen: IEEE), 358–364.

Parent, M., Tiwari, A., Albuquerque, I., Gagnon, J.-F., Lafond, D., Tremblay, S., et al. (2019). "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 4131–4136.

Park, M. W., Kim, C. J., Hwang, M., and Lee, E. C. (2013). "Individual emotion classification between happiness and sadness by analyzing photoplethysmography and skin temperature," in *2013 Fourth World Congress on Software Engineering* (Hong Kong: IEEE), 190–194.

Parveen, S., and Green, P. (2004). "Speech enhancement with missing data techniques using recurrent neural networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1* (Montreal, QC: IEEE), I-733.

Pascual, S., Bonafonte, A., and Serra, J. (2017). Segan: Speech enhancement generative adversarial network. *arXiv:1703.09452.* doi: 10.21437/Interspeech.2017-1428

Patamia, R. A., Jin, W., Acheampong, K. N., Sarpong, K., and Tenagyei, E. K. (2021). "Transformer based multimodal speech emotion recognition with improved neural networks," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)* (Chengdu: IEEE), 195–203.

Patrick, K., and Lavery, J. F. (2007). Burnout in nursing. *Aust. J. Adv. Nurs.* 24, 43.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (2017). "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)* (Vancouver, BC), 873–883.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: a multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint* arXiv:1810.02508. doi: 10.18653/v1/P19-1050

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019). Emotion recognition in conversation: research challenges, datasets, and recent advances. *IEEE Access* 7, 100943–100953. doi: 10.1109/ACCESS.2019.2929050

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint* arXiv:2010.16061. doi: 10.48550/arXiv.2010.16061

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice).

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 2* (Salt Lake City, UT: IEEE), 749–752.

Sangwan, S., Chauhan, D. S., Akhtar, M., Ekbal, A., Bhattacharyya, P., et al. (2019). "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *International Conference on Neural Information Processing* (Vancouver, BC: Springer), 662–669.

Sebe, N., Cohen, I., and Huang, T. S. (2005). Multimodal emotion recognition. *Handbook Pattern Recogn. Comput. Vis.* 4, 387–419. doi: 10.1142/9789812775320_0021

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., and Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* 8, 176274–176285. doi: 10.1109/ACCESS.2020.3026823

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E.-M., et al. (2021). The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. *arXiv preprint* arXiv:2104.07123. doi: 10.1145/3475957.3484450

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* 19, 2125–2136. doi: 10.1109/TASL.2011.2114881

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint* arXiv:1905.06316. doi: 10.48550/arXiv.1905.06316

Thiemann, J., Ito, N., and Vincent, E. (2013). "The diverse environments multichannel acoustic noise database (demand): a database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013, volume 19* (Montreal, QC: ASA).

Triantafyllopoulos, A., Keren, G., Wagner, J., Steiner, I., and Schuller, B. W. (2019). "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Interspeech* (Graz), 1691–1695.

Trinh, V. A., Kavaki, H. S., and Mandel, M. I. (2021). Importantaug: a data augmentation agent for speech. *arXiv preprint* arXiv:2112.07156. doi: 10.1109/ICASSP43922.2022.9747003

Tripathi, S., Tripathi, S., and Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint* arXiv:1804.05788. doi: 10.48550/arXiv.1804.05788

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* 11, 1301–1309. doi: 10.1109/JSTSP.2017.2764438

Valstar, M. (2016). "Avec 2016: depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (Amsterdam: ACM), 3–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wu, S., Falk, T., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* 53, 768–785. doi: 10.1016/j.specom.2010.08.013

Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., et al. (2020). Impact of covid-19 pandemic on mental health in the general population: a systematic review. *J. Affect. Disord.* 277, 55–64. doi: 10.1016/j.jad.2020.08.001

Xue, W., Cucchiarini, C., van Hout, R., and Strik, H. (2019). "Acoustic correlates of speech intelligibility: the usability of the egemaps feature set for atypical speech," in *Proceedings of 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)* (Graz), 48–52.

Yang, K., Lee, D., Whang, T., Lee, S., and Lim, H. (2019). Emotionx-ku: BERT-max based contextual emotion classifier. *arXiv preprint* arXiv:1906.11565. doi: 10.48550/arXiv.1906.11565

Yang, Y., and Cui, X. (2021). Bert-enhanced text graph neural network for classification. *Entropy* 23, 1536. doi: 10.3390/e23111536

Zeng, S. (2007). Audio-visual affect recognition. *IEEE Trans. Multimedia* 9, 424–428. doi: 10.1109/TMM.2006.886310

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., and Zhou, G. (2019). "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *IJCAI* (Macao), 5415–5421.

Zhang, Y., Wang, Q., Li, Y., and Wu, X. (2018). Sentiment classification based on piecewise pooling convolutional neural network. *Comput. Mater. Continua* 56, 285–297.

Zhao, H., Zarar, S., Tashev, I., and Lee, C.-H. (2018). "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 2401–2405.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d and 2d cnn lstm networks. *Biomed. Signal Process. Control.* 47, 312–323. doi: 10.1016/j.bspc.2018.08.035

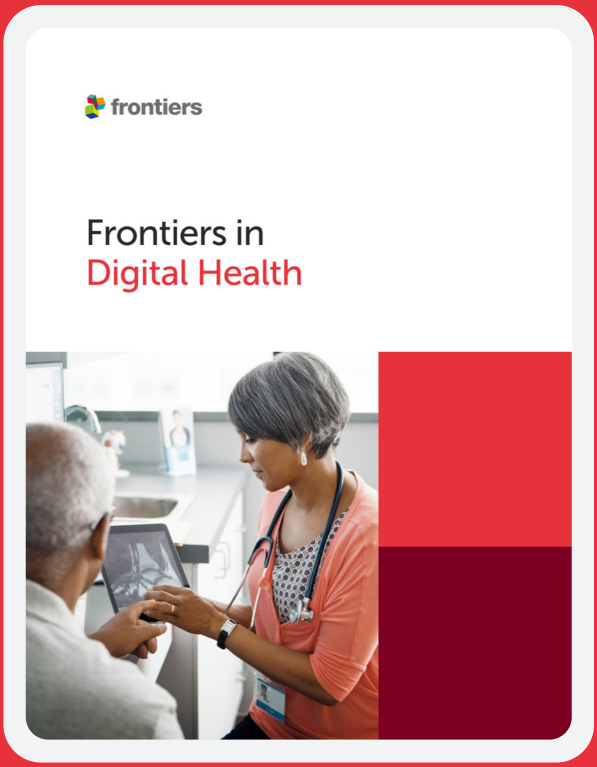
Frontiers in Digital Health

Explores digital innovation to transform modern healthcare

A multidisciplinary journal that focuses on how we can transform healthcare with innovative digital tools. It provides a forum for an era of health service marked by increased prediction and prevention.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Digital Health

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact