frontiers | Frontiers in Signal Processing

TYPE Review
PUBLISHED 16 August 2024
DOI 10.3389/frsip.2024.1339159

Check for updates

OPEN ACCESS

EDITED BY
Ashwani Kumar,
Bennett University, India

REVIEWED BY
Shahan Nercessian,
iZotope, United States
Nikola Simić,
University of Novi Sad Faculty of Technical Sciences, Serbia

*CORRESPONDENCE
Anders R. Bargum,
✉ arba@create.aau.dk

RECEIVED 15 November 2023
ACCEPTED 15 July 2024
PUBLISHED 16 August 2024

CITATION
Bargum AR, Serafin S and Erkut C (2024) *Reimagining speech*: a scoping review of deep learning-based methods for non-parallel voice conversion.
Front. Sig. Proc. 4:1339159.
doi: 10.3389/frsip.2024.1339159

Reimagining speech: a scoping review of deep learning-based methods for non-parallel voice conversion

Anders R. Bargum[1,2]*, Stefania Serafin[1] and Cumhur Erkut[1]

[1]Multi-Sensory Experience Laboratory, Department of Architecture, Design and Media Technology, Aalborg University, Copenhagen, Denmark, [2]Heka, Khora VR, Copenhagen, Denmark

Research on deep learning-powered voice conversion (VC) in speech-to-speech scenarios are gaining increasing popularity. Although many of the works in the field of voice conversion share a common global pipeline, there is considerable diversity in the underlying structures, methods, and neural sub-blocks used across research efforts. Thus, obtaining a comprehensive understanding of the reasons behind the choice of the different methods included when training voice conversion models can be challenging, and the actual hurdles in the proposed solutions are often unclear. To shed light on these aspects, this paper presents a scoping review that explores the use of deep learning in speech analysis, synthesis, and disentangled speech representation learning within modern voice conversion systems. We screened 628 publications from more than 38 venues between 2017 and 2023, followed by an in-depth review of a final database of 130 eligible studies. Based on the review, we summarise the most frequently used approaches to voice conversion based on deep learning and highlight common pitfalls. We condense the knowledge gathered to identify main challenges, supply solutions grounded in the analysis and provide recommendations for future research directions.

KEYWORDS

voice conversion, voice transformations, voice control, deep learning, disentanglement, speech representation learning

1 Introduction

Voice transformations (VT) describe the act of controlling non-linguistic characteristics of speech, such as the quality or the individuality of a vocal signal (Stylianou, 2009). The expression “transformation” is used as an umbrella term referring to the modifications made in a speech-to-speech scenario where an application or a technical system is used to map, modify, or modulate specific characteristics of a voice, be it pitch, timbre, or prosody.

A sub-task of VT is the topic of voice conversion (VC). More specifically, VC seeks to render an utterance from one speaker to sound like that of a target speaker. In the past decade, it has become a prominent research subject within the field of artificial intelligence (AI). Most commonly, voice conversion refers to the process of changing the properties of speech, such as voice identity, emotion, language or accent, and the process has, in the past years, made a major impact on several real-life applications such as personalised speech synthesis, communication aids for speech impaired, or simple voice mimicry. It should be noted that VC is also used to describe the conversion procedure of a text-to-speech (TTS) pipeline in which a user chooses specific speaker characteristics that the written text should sound like (Sisman et al., 2020). This review considers only the former definition.

The VC pipeline can be divided into three main stages, each aiming at solving a specific problem: 1) *the speech analysis* stage aims at breaking down speech signals into intermediate representations, facilitating efficient manipulation or modification based on the acoustic properties of speech. For prosody modifications, it is important to decompose pitch and rhythm-related properties, whereas, for identity conversion, it is important to extract and disentangle linguistic content from speaker timbral information, 2) *the mapping stage* where a system transfers the decomposed information from stage 1 towards a representation that matches the qualities of a specific target speaker, and 3) *the reconstruction and synthesis stage,* where the transformed intermediate representation is processed and re-synthesised into the time domain using a generator or vocoder (Walczyna and Piotrowski, 2023). All three stages may be carried out using traditional signal processing or statistical modelling techniques. During the *speech analysis* stage, one may represent a speech signal as overlapping segments of pitch periods using the pitch synchronous overlap and add (PSOLA) method or as frames of pitch varying excitation signals and vocal tract filters based on the mel-log spectrum (Sisman et al., 2020). Assuming that parallel data, that is, the same utterances spoken by both the input and target speaker, are available, the mapping stage may be carried out through the procedure of *prosody and spectrum mapping*. This has most commonly been accomplished by Gaussian mixture models (GMM) (Stylianou et al., 1998), non-negative matrix factorisation (NMF) (Wu et al., 2013), and regression-based clustering methods (Zhou et al., 2020). Lastly, the syntheses have traditionally been executed through techniques based on the inverse Fourier transforms or PSOLA (Valbret et al., 1992).

Nonetheless, the traditional approaches suffer from several limitations. The manipulation of time-domain signals, as done through PSOLA, is complex and rarely results in good audio quality as it mostly ignores phase relationships when mapping acoustic features (Valbret et al., 1992). Simultaneously, assuming a stationary process in time-invariant linear source-filter methods often gives rise to unnatural-sounding voices. The progress in artificial intelligence deep learning modules has, therefore, gradually been incorporated as primary foundational elements for each stage in the VC pipeline. There are several benefits to this. First, the feature-mapping processes are generally non-linear, making non-linear deep learning operations more compatible with human speech than methods based on linear operations such as GMMs. Second, neural networks are end-to-end compatible and can learn from and generalise to large datasets. Nearly all contributions to the biannual Voice Conversion Challenge (VCC) incorporate neural networks at some point in the VC pipeline, with a predominant number of submitted works being entirely based on deep learning principles (Yi et al., 2020). With the increasing interest in deep learning, new non-parallel end-to-end training methods, novel mapping functions, and vocoding techniques have been promoted. Consequently, these advancements have led to substantial improvements in the quality and fidelity of VC results in terms of naturalness, realism, and conversion quality. Deep learning has, therefore, become “a new standard” for carrying out voice conversion today, which is why this review has chosen to focus solely on deep learning VC techniques, mainly in the realm of non-parallel VC.

Few works have reviewed the field of VC. Sisman et al. (2020) provide a comprehensive overview of the history of VC technology, including statistical approaches and neural networks, and identify common deep learning modules. Walczyna and Piotrowski (2023) extend this work by focussing on frequently used deep learning models for analysis, mapping, and synthesis. However, both reviews have limitations. Sisman et al. (2020) emphasise the historical context but lack coverage of contemporary methods and their integration within broader settings. Walczyna and Piotrowski (2023) focus on current techniques but do not connect them with the diverse challenges in VC. Both reviews also lack a forward-looking perspective, offering limited suggestions for future research directions. This is significant as identifying emerging challenges and opportunities is crucial for advancing VC technology. Our contribution supplements these reviews by analysing key problem areas in current VC research and focussing on foundational elements for targeted solutions. We aim to bridge the gap by highlighting contemporary methods and connecting them to broader challenges and applications in VC. Additionally, we propose a roadmap for future research, emphasising interdisciplinary approaches and novel applications. We seek to provide an overview of research trends, techniques, and challenges in VC, especially covering the rapid growth in the last few years.

Specifically, we offer an illustrative and statistical examination of the prevalence and applications of deep learning-based methods used within current VC pipelines. To accomplish this, we undertake a scoping review, building upon a succession of similar reviews conducted within the realm of multi-sensory audio signal processing. Each of these prior reviews addressed distinct research inquiries and employed unique methodological approaches. Brice et al. (2023) developed a framework for contemporary hearing care using a PRISMA approach to identify service and product delivery options. Paisa et al. (2023) focused on tactile displays for auditory augmentation, categorising devices based on physical, auditory, perceptual, purpose, and evaluation domains. Salinas-Marchant and MacLeod (2022) explored audiovisual speech perception in children, identifying key gaps in research. Our review aligns with these works in its methodological rigour but diverges in its focus on deep learning-based VC. By extending the library of encoders and loss functions, we contribute to a more comprehensive understanding of the current landscape and potential future directions in this rapidly evolving field.

In this scoping review, we analyse 130 papers published between 2017 and 2023 to identify common approaches and areas needing development in VC. Our analysis includes a quantitative examination of training configurations and thematic analysis guided by a 14-code codebook. We offer an intuitive overview of the VC pipeline and provide a dataset with reviewed papers, codes, and keywords. Graphical representations illustrate work distributions, while a detailed analysis addresses specific VC challenges. The review concludes with insights into future research directions, focussing on identity conversion, interpretability, and real-time control. We aim for this review to be a valuable resource for newcomers to the field of voice conversion, aiding understanding of common techniques and guiding research efforts.

2 Background

Before presenting the methodology, we introduce general terminology and describe the sub-blocks and typical flow of a traditional deep learning-based VC pipeline. VC occurs in various forms, including one-to-one, one-to-many, or many-to-one conversions. In one-to-one conversion, the voice of one speaker is converted to another specific speaker. One-to-many conversion involves converting the voice of a single speaker to multiple other speakers, while many-to-one conversion changes the voices of multiple speakers to a single target speaker. Additionally, there exists one-shot, many-to-many conversion, sometimes referred to as any-to-any or zero-shot voice conversion. In this approach, a model can generalise from only one utterance of several speakers, including both those seen and unseen during training. Although certain studies encountered in this review may involve one-to-one or one-to-many conversion schemes, specifically for spectral mapping, the predominant focus will be centred on methods pertinent to many-to-many or zero-shot voice conversion. These instances not only offer generalisability to the remaining cases but also stand at the forefront of harnessing the non-parallel benefits facilitated by deep learning. The following section serves as a brief overview of the non-parallel voice conversion pipeline as well as the codes and descriptions employed throughout the coding, analysis, and synthesis procedures.

2.1 The voice conversion pipeline

Research on speech analysis and synthesis has been conducted since 1922 when Stewart (1922) commented that the difficult problems involved in artificial voice production are found in the manipulation of the apparatus producing it rather than the actual production of the speech itself. In order to understand and implement speech analysis, production, and manipulation blocks in such an apparatus, we utilise the ability to characterise speech by different factors. First, one can define speech by its *linguistic factors*, which are reflected in sentence structure, lexical corpus, and idiolect, for example, words and speech habits. Second, we divide it into *supra-segmental factors*, which are the prosodic attributes of speech, such as intonation, stress, and rhythm, and *segmental factors*, which are related to speaker identity and timbre, such as the spectrum, spectral envelope, and formants (Sisman et al., 2020). These aspects can be interchanged and mapped in various ways. Most deep learning-powered VC approaches train a conversion model that transforms either the segmental factors, that is, changes the timbre in order to match that of a target speaker, or the supra-segmental factors, that is, transforming prosody, such as pitch and rhythm. The former aspects result in pure identity conversion, whereas the latter are important when converting input speech to an accent or emotion of another speaker. Both processes are transformed with the main goal of keeping the linguistics unchanged. We illustrate the typical voice conversion pipeline, including analysis, mapping, and reconstruction modules in Figure 1 and divide the stages depicted into even more specific tasks.

Most VC pipelines start by extracting information about linguistic content, prosody, and speaker identity individually. In the field of VC, we denote the segmental and supra-segmental aspects of a speaker as *speaker-dependent* features, that is, features that capture the specific vocal characteristics of an individual speaker, while the remaining linguistics are denoted as *speaker-independent* features, that is, features that describe spoken content universal to any spoken language. Once the speaker-dependent and speaker-independent factors are extracted, the voice conversion process can be recast as a style transfer problem where speech characteristics are regarded as styles, and speaker-independent factors are regarded as domains (Qian et al., 2019). The key idea behind the style transfer formulation is to achieve full disentanglement between the styles and domains from which one can manipulate and/or replace the different styles, often represented as "embeddings." When replacing the styles, one can incorporate timbral or prosody information from the input itself, which will result in pure reconstruction, or involve the speaker embedding and prosody information from other speakers, which will result in actual conversion, matching the characteristics of the speaker inserted (Walczyna and Piotrowski, 2023).

Non-parallel VC differs from parallel VC in several approaches. Parallel voice conversion is simpler due to the availability of aligned training data, that is, the same utterances spoken by different individuals. This allows for straightforward learning of direct mappings between source and target features, often involving statistical models like GMMs, vector quantisation (VQ), and dynamic time warping (DTW). Additionally, parallel VC primarily focuses on learning a spectrum mapping between seen speakers. The techniques employed in these scenarios do not generalise well to unseen data and do not provide useful information for one-shot or zero-shot conversion, such as speaker embeddings. In contrast, the absence of alignment in training data is a significant challenge in non-parallel VC. Without paired utterances, a VC model trained on non-parallel data does not have explicit examples of how to map features from the source speaker to the target speaker. From a deep learning perspective, parallel training simplifies the task by providing models with the same linguistic content for both source and target speakers, reducing the problem to primarily learning a timbral mapping. In contrast, non-parallel models must concurrently interpret and align linguistic content, timbre, and prosody without the benefit of aligned utterances, thereby increasing the complexity of the task. One approach to addressing this challenge involves representing speaker-independent features explicitly, for example, through phonetic posteriorgrams (PPGs) that establish an intermediary phonetic representation of both the source and target speakers (Sun et al., 2016) or speaker embeddings representing speaker identity as a string of data. As will be subsequently discussed, several alternative methods are available.

Numerous models have been proposed in the literature to extract and map the above-mentioned features using deep learning and neural networks. *Generative adversarial networks* (GANs) and *(variational) auto-encoders* (AEs/VAEs) are particularly popular choices. Lu et al. (2021) and Zhao W. et al. (2019) utilise traditional GAN-based training schemes with timbre representation losses to train a system that can extract and match the timbre characteristics of many speakers. Differently, Dhar et al. (2023) extend the generator of a traditional GAN network with

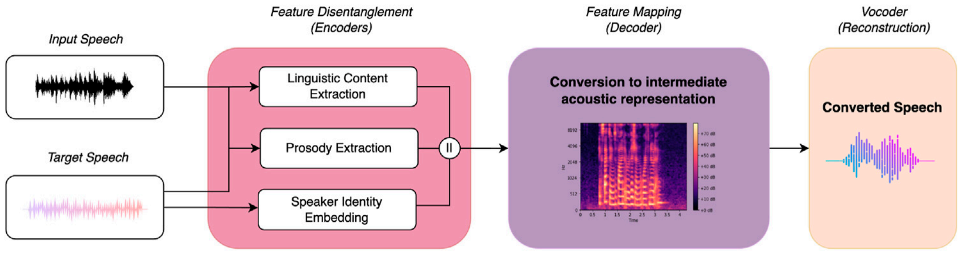
Input Speech
Target Speech
Feature Disentanglement (Encoders)
Linguistic Content Extraction
Prosody Extraction
Speaker Identity Embedding
||
Feature Mapping (Decoder)
Conversion to intermediate acoustic representation
Vocoder (Reconstruction)
Converted Speech

FIGURE 1
Illustration of a traditional deep learning-powered voice conversion pipeline and its three main stages. "Prosody extraction" includes features related to pitch, energy, and rhythm.

adaptive learning using a dense residual network (DRN) to enhance the feature learning ability, that is, speaker generalisation, of the proposed model. Ferro et al. (2021) use adversarial weight-training paradigms to map different features to more realistic representations by creating balance in the sometimes unstable nature of a GAN. This is done by giving more attention to samples that fool the discriminator and allowing the generator to learn more from "true" samples than "fake" ones. The study further imparts an inductive bias by using spectral envelopes as input data for the generator. By doing this, they limit the conversion task to subtle adjustments of the spectral formants, promoting ease of learning in the often-challenging training scheme of GANs. In contrast to the GAN-based approaches, Du et al. (2022a), Nikonorov et al. (2021), and Tang et al. (2022) use the benefits of representation learning in AEs and VAEs. In this process, the models segregate linguistic and timbre details by creating an information bottleneck. Variations in the size and characteristics are utilised to represent the latent spaces, as well as the explicit control provided by these models. Du et al. (2022a) use traditional content and speaker embeddings to condition the decoder, which in turn produces a mel spectrogram to be synthesised by the vocoder. In contrast, Nikonorov et al. (2021) focus on learning a latent representation from which the decoder can create harmonic and noise components matching that of the target speech. Lastly, Tang et al. (2022) encode a broader range of information, including speaker, content, style, and pitch (F0), making it easier to force disentanglement and interchange chosen features in the conversion process.

Despite the improved efficiency of AE and VAE-based representations, Wu et al. (2020) note that it may produce imperfect disentanglement in some cases, harming the quality of the output speech. This happens because weaknesses in any intermediate and individual module will cascade errors in the overall system. To address this, Wu et al. (2020) further extend the auto-encoder-based VC framework with a U-Net architecture and force a strong information bottleneck using VQ on the latent vectors. The latter is done to prevent the U-Net from overfitting on the reconstruction task and will later be shown to be a popular choice in regularising the latent space (see Sections 5.2.2 and 4.2).

As noted in Figure 1, the speech is finally reconstructed by synthesising the intermediate acoustic representation back into the time domain. Although this classically has been achieved by the Griffin–Lim algorithm or inverse Fourier transforms, current work utilises neural vocoders such as the WaveNet (van den Oord et al., 2016) or the HiFi-GAN (Lian et al., 2022). These processes are known for their high fidelity and robustness toward modifications in the intermediate representations – aspects that are crucial for high output quality. The use and inclusion of vocoders will be examined in Section 4.3.

In the context of DL-based VC, it is important to acknowledge the significance of language models (LMs) as the recent surge in LM research has demonstrated promising outcomes for VC, particularly for feature mapping and general robustness. Wang et al. (2023) propose "LM-VC," in which the usual embeddings are substituted by tokens known from language representations. Here, a two-stage masked language model generates coarse acoustic tokens for recovering both the source linguistic content and the target speaker's timbre. The approach is shown to outperform competitive systems for speech naturalness and speaker similarity; however, the model is restricted to the use of well-known tokenisers, which often contain millions of parameters (Hsu et al., 2021). VC systems relying on LMs are, therefore, inherently intricate, lack interpretability, and demonstrate inefficiency during inference. Therefore, they may not necessarily always contribute positively to the VC process. Although the methodology surrounding LMs presents intriguing avenues for VC research, the expansive proliferation of LM studies within the broader realm of AI has rendered it impractical to encompass its entirety within this review. Consequently, we will not directly search for LMs nor extensively cover LMs and transformer models as standalone subjects. We believe this would require an individual review. Instead, we will concentrate on exploring techniques intrinsic to LM research that are employed within the encountered VC pipelines, such as attention and masking. Further elaboration on this focus will be provided in Section 4.2.

In addition to complexity, freedom, and modularity, the introduction of deep learning signifies a departure from the conventional analysis-mapping-reconstruction pipeline. The above-mentioned techniques may all be trained in an end-to-end manner, substituting each sub-task with other neural processes either from similar VC work or from completely different speech processing fields. Subsequent sections of this article will navigate deeper into the intricacies of these techniques. The forthcoming sections will serve as a bridge to the results, offering a granular perspective on the approach taken when choosing and extracting our data.

3 Materials and methods

As earlier mentioned, the decision to undertake a scoping review in the domain of deep learning-powered VC has been informed by the transformative nature of deep learning. Sisman et al. (2020) emphasized that differentiable techniques have shifted the paradigm away from the traditional analysis-mapping-reconstruction pipeline. This shift enables end-to-end training, providing flexibility and improved target matching; however, challenges arise depending on the new methods incorporated. We seek to review current techniques and survey future challenges mentioned in the current literature. Unlike other review types, scoping reviews aim to "identify and map the available evidence" (Munn et al., 2018) and thus focus on the quality and quantity of key features rather than answering specific questions (Grant and Booth, 2009).

Although no formal quality assessment is needed in a scoping review, Colquhoun et al. (2014) recommend following a few simple guidelines to ensure consistency in the analysis and synthesis phases. As suggested by Colquhoun et al., we have chosen to follow Arksey and O'Malley's framework stages for the conduct of scoping reviews combined with the Levac et al. enhancements (Colquhoun et al., 2014). In the review process, we take the following steps: 1) Identify the research question, 2) Identify relevant studies, 3) Select and screen relevant studies, 4) Chart the gathered data, 5) Collate, summarise, and report the results (Colquhoun et al., 2014). We furthermore integrate the PRISMA checklist for scoping review (PRISMA ScR) into the guidelines, ensuring consistency and objectivity throughout the iterative reviewing process (Tricco et al., 2018). The latter aspects have been specifically important as the reviewing process has been carried out by fewer authors than recommended, making the synthesis and results receptive to subjective bias. Part of the initial paper analysis and code extraction process (step 3 and 4) was carried out using the generative AI tool "elicit"[1]

3.1 Research questions

We guide our review of deep learning-based VC by the following research objectives: *1) identify the current state of the art in the field of deep learning-based VC, 2) identify the commonly used tools, techniques, and evaluation methods in deep learning-based VC research, and 3) gain a comprehensive understanding of the requirements and existing gaps in different VC frameworks*. To accomplish these objectives, our review will address the following research questions.

- What are the fundamental components that comprise high-fidelity VC pipelines?
- What are the primary areas of concern addressed in research on VC, and which challenges are they trying to solve?

More specifically, our review will examine research findings and standardised methodologies in the domain of VC. We aim to provide a quantitative analysis of the approaches employed at each stage of the conversion pipeline, clarifying the rationale behind the selection and application of these techniques.

3.2 Keyword identification

Relevant studies were retrieved using research-specific keywords. We identified the keywords using a data-driven approach where one main keyword guided the search for related keywords. We did this to ensure objectivity and overcome limitations regarding knowledge gaps or biases towards terms we would use to describe the research objectives at hand. To find deep learning terms connected to VC, we searched for relevant keywords in the 2022 proceedings of two machine learning and audio-related conferences (ICASSP[2] and NeurIPS[3]) using the main keyword "voice conversion." For all papers retrieved, Author 1 screened the relevance of the results by reading the full title and abstracts, where-after the global keyword list was updated by the author keywords from each paper. In total, 18 relevant papers on voice conversion and deep learning were found. In these papers, 16 unique keywords were repeated more than once. The complete list is shown in Table 1, with each keyword sorted into subtopics. We specifically excluded the keyword "text" to avoid searching for studies focussing on TTS-based VC. Lastly, we added feature-based keywords like pitch, timbre, formant, energy, and dynamics to further force audio domain-specificity.

3.3 General search

We ensured that the voice conversion content was limited to deep learning techniques using AND operators between the first column and the last column of the keyword list, while OR operators were used between the remaining columns and rows of Table 1. This meant that articles searched for all contained the keyword "voice conversion" in addition to popular deep learning methods, subtopics, and features. We queried the Scopus®[4] and the Web Of Science[5] databases due to the former's high scientific journal rankings and the latter's indexing of conferences, such as the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the Conference and Workshop on Neural Information Processing Systems (NeurIPS). We searched for the keywords in all material of the Scopus archive and limited the search to title, abstract, and keywords for the Web of Science archive. Initially, 621 papers were retrieved, 422 from Scopus and 199 from Web of Science. Despite the long history of VC, deep learning-based solutions only started to gain popularity in the mid-2010s. The launch of the Voice Conversion Challenge in 2016 highlights this.

[1] https://elicit.com/welcome

[2] International Conference on Acoustics, Speech, and Signal Processing: https://ieeexplore.ieee.org/xpl/conhome/9745891/proceeding, accessed 05.07.2023.

[3] Advances in Neural Information Processing Systems:

[4] Scopus: https://www.scopus.com/search, accessed 23.10.2023.

[5] Web of Science: https://www.webofscience.com, accessed 23.10.2023.

TABLE 1 *Keyword list for literature search. The search is done by placing an AND between the first and remaining columns and an OR between the rows of each column.*

Keyword 1	Keyword 2 (method)	Keyword 3 (subtopic)	Keyword 4 (feature)
Voice conversion	Deep learning	Style transfer	
	Convolutional NNs	Speech synthesis	Pitch
	Generative adversarial	Disentanglement	Timbre
	Unsupervised	Vocoder	Formant
	Adversarial (training*)	Zero-shot	Energy
	Self-supervised	Conditioning	Dynamics
	Vector quantisation	End-to-end	Prosody
	Autoencoder	Speaker embedding	

Therefore, we carried out an additional filtering process in which the search was limited to 6 years from 2017 to 2023. We also filtered out reviews, surveys, book chapters, letters, and thesis papers. The initial filtering resulted in 573 papers.

We implemented a three-stage screening process to manage the extant literature, comprising 1) the title phase, 2) the abstract phase, and 3) the full-text phase. The initial phase involved the removal of duplicated publications. Subsequently, Author 1 screened the remaining papers based solely on their titles. This phase applied two primary criteria: first, manuscripts must be in English, and second, their titles must pertain to the realm of VC, excluding materials concerning regulation, detection, or anti-spoofing. This stage functioned as a supplementary filtration step, addressing any oversights in the initial filtering stage that might have arisen from inaccuracies in metadata. Phase 1 led to the exclusion of 357 papers primarily due to their lack of relevance. In the subsequent phase, abstracts of the remaining 216 works underwent review and assessment against various exclusion/eligibility criteria (EC) formulated iteratively throughout phase 1.

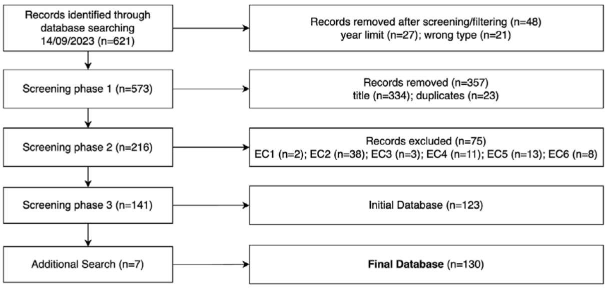
Records identified through database searching 14/09/2023 (n=621)
Records removed after screening/filtering (n=48) year limit (n=27); wrong type (n=21)
Screening phase 1 (n=573)
Records removed (n=357) title (n=334); duplicates (n=23)
Screening phase 2 (n=216)
Records excluded (n=75) EC1 (n=2); EC2 (n=38); EC3 (n=3); EC4 (n=11); EC5 (n=13); EC6 (n=8)
Screening phase 3 (n=141)
Initial Database (n=123)
Additional Search (n=7)
Final Database (n=130)

FIGURE 2
Prisma-ScR chart documenting the retrieval process of identified sources of evidence used for data extraction and analysis.

EC1 *Modality:* The main focus of the article is on other modalities, such as video information or text-to-speech systems. Only direct voice conversion using speaker-to-speaker or reconstruction methods (audio-to-audio) should be included.
EC2 *Purpose:* The article has a bigger purpose than feature-based VC. For example, it aims to achieve speaker recognition and identification, recreate pathological voices, or convert whispers and screams.
EC3 *Synthesis:* The paper deals with speech synthesis/neural vocoding only.
EC4 *Method:* The paper does not include any deep learning techniques (GANs, AEs, VAEs, RNNs, attention mechanisms, etc.).
EC5 *Singing Voice:* The system is focused on singing voice synthesis or conversion.
EC6 *Lack of VC information:* The paper lacks general information on the VC process; for example, it focuses on evaluation methods.

It is imperative to acknowledge that EC2 was incorporated to curate a more streamlined dataset. Given the thematic focus of our review on feature conversion, which encompasses dimensions such as timbre, emotion, and accent conversion, we regard investigations involving these elements as foundational for tackling additional conversion challenges, including pathological conversion. Simultaneously, we incorporated EC5 based on two principal rationales: First, singing voice conversion (SVC) and speech-based VC present distinct goals leading to architectural modifications and challenges during the encoding and analysis stages. Owing to the musical elements inherent in singing voices, SVC necessitates precise control over musical attributes such as melody, harmony, and rhythm. The enhancements within an SVC pipeline, therefore, predominantly occur during the recogniser training and feature extraction phases, capturing extended pitch ranges and dynamic expressions. Second, despite the alterations in vocal characteristics, SVC pipelines adhere to the same decoding and vocoding structures as VC. Consequently, we do not anticipate that SVC will introduce substantial novelty in this context.

We excluded 75 articles in phase 2 of our screening process, resulting in an initial corpus of 141 papers designated for full-text analysis and coding. Subsequently, during the code extraction phase, an additional 18 papers were deemed either irrelevant or unattainable. Following the retrieval of the initial database, it was observed that VC research using diffusion methods was omitted and not indexed using keywords such as "deep learning" or "style transfer." To address this issue, an additional search was conducted, exclusively querying the databases for diffusion-based VC work using the hash "(voice conversion AND diffusion)." This search resulted in seven more papers fitting our exclusion criteria. A

TABLE 2 Codes used for data extraction.

Research objective	Deep learning methods	Evaluation & misc	Training specs
C1 category	*C4* global structure	*C8* objective evaluation	*C12* loss function
C2 main goal	*C5* analysis features	*C9* subjective evaluation	*C13* optimiser used
C3 contributions	*C6* encoders used	*C10* dataset used	*C14* sampling rate
	C7 vocoder type	*C11* manipulation	

final pool of 130 papers was included in the review. We picture the full source selection process using the PRISMA diagram for scoping reviews (PRISMA-ScR) (Tricco et al., 2018) in Figure 2.

3.4 Data items and code book

The 130 papers were carefully read, analysed, and summarised. We charted the papers based on three main topics: 1) *Research objective and contributions*: what was the goal of the authors, and what did they achieve? 2) *Methods and techniques used*: How did the authors achieve their goal, and which deep learning methods and intermediate features were used? 3) *Evaluation and miscellaneous*: How did the authors evaluate their work, what were the results, and did they apply any manipulation techniques (e.g., augmentation, perturbation, regularisation)? Data items related to each coding topic can be seen in Table 2, while the complete code book can be found in the Supplementary Material.

4 Results

The following section presents a combined analysis and discussion of the results of this review. First, we summarise the papers' research directions and distributions of deep learning methods used. Second, we provide an exposition of the papers' relationship with the traditional VC pipeline explained earlier, substantiating our analysis with quantitative data and illustrative materials. Third, the main problem areas addressed in the analysed work are outlined, including a discussion of topics such as interpretability, prevalent conditioning features, and the challenges encountered in integrating explicit control mechanisms. As we aim to reveal, compare, and discuss general methods and tendencies, describing all the included papers in detail is out of the scope of this work. Rather, a full list of papers is provided in the Supplementary Material as a resource for future in-depth analysis. We further refer to the Supplementary Material for the codes and data extraction used to synthesise the results.

4.1 Overview of papers

The final corpus, consisting of 130 unique papers published in 38 different publication venues, with INTERSPEECH (43), ICASSP (22), APSIPA (8), IEEE-ACM Transactions on Audio, Speech and Language Processing (5), and ISCSLP (4) being the most popular platforms. Figure 3 shows that research on VC has exhibited a steady increase since 2017, reaching its peak in 2022 with 39 papers centred on the topics examined. This trend serves as compelling evidence that interest in VC continues to flourish. Most of the work concentrates on systems using low dimensional data, that is, input sampled at 16 kHz or 2205 kHz, which is suitable for speech because the human vocal range does not exceed the Nyquist frequency for these sampling rates. However, this is not sufficient for modern musical applications where such sampling rates are considered low-quality. Consequently, only one study incorporates a high sampling rate, that is, 48 kHz. It is additionally noticed that there is consensus on training the VC systems on the VCTK dataset (Yamagishi et al., 2019) that provides speech data uttered by 110 English speakers with various accents (n = 53). However, datasets such as the VCC 2018 (n = 19) and the CMU-ARCTIC dataset (n = 16) are other popular choices. Finally, most of the VC pipelines are constructed following the unsupervised paradigm (n = 111), whereas 18 works include aspects of self-supervision. One work is considered semi-supervised (Stephenson et al., 2019). As outlined in Section 2, this is consistent with the observation from related work stating that most VC systems incorporate either AEs or GANs, which inherently operate in an unsupervised manner.

4.2 An overview of the structures employed

A fundamental aspect of the VC pipeline concerns the structure used for learning the conversion process. As shown in Figure 4, most structures used in the studies we have analysed are in the realm of AEs, with the CAE (n = 27), conventional AE (n = 19), and VAE (n = 15) being the most frequently used structures. We distinguish between CAEs and conventional AEs in their use of external conditioning features.

Among many works, Hwang et al. (2022), Qian et al. (2020b), and Kim et al. (2022) explicitly condition the decoder on speaker and pitch embeddings to inform appropriate content and style representations. The conditioning can be used to tune the bottleneck as well as constrain the information flow of the speech components, forcing the conditioning feature to be disentangled on the AE input. Given the potential entanglement of speaker style, prosody, and linguistics in the latent space, the conditioning features serve the purpose of supplying additional information to the decoder. This, in turn, guides the encoder to focus on learning only the essential and speaker-independent representation, thereby disentangling the intertwined aspects of the input data. As seen in the elaborate VC pipeline illustration in Figure 5, the conditioning is based on the input during training and substituted with that of the target speaker during conversion or inference. The non-conditioned AEs and VAEs often differ from the CAEs in their end goal. Most

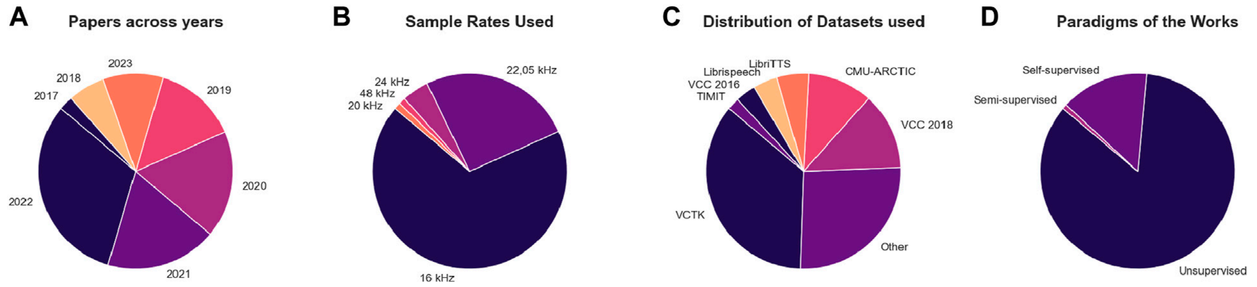

FIGURE 3
Illustrations of the papers across; **(A)** years, **(B)** sample rates used, **(C)** distribution of datasets and **(D)** paradigms included in the works.

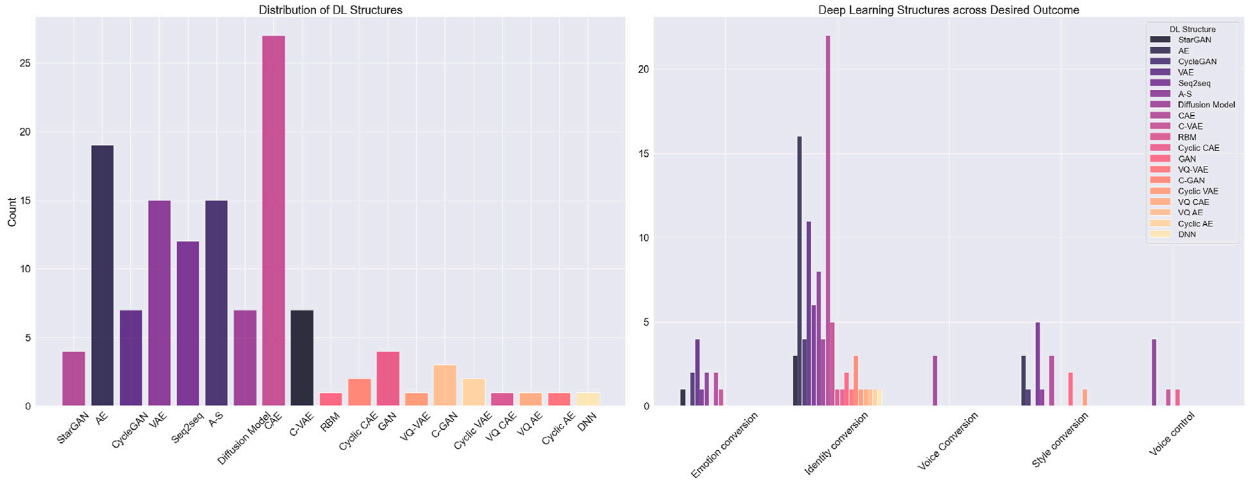

FIGURE 4
Deep learning structures used across the different papers (left) and deep learning structures distributed across their use for specific goals and outcomes (right). "Style conversion" is a combination of both accent and cross-lingual voice conversion used for brevity.

frequently, they consider one-to-one or many-to-one conversion, limiting the need for external information (Cao et al., 2020; Zang et al., 2022).

In the right-side plot of Figure 4, we see that AEs and VAEs are employed for style and emotion conversion, whereas their conditioned counterparts, in particular, are popular for voice and speaker identity conversion. This phenomenon is predominantly attributed to the requirements for external and speaker-dependent conditioning needed in zero-shot and many-to-many identity conversion scenarios. In general, it is clear from the right-hand side of Figure 4 that most of the VC field concentrates on identity and timbre conversion, compared to emotional and style conversion.

CAEs are closely related to the "analysis-synthesis" structure (A-S). A-S structures differ in that they are not inherently obligated to encode the parsed conditioning parameters. In contrast to CAEs, where pitch and speaker information is fed through individual encoders, A-S systems may include encoders but are not limited to doing so – explicit information extracted from DSP processes is often enough. A-S structures often decompose a signal into several attributes counting content, timbre, pitch and energy, and may, as mentioned by Wang D. et al. (2021), Nercessian (2021), Choi et al. (2021), and Xie et al. (2022), concatenate the pitch and/or energy information directly with the content and speaker embeddings before being fed to the decoder. Doing so introduces constraints on the intermediate and mid-level representations. Transparent control mechanisms are distinctive to A-S systems. Given that their parameters are explicitly accessible and employed for training the underlying decoder or generator, A-S architectures exhibit a high level of controllability and frequently yield an interpretable control space. Further elaboration on this topic will be provided in Section 5.2.4.

In addition to the AE-based structures, Figure 4 illustrates a small representation of VC systems based on StarGAN (n = 4) and CycleGAN (n = 7). Both methods come from image-to-image translation, where a cycle-consistency loss enables training without the need for paired data. GAN-based systems directly manipulate the input data to generate data that closely conform to the distribution of the target speaker. As shown in Figure 5, a GAN will often skip the detailed speech representation stage, relying only on content information when performing feature mapping. Compared to AEs, this approach does not necessarily disentangle speaker information from linguistic information. Rather, it relies on

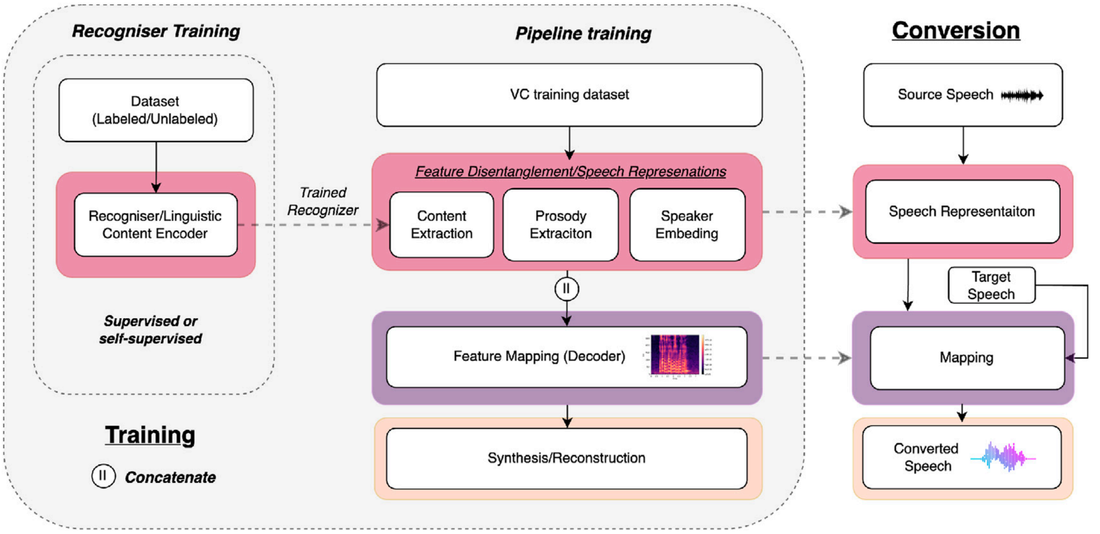

FIGURE 5
Extended illustration of the VC pipeline in its training and conversion stages.

the discriminator's ability to capture the human perception of speaker identity and the generator's ability to create an output that can deceive the discriminator. Given that GANs operate on distributional properties and are not constrained to compel latent representations to be disentangled, they can generate more natural speech. However, while the optimisation of the discriminator and the adversarial losses may yield an output that resembles the distinctive characteristics of the target speaker, it does not always guarantee that the contextual information of the source speaker is kept intact.

To take account of the missing linguistics, Chun et al. (2023), Kaneko et al. (2019a), and Liang et al. (2022) utilise cycle-consistency loss. In cycle-consistency loss, a supplementary generator carries out an inverse mapping of the target y to the input x during training. This procedure encourages the two generators to find (x, y) pairs with the same contextual information, forcing the transformed output to match the linguistics and the timbre of the target. In contrast, StarGAN-based VC systems compress the CycleGAN structure into one generalisable generator-discriminator pair. Using spatially replicated domain codes, often in the style of one-hot vectors, StarGAN conditions the system on speaker information. This allows the model to learn more than one speaker-configuration (Kaneko et al., 2019b; Baas and Kamper, 2020). Although the generators employed in VC architectures based on StarGAN and CycleGAN frequently exhibit AE-like characteristics, their dimensionality reduction and lack of external conditioning make them highly non-interpretable. To address this challenge, one may combine CAEs and GANs into "adversarial auto-encoders." These are built similarly to the traditional CAE structure but are guided by discriminators and adversarial losses and have become popular choices for VC. This is evident as 33 of the 81 AE-based structures analysed in our review (CAE, AE, VAE, and A-S) included adversarial components.

Similar to the use of adversarial components, several studies incorporate VQ as a foundational element in the VC pipeline. The application of VQ in VC can be traced back to the 1980s, when Abe et al. (1988) discretised speech features from parallel data into codebooks and learned a mapping between the codebooks of two speakers to perform the conversion. Currently, VQ is being applied to non-parallel data to disentangle content and speaker embeddings, as the compression capabilities of VQ can discard speaker information from the content code. Tang et al. (2022) use a 512-dimensional learnable codebook to quantise continuous data from a content encoder into a discrete latent space. Regarding a discretised utterance as the related content embedding, they further retrieve the speaker embedding by calculating the mean difference between the discrete code and the continuous encoder output. In other words, the speaker information is considered to be what remains after quantisation. These ideas are elaborated in Wang D. et al. (2021). Speech representation disentanglement is challenging because correlations between speaker and content representations can cause content information to leak into the speaker representation. To address this issue, the authors implement several techniques, including adding VQ to the content encoder. This addition creates an information bottleneck that filters out non-essential details from the representation, thereby aligning the output more closely with underlying linguistic information. Wu et al. (2020) further incorporate the VQ technique into a U-Net-like structure that links each downsampling layer of the content encoder to the corresponding up-sampling layer of the decoder. As highlighted by Wu et al. (2020), U-Net-like structures are seldom used to decode spectrograms in VC pipelines due to their tendency to overfit. However, adding VQ to the skip connections in the U-Net can create a sufficiently strong bottleneck to prevent it from overfitting on the reconstruction task.

Twelve studies follow a sequence-to-sequence (seq2seq) procedure traditionally known from natural language processing (NLP). Although most seq2seq models are based on encoder–decoder structures (Sutskever et al., 2014), we distinguish between these models and traditional AEs as they differ significantly in their processing and training steps.

Following seq2seq modelling, Huang et al. (2020b) encode raw input speech into discretised features, which are represented as indices. A target-dependent seq2seq model then learns a mapping between the source feature sequence and a target sequence. Zhang J.-X. et al. (2020) extract a content embedding using a seq2seq recognition encoder. The encoder is here guided by an embedding derived by a text encoder that is fed with phoneme transcriptions. The seq2seq structure is used as an external module "adopted for obtaining disentangled linguistic representations" (Zhang J.-X. et al., 2020) and is one of many use cases of seq2seq-based models. Such models are often pretrained on large, multi-speaker datasets, and although they create a robust many-to-one pipeline, they are limited by sequential modelling and complex intermediate structures.

Similar to NLP-inspired seq2seq modelling, few studies employ transformer and attention mechanisms (n = 14), often for feature disentanglement or generation tasks. Fu et al. (2022) augment the generator of a CycleGAN pipeline with a transformer in order to capture temporal relationships among the down-sampled time steps. Long et al. (2022) and Shi et al. (2022) insert self-attention into the decoder, guiding it to focus on important regions, taking account of non-local and long-range information. As reported in both works, the attention improves the zero-shot many-to-many pipeline, boosting model performance while achieving faster convergence. Furthermore, the aspect of "masking" is introduced in two of the analysed studies. Gu et al. (2023) add robustness to their model by adding learnable similarity-guided masking (LSGM) to the content encoder, masking redundant input frames when performing inter-frame compression. In contrast, Wang J. et al. (2021) integrate an adversarial mask-and-predict (MAP) network, drawing inspiration from the deep bidirectional transformer model "BERT" to enhance feature disentanglement. Specifically, they employ a random masking and prediction approach, masking one of the four speech representations (content, timbre, rhythm, and pitch) for each speaker during training. By masking one feature space and predicting it from the remaining representations, MAP enables selective modification of specific features and enhances the disentanglement capabilities of the model.

When analysing the widespread application of self-supervised methods in connection with transformers, masking, and attention mechanisms, it becomes evident that of the 18 articles that included self-supervision, the predominant focus lies on leveraging self-supervision for speech representation learning. Consequently, self-supervision serves to strengthen the comprehensive VC pipeline, either facilitating better disentanglement or enhancing general performance. As shown in Figure 5, this often occurs during the recogniser training stage. Yang et al. (2023), Chun et al. (2023), and Dang et al. (2022) use networks pretrained in a self-supervised manner to extract high-level speech representations of different kinds. Yang et al. (2023) use the Mockingjay model, a bidirectional transformer network known to capture content information and speaker information to condition the decoding process (Liu et al., 2019). Chun et al. (2023) and Dang et al. (2022), on the other hand, use outputs of pretrained wav2vec models as content embeddings. In general, it is evident that self-supervised feature extraction processes are attractive due to their potential to replace expensive supervised content representations such as PPGs or non-generalisable speaker embeddings such as the X-vector. This will be discussed further in Section 4.4.

Lastly, seven of the 130 studies reviewed incorporate diffusion probabilistic modelling into the VC pipeline. Diffusion modelling entails progressively introducing noise to a prior distribution and, subsequently, reversing it to generate synthetic data. In these works, this occurs through denoising diffusion probabilistic models (DDPM), where a forward process systematically introduces noise to an input, and a backward process, through a generative model, iteratively reconstructs the input data distribution by predicting the noise. In most of the VC pipelines encountered, the diffusion process is added to the decoder (Baas and Kamper, 2023; Choi H.-Y. et al., 2023; Zhang et al., 2023). Choi H.-Y. et al. (2023) introduce a conditional diffusion model as an extraneous decoder to ensure high-quality speech synthesis. During training, the DDPM process is used to noise and denoise an approximate acoustic representation produced by a source-plus-filter encoder. More specifically, two individual encoders, a source and a filter encoder, reconstruct an intermediate mel spectrogram from disentangled speech representations. Using diffusion, the intermediate representation is thereafter transformed into a high-quality equivalent that can be fed to a neural vocoder. The inserted diffusion stage uses the source-filter encoder output as a data-driven prior and uses pitch and speaker representations as conditions to maximise overall speaker adaptation capacity. A comparable approach is adopted by Popov et al. (2022). In this study, a transformer-based content encoder is trained on average phoneme-level mel features to generate an "average voice" speech representation. Subsequently, this speech representation is combined with the output of a speaker conditioning network and a noise variable t sampled from a uniform distribution before being inputted into a diffusion-based U-Net decoder. Analogous to this research, the diffusion process yields a high-quality acoustic representation primed for a neural vocoder. Moreover, the authors introduce a novel stochastic differential equations solver tailored for rapid sampling, thereby facilitating fast synthesis. As in Popov et al. (2022), the U-Net architecture is commonly adopted for processing inputs represented in the time-frequency domain. This is attributed to its demonstrated success across a range of image processing and object detection applications, allowing one to treat speech as images and objects rather than one-dimensional sequences. In addition to the research conducted by Popov et al. (2022), the U-Net architecture is employed by Wu et al. (2020) and subsequently extended by Liu et al. (2021) for natural one-shot conversion. More specifically, the U2-Net in Liu et al. (2021) is a two-level nested U-Net structure comprised of residual u-blocks (RSU) inserted into a classic U-Net pipeline. The RSU blocks operate at different scales, allowing the network to extract local and multi-scale features from the input, thereby enhancing the naturalness of the converted speech. In contrast to related work, the U2-Net conditions the decoder on target speaker information through its skip connections, meaning that the encoder takes the source and target spectrograms as input.

4.3 The use of vocoders

Vocoders are crucial to the voice conversion process, as they enable the generation of audio based on the intermediate representation. Like the difference in deep learning structures

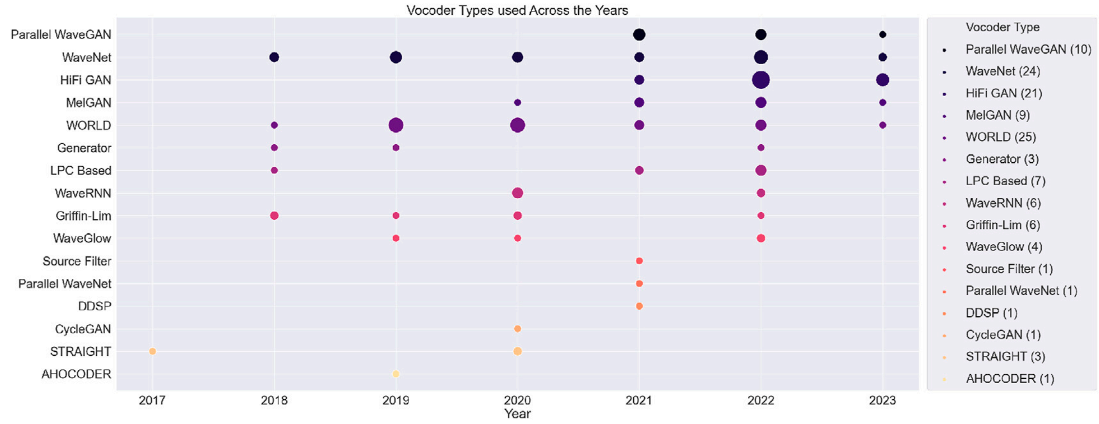

FIGURE 6
Distribution of vocoders used from 2017–2023. The size of the shape represents the number of similar vocoders used in that year.

employed, the choice of the vocoder differs depending on the use case. In general, we can divide vocoders into three main classes: 1) concatenate, signal-based models, such as the harmonic plus noise model (Stylianou, 2001), 2) hand-designed vocoders, or source-filter models, such as the STRAIGHT (Kawahara et al., 1999) and WORLD (Morise et al., 2016) models, and 3) neural vocoders such as the WaveNet (van den Oord et al., 2016) or the WaveRNN (Kalchbrenner et al., 2018) models.

Neural vocoders have grown increasingly popular as their data-driven approach and high-quality output allow for very expressive synthesis. They simultaneously only require an intermediate acoustic representation as input, often in the form of a mel spectrogram, and are, therefore, highly flexible, allowing them to be inserted in almost any end-to-end pipeline. The statistics of the analysed work in this review support these arguments. Figure 6 shows that 75 of the 130 papers used a neural vocoder as its synthesis back-end, with the WaveNet (n = 24) vocoder being the most popular.

In general, neural vocoders have become state of the art in terms of audio quality. WaveNet (van den Oord et al., 2016) paved the way in 2016 with its auto-regressive nature and is still widely used today. Yang S. et al. (2022) and Bonnici et al. (2022) used it as the main synthesis block in an AE-based pipeline, whereas Zhang J.-X. et al. (2020) used it in a seq2seq modelling pipeline. Like most neural vocoders, the WaveNet is conditioned via acoustic features; however, Tan et al. (2021) and Wu et al. (2021) extended the WaveNet with fundamental frequency (F0) conditioning to force decoupling between the pitch and content. In Tan et al. (2021), this is done by substituting the predicted mel spectrogram with simple acoustic features (SAF), such as the mel-cepstral coefficients (MCCs) and log-F0 information. Differently, Wu et al. (2021) extend the WaveNet implementation itself through pitch-dependent dilated convolution neural networks (PDCNN) and auxiliary F0 conditioning.

Even today, WaveNet's sequential generation remains prohibitively costly, driving the need for more efficient neural vocoders. GAN-based vocoders emerge from this necessity, aiming to enable non-autoregressive generation architectures capable of synthesising high-quality speech. As a result, GAN-based vocoders have set a new benchmark, characterised by their fast inference speed and lightweight networks (Sisman et al., 2020). In the analysed studies, the most widely used GAN-based vocoders are the HiFi-GAN (n = 21) and the parallel WaveGAN (n = 10). Figure 6 shows that their use has become more frequent after the invention of the MelGAN in 2019. In addition to the inclusion of adversarial training, the inputs to and usage of GAN-based vocoders do not differ significantly from other neural vocoders such as the WaveNet, and the papers examined rarely justify the type of vocoder chosen. However, it is often mentioned that GAN-based vocoders are included due to their "better speech quality and much faster inference speed" (Lian et al., 2022).

Even though neural vocoders have become increasingly popular, parametric and hand-designed vocoders are still used. Figure 6 shows that the WORLD vocoder was used more than its neural counterparts in 2019–2020 (n = 25). The WORLD vocoder is a high-quality speech synthesis and analysis tool used to extract and synthesise waveform information. In the work analysed, it is mainly used for two reasons: first, its inherent capability to extract pitch and timbre information provides a strong foundation for subsequent disentanglement efforts; second, the substantial amount of acoustic data it offers facilitates a straightforward guidance of a WORLD synthesis process. Huang et al. (2020a) and Kaneko et al. (2019b) use the WORLD vocoder to extract aperiodicity signals (APs), F0 features, and 513-dimensional spectral envelopes. The spectral envelopes are further reduced to more specific MCCs and encoded for linguistic/content information. The F0 features are linearly transformed to match the target and can, together with the extracted AP information, be carried over to the inverse WORLD synthesis stage directly. This simplifies the conversion task to be a non-linear transformation of the remaining source spectrum only (often conditioned on extra target speaker information), creating an inductive bias. Almost the same procedure can be carried out for the STRAIGHT vocoder; however, we only see the use of this vocoder in three of the analysed works. Although these parametric synthesisers offer robustness and flexibility, they are limited to monophonic reproduction. They are simultaneously limited by their internal synthesis mechanism, which often produces artefacts (Nguyen and Cardinaux, 2022).

The signal-based vocoders used in two of the studied articles are closely related to the hand-designed vocoders. An example is the continuous sinusoidal model utilised in the synthesis stage by Al-Radhi et al. (2021). Here, a neural network converts sinusoidal parameters, constructing speech frames from a voiced and a noise component, respectively. Because the synthesis stage is vocoder-free, this approach simplifies the learning process and limits the model to learn the reconstruction of the intermediate representation only. Similarly, a harmonic plus noise model is used by Nercessian (2021). Inspired by the differentiable digital signal processing approach (DDSP) (Engel et al., 2020), a feature transformation network learns to map input attributes to parameters that control a differentiable harmonic plus noise (H + N) synthesiser. More specifically, the network predicts the harmonic distribution for an additive sinusoidal synthesiser and 65 noise filter taps used to filter the noise part of the produced speech signal. Despite being efficient and lightweight, the inclusion of the DDSP framework additionally introduces an inductive bias as the sinusoids may be directly controlled by the input pitch. In contrast, the output quality is limited by the capabilities of H + N synthesis, forcing post-filtering or extra processing.

Although most of the work examined in this review uses actual vocoders and thus adheres to an encoder-decoder-vocoder structure, a few works are taking a more immediate approach, using the generator to produce time-domain data directly (n = 3). With the reasoning that traditional VC highly depends on the quality of the intermediate representation and the vocoder itself, the NVC-Net in Nguyen and Cardinaux (2022), for example, performs “voice conversion directly on the raw audio waveform.” This is done by combining the decoder and the vocoder into a single generator inspired by the MelGAN (Baevski et al., 2020). More specifically, the generator upsamples the latent embedding using transposed residual convolutions, with each residual connection being conditioned by a target speaker identity. By limiting the NVC-Net to exploit only its internal representation, the authors provide a high-quality, condensed and fast framework claimed to “generate samples at a rate of 3661.65 kHz on an NVIDIA V100 GPU in full precision and 7.49 kHz on a single CPU core” (Nguyen and Cardinaux, 2022). These methods are highly useful in low-latency scenarios but are also favourable for zero-shot VC, as speech production is independent of intermediate representations and mismatch problems.

4.4 Choices of feature extraction

The selection of encoder structures for feature extraction and feature embeddings constitutes another significant aspect of the VC pipeline. As previously noted, it is imperative to segregate content, prosody, and speaker identity to disentangle linguistic attributes from other characteristics.

4.4.1 Content embeddings

Figure 7 shows a consensus on the networks used for retrieving linguistic content information, with traditional CNNs and related structures being the most popular. In the analysed work, singular CNNs are used to extract the content embedding in 28 of the cases, whereas the combination of CNNs and BiLSTMs are incorporated nine times. The combination of CNNs and recurrent neural networks (RNNs), such as the BiLSTM, is interesting as it is inspired by the field of automatic speech recognition (ASR). In ASR systems, CNNs are employed as they are beneficial in modelling local acoustic patterns, either in the audio signal or the spectrogram, while the RNN is advantageous in capturing temporal dependencies. Overall, the analysed work represents ASR-based content information as 1) Linguistic embeddings using condensed ASR blocks such as the CNN + BiLSTM or CNN + LSTM encoders trained end-to-end by Tan et al. (2021), Choi and Hahn (2021), Wang Q. et al. (2022), or 2) As PPGs obtained from speaker-independent ASR systems as done by Mohammadi and Kim (2019) and Chen et al. (2022). The latter is efficient as it uses pretrained models for the extraction task, often trained on large, multi-lingual datasets such as the “Kaldi speech recognition toolkit” (Povey et al., 2011), the “Julius dictation kit,”[6] or through the conformer model (Gulati et al., 2020). Obtaining linguistic content through speech recognition models is useful as it decodes linguistic discriminant information from speech without considering who is speaking, thus creating a complete speaker-independent representation. Overall, the ASR-based approach “frees up the conversion network from using its capacity to represent low-level detail and general information” (Sisman et al., 2020). Instead, the network can focus on the high-level semantics necessary for the conversion.

Although PPGs encapsulate linguistic content, they may still contain errors stemming from the recognition model, which can result in mispronunciations in the converted speech output. Furthermore, they are labour-intensive, as PPG and related ASR-based content encoders necessitate a substantial volume of labelled data for effective training. As depicted in Figure 7A, minor representations of wav2vec models are thus used for extracting the linguistic content information (n = 5). Wav2vec models, available in unsupervised (Schneider et al., 2019) and self-supervised forms, are frequently used to generate speech representations within ASR pipelines (Baevski et al., 2020). Like the aforementioned PPG-based ASR systems, wav2vec models provide linguistic embeddings that are time-aligned and speaker-independent. Their high-level analysis features make them superior on downstream tasks, especially for low-resource languages (Choi et al., 2021). Chun et al. (2023) and Choi et al. (2021) derive the content embedding from the 12th layer of a pretrained XLSR-53 wav2vec model. This particular representation has been chosen due to its reported significance in encapsulating essential pronunciation-related characteristics (Singla et al., 2022) and outperforms related work when inserted into an unsupervised VC pipeline (van Niekerk et al., 2022). Finding alternative self-supervised ways to represent content information is thus of high interest in the VC and speech representation community. In addition to employing wav2vec features, the analysed work uses methods such as the unsupervised contrastive predictive coding (CPC) algorithm (n = 4) and the self-supervised HuBERT model (n = 2). CPC and HuBERT are used to directly represent context and linguistic-

[6] https://github.com/julius-speech/dictation-kit

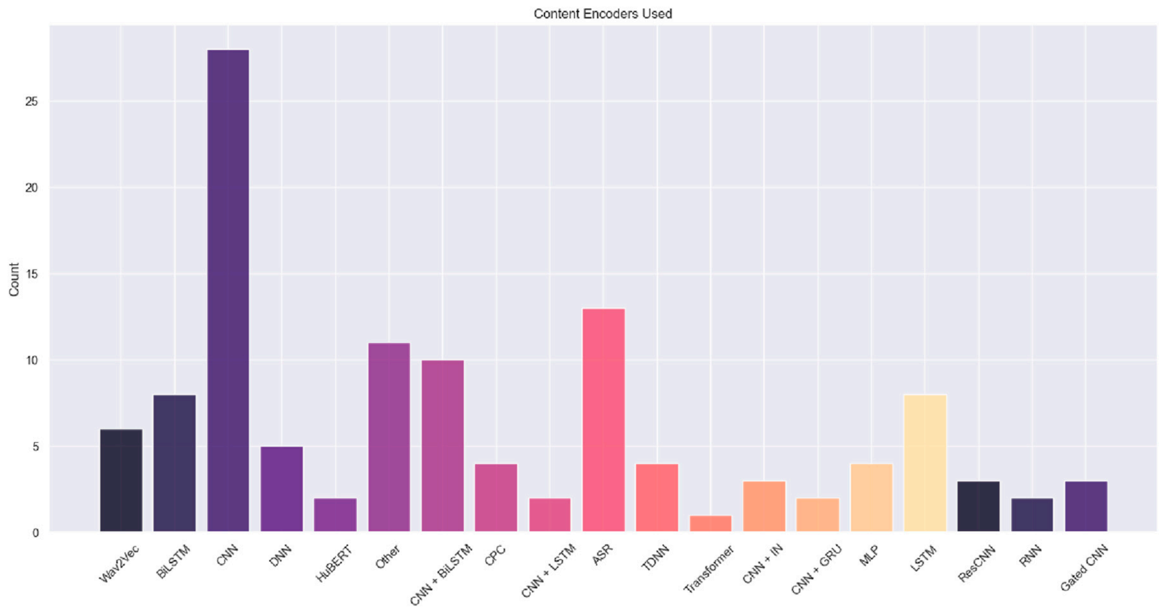

FIGURE 7
Distribution of content encoder networks used across all the analysed studies. For brevity, we have collected encoders appearing once in the "other" category. This category, among others, contains pyramid attention modules, feed-forward transformers, and dynamic time-warping models.

related features from the raw waveform. CPC is employed to further encourage a latent content vector to capture local structures (Yang H. et al., 2022; Huang et al., 2022), whereas the HuBERT model is used to form discrete speech units that represent linguistic content in the latent space (Hsu et al., 2021). More precisely, discrete speech units refer to transformer-produced latent variables that have been influenced by an acoustic unit discovery system, such as a system that clusters MFCCs with phonetic similarities using k-means clustering (Li et al., 2023). Compared to the different hidden layers of the wav2vec models, encoders based on self-supervised representations, such as those from a HuBERT or CPC, are reported to contain large amounts of speaker information, which in some cases may render them unsuitable for VC (Li et al., 2023). van Niekerk et al. (2022) employ a soft content encoder. A soft content encoder incorporates a linear projection head on top of the HuBERT model that, in a supervised manner, is trained to predict a distribution over the discrete units. This offers an intermediary solution between raw continuous features and rigid discrete units without containing speaker-related information. Compared to related ASR models, it simultaneously provides efficient supervised labelling.

van Niekerk et al. (2022) demonstrate that an unsupervised VC pipeline can be enhanced from discrete and soft content encoders, with the latter exhibiting superiority in objective metrics and subjective evaluations. Finally, Gu et al. (2023) integrate the masking procedure initially introduced in HuBERT directly into an autoencoder architecture. As mentioned earlier, they introduce LSGM, compelling the encoder to discern and infer masked frames based on neighbouring similar points. This method serves multiple objectives: first, it achieves feature compression through masking rather than dimensionality reduction, thereby enhancing robustness; second, it introduces self-supervision directly into the end-to-end training process, reducing reliance on large pretrained models.

4.4.2 Speaker embeddings

In contrast to the content encoders, the techniques used to retrieve speaker embeddings and timbre characteristics differ significantly across the field. Few works obtain speaker embeddings by averaging the frame-level characteristics of different speaker utterances or by downsampling the input to one-hot vectors (Kaneko et al., 2019b), while others incorporate feature vectors and codebooks directly (Reddy and Rao, 2020; Ho and Akagi, 2021). Few works use the ECAPA-TDNN architecture trained for speaker verification (Zhang et al., 2021) or the earlier mentioned XLSR-53 model, whose first layer forms clusters for each speaker (Choi et al., 2021). Du and Yao (2023) further extend the ECAPA-TDNN by pretrained X-vector networks to address the differences in their distributional variations. More simple and traditional approaches to speaker embeddings are also taken in the analysed work. Dang et al. (2022) use a 12-layer CNN to encode the input mel spectrogram, while Chen et al. (2022) use a BiLSTM-based speaker encoder pretrained for speaker classification. It is essential to acknowledge that many of the less intricate algorithms are frequently supported by adversarial, classification, or cycle-consistency losses, ensuring that the converted output matches the target characteristics. We will discuss this in Section 4.5.

4.4.3 Additional embeddings

In addition to the content and speaker information, the works analysed include the following features for conditioning and further embedding: pitch (n = 52), rhythm (n = 8), and energy (n = 6), which all relate to prosody. As mentioned earlier, pitch extraction is commonly applied to the VC pipeline. Even though speaker characteristics may be disentangled from the input content in many cases, a significant amount of prosodic information, such as volume or source F0, is often still entangled in the content embedding. This may leak into the intermediate representation, causing mismatch problems or making the converted F0 fluctuate. In this context, the inclusion of supplementary pitch-related data can prove advantageous, serving a dual purpose: first, it enables the content encoder to concentrate exclusively on linguistic aspects; second, as demonstrated when integrating the WORLD vocoder, it can introduce an inductive bias, which can be directly incorporated into the synthesis stage. Like pitch changes in VT, individual prosodic features such as pitch may thus be modified directly during the conversion. This is done by Xie et al. (2022), where the pitch is manipulated after adding extra information to the latent space. Nercessian (2021) transforms the source pitch to the register of the target before using it to modify a source-filter-based vocoder. In contrast, Zhou et al. (2021) and Nguyen et al. (2022) use pitch and energy information as an important factor for prosody transformations in emotion and accent conversion, respectively.

In general, the act of emotional voice and accent conversion (EVC and AC) adds yet another sub-process to the VC pipeline, often in the shape of another encoder. Like the content and speaker disentanglement in identity/timbre conversion, EVC and AC aim at disentangling prosody information from the remaining content, transforming it to match a target. Cao et al. (2020) extract emotion prosody in a supervised manner and further train it using cycle-consistency losses. Du et al. (2022b) disentangle emotional style by employing an emotional style encoder directly on the input mel spectrogram. In general, EVC and AC promote challenges that are different from timbre and identity conversion. Emotion is inherently supra-segmental and hierarchical in nature, making it highly complex, with multiple attributes entangled within the spectrum and prosody (Zhou et al., 2020). It is often not possible to operate at the frame level or from the spectrum alone. To take account of this, Zhou et al. (2020) decompose F0 information into different temporal scales using a continuous wavelet transform (CWT) prior to feeding F0 to the prosody encoder. The accent is, on the other hand, often converted by blending spectral components from two different speakers. This is done using PPGs, bottlenecks, and acoustic models from the field of ASR (Zhao G. et al., 2019; Wang et al., 2020).

In Table 3, we provide an overview of the VC pipelines that include four or more different features and note the techniques used to extract the features outside of the traditional content and speaker conventions. We additionally summarise the overall goal of doing so. As seen in Table 3, the studies listed take a more explicit approach to disentangled speech representation learning, including pitch and prosody information in the process. It is evident that rhythm embedding is often carried out by encoder structures similar to the ones utilised for content representations in ASR. Pitch, however, may be retrieved in terms of pitch contour using classical signal processing methods like the YIN (Cheveigné and Kawahara, 2002), RAPT (Talkin and Kleijn, 1995), or CREPE (Kim et al., 2018) algorithms. In all cases, the pitch contour is further processed by a pitch encoder. Xie et al. (2022) allow the pitch embedding to be influenced by information from the target speaker, whereas Wang J. et al. (2021) and Nercessian (2020) create a more condensed representation. When pitch encoders are used to compress the analysed pitch contour, they often use similar structures as the other encoder models used in the given pipeline, including residual CNN networks (Xie et al., 2022) and combinations of CNNs, group normalisation and BiLSTMs (Wang J. et al., 2021). However, one work uses diffusion-based modelling to further process the input pitch contour. Choi H.-Y. et al. (2023) introduce a diffusion model to effectively transform the normalised F0 of the source speech to the target pitch representation. Like other diffusion-based work, a speaker conditioned pitch encoder creates a pitch-based prior of the normalised input-F0 matching the register of the target. This is then refined by the diffusion model and fed to the decoder. Such a process introduces a highly precise pitch transformation, supporting zero-shot conversion and any mismatch problems.

Looking at Table 3, we lastly see that the encoding of energy information is extracted from signal processing techniques, often directly on the input waveform. Nonetheless, energy is most commonly conveyed as an additional, unaltered conditioning feature, owing to its classification as less critical data.

4.5 Losses, costs, and errors

As mentioned, the success of VC primarily depends on the deep learning structures employed like the information bottleneck principle of the AE system or the generative capabilities of the GAN. However, refining their results and supporting intermediate tasks can be achieved through loss functions tailored to the tasks at hand. Here, an obvious choice is to train the system using a "reconstruction loss," which in the analysed studies was mentioned 77 times. Most of the work aims at minimising the reconstruction loss in the acoustic feature space, that is, the difference in spectral envelopes (He et al., 2021), mel-cepstral coefficients (Kaneko and Kameoka, 2018), or the difference between the input mel spectrogram X_1 and the predicted mel spectrogram $\hat{X}_{1\to1}$, before vocoding. The reconstruction loss can be generalised by the Equation 1:

$$L_{recon} = \mathbb{E}\left[\|\hat{X}_{1\to1} - X_1\|\right], \tag{1}$$

where the difference between the prediction and the ground truth may be calculated using the L1 or the L2 distance, the mean squared error (MSE), or the mean absolute error (MAE). Rather than minimising the error between the acoustic features, a few works calculate the reconstruction loss in the time domain (Du and Yao, 2023). Some studies additionally extract a perceptually based spectral loss from the produced waveforms (Choi et al., 2021). The spectral loss compares the input spectrogram with the spectrogram of the time-domain output and helps the generator obtain the time-frequency characteristics of the produced speech. The predicted and ground-truth spectrograms may be compared individually (Nguyen and Cardinaux, 2022) or at multiple resolutions, as done by Nercessian (2021). In both the case of the time domain and the perceptual reconstruction losses, the vocoder is incorporated into the learning process, giving more degrees of freedom.

As an alternative approach to the reconstruction loss, the VC system may be trained using a feature-matching loss (FM loss). The FM loss is an adversarial method that incorporates a discriminator

TABLE 3 *Overview of analysed VC pipelines that focus on disentangled speech representation learning using four or more explicit features. "Main goals" describe the desired outcome of the work, "pitch and feature representation" describes the methods used to represent the given feature, and "main contribution" describes the primary technique used to achieve the end goal.*

Reference	End goal	Pitch representation	Feature representation	Main contribution
Yang et al. (2022b)	One-shot VC	RAPT → Z-Norm → Enc	Rhythm: Enc (BiLSTM)	Mutual information learning
Wang et al. (2021b)	Any-to-many VC	Contour → RS → Enc	Rhythm: Enc (CNN, BiLSTM)	Adversarial MAP
Luo et al. (2023)	Emotional VC	Contour → RS → Enc	*Rhythm*: Enc (CNN, BiLSTM)	Source-filter-based
Wang et al. (2022c)	Voice control	Swipe + CREPE → Abs-Norm	*Energy*: Time domain	Adv. Training and AIC Loss
Choi et al. (2021)	Voice control	Yingram (Yin spectrogram)	*Energy*: Avg. Log-Mel spec	Information perturbation
Wang et al. (2022b)	Any-to-any VC	Enc (RankNet)	*Energy*: Enc (RankNet)	Self-supervision
Nercessian (2020)	Any-to-any VC	CREPE → Log	*Energy*: A-weighted spec	Explicit conditioning
Xie et al. (2022)	Any-to-any VC	Contour → Enc	*Energy*: Time domain	Information perturbation

to "guide" the reconstruction. The global network is then trained in conjunction with the discriminator, while the generator is updated based on the similarity between the prediction and the ground-truth feature maps produced by the discriminator:

$$L_{FM} = \mathbb{E}_{(c,z,x)}\left[\sum_{i=1}^{L}\frac{1}{N_i}\|D_i(x) - D_i(G(c,z)\|_1\right]. \tag{2}$$

In Equation 2, D_i denotes the feature map output of N_i units from the discriminator at the ith layer (Nguyen and Cardinaux, 2022). Eight studies incorporate FM loss, which is used for GAN-based pipelines (Lu et al., 2021) and for adversarially tuned CAEs and C-VAEs (Huang et al., 2022; Xie et al., 2022). In addition to FM loss, the reconstruction loss may be extended by a content loss. The content loss is particularly useful for constraining the information capacity of the bottleneck, as such systems often are invariant to self-to-self reconstruction. Content loss is applied in the works of Wang Y. et al. (2022), Li and Wei (2022) and Shi et al. (2022), among others, and is given by the following Equation 3 (Qian et al., 2019):

$$L_{content} = \mathbb{E}\left[\|E_c\left(\hat{X}_{1\to1}\right) - C_1\|\right], \tag{3}$$

where $E_c(\hat{X}_{1\to1})$ is the content embedding of the prediction, and C_1 the actual content embedding of the input, for example, $E_c(X_1)$. Incorporating feature-specific loss functions like content loss is a commonly employed approach. This is evident as 10 papers use feature-specific loss functions in addition to the ones based on content. In Yang S. et al. (2022), a pitch-contour loss is used to compare the pitch of the analysed input with the pitch of the reconstruction to enforce pitch coherence. Like the content loss, Lee et al. (2021) compare the style embedding of the input with the prediction fed through the style encoder. Lastly, studies optimise speaker embeddings either by the inclusion of classification losses (n = 9), cross-entropy losses (n = 9), or speaker-ID losses (n = 4). Ho and Akagi (2021) and Ding et al. (2022) train an auxiliary classification model to help the generator "produce fake data with the correct target speaker voice." The classifier is trained using the output log-likelihood that the acoustic features coming from the generated audio belong to the target speaker. Chen et al. (2022) and Dang et al. (2022), on the other hand, improve speaker similarity by optimising the speaker encoder with a speaker classification loss based on the ground-truth speaker identity label in one-hot vector format. Based on this, we define the general speaker identity loss in Equation 4:

$$L_{speaker} = CE\left(x_{id}, softmax\left(V * E_s(x)\right)\right), \tag{4}$$

where CE is the cross-entropy loss, x_{id} is the ground-truth speaker identity vector, V is a trainable weight matrix, and E_s is the speaker encoder (Chen et al., 2022). The speaker ID additionally plays a role in what is called the "cycle-consistency loss" discussed next.

Adversarial and cycle-consistency loss functions are incorporated in 36 and 21 of the studies, respectively. As earlier mentioned, the use of adversarial losses is not limited to GANs. Huang et al. (2021) use a very traditional adversarial loss guided by a patchGAN discriminator in an AE pipeline. Xie et al. (2022) incorporate an adversarial loss based on the parallel WaveGAN model in conjunction with a multi-period discriminator (MPD), multi-scale discriminator (MSD), and multi-resolution spectrogram discriminator (MRSD) to steer the A-S process. Here, the model parameters are optimised based on the generator's ability to deceive all discriminators. Lastly, Hwang et al. (2022) extend the discriminative process by a pitch-based discriminator that, in addition to the real/fake probability prediction, predicts how much the pitch of the reconstruction is similar to that of the target speaker. Most of the cycle-consistency losses included take a form based on the StarGAN paradigm; rather than incorporating another generator, they use the main generator to map the prediction back to its original form by including a speaker label (often referred to as a domain classifier). This is done by Zhang Z. et al. (2020) and Huang et al. (2021), among others, following Equation 5:

$$L_{cyc} = \mathbb{E}_{c\sim(c),x\sim\left(x|c'\right),c\sim(c)}\left[\|G\left(G(x,c),c'\right) - x\|\right], \tag{5}$$

with c representing the attribute label that classifies the domain and G being the generator. Cycle-consistency loss aims to enhance the contextual robustness of the encoder. When the adversarial loss forces output to follow the target-data distribution, the cycle-consistency loss is used to preserve the composition in the conversion.

Although cycle consistency provides a constraint and encourages the forward and inverse mappings to find (x, y) pairs with the same contextual information, it does not guarantee that the mappings always preserve linguistic information. In order to do so, the identity-mapping loss is included. We find identity-mapping losses in 13 studies. Identity-mapping losses are usual in cycleGAN-based pipelines and are equivalent to the content loss used to preserve linguistic information without relying on extra modules (Kaneko and Kameoka, 2018). In general, the identity-mapping loss is adopted to regularise the generator to be close to an identity mapping when one converts the input to that of the same speaker. As mentioned by Cao et al. (2020), the intuition behind this is that "the model is supposed to preserve the input if it already looks like that from the target domain." We can represent the identity-mapping loss by the Equation 6:

$$L_{id} = \mathbb{E}_{y\sim p_y(y)}\left[\|G_{X\rightarrow Y}(y) - y\|_1\right] + \mathbb{E}_{x\sim p_x(x)}\left[\|G_{Y\rightarrow X}(x) - x\|_1\right]. \quad (6)$$

In this context, let $G_{X\rightarrow Y}$ represent the generator responsible for the transformation from the source domain to the target domain, and $G_{Y\rightarrow X}$ denote an auxiliary generator tasked with performing the reverse transformation. The role of the auxiliary generator $G_{Y\rightarrow X}$ is to facilitate the preservation of composition between the input and output domains. This encourages the primary generator, $G_{X\rightarrow Y}$, to discover the mapping that effectively maintains the compositional integrity throughout the transformation process.

4.9 Metrics and performance evaluation

With the amount of VC work available, effective evaluation of the different results is required. This is needed to validate the voice quality of the system proposed and to compare and benchmark results against state-of-the-art work and related techniques. Most commonly, the works analysed utilise the mel-cepstral distortion (MCD) metric (n = 48) to objectively evaluate the overall audio quality of the system output, compared to a reference speech sample. MCD is a measure of the difference between two sequences of mel cepstra, and although it is not always correlated with human opinions, it evaluates perceptually relevant features like the MCEPs of the two signals. Figure 8 shows that different versions of the root mean squared error (RMSE) are used to examine specific attributes of the system outputs. RMSE-F0, RMSE-energy, and general RMSE metrics are respectively used in 12, four, and three of the studies, highlighting the fact that the performance in reconstructing prosodic information is of high importance in many VC systems.

Second, we see a high use of metrics borrowed from the speech recognition or machine translation field. The word error rate (WER) metric, which measures the percentage of words that are not correctly transformed during conversion, is included in 15 of the works analysed. Similar tendencies are experienced for the character error rate (CER) that is included in 14 papers. CER measures the percentage of characters from the output that are transcribed incorrectly. For measuring the WER and the CER, it should be noted that the output and the reference samples are fed through an external speech-to-text (STT) engine to evaluate the intelligibility of the produced speech. However, some databases, such as the VCTK dataset, may include transcriptions of the audio itself. Finally, speaker-based metrics are employed to assess the system's capability in either transforming or retaining speaker identity. Notably, the most commonly used objective metrics in this regard are speaker classification (n = 4) and equal error rate (EER) originating from the domain of speaker verification (n = 6).

Even though the above objective metrics are good at evaluating important aspects related to the quality of the conversion, they have difficulties measuring the naturalness of the reconstruction and are unable to judge its perceptual similarity to a target sample. To measure this, the VC community opt for subjective metrics such as the mean opinion score (MOS), multiple stimuli with hidden reference and anchor (MUSHRA) scales, or AB preference tests. The MOS is a listening test that asks participants to rate the conversion quality from 1 to 5, often based on the aspect of naturalness. MOS is a standard benchmark used in 112 of the 130 analysed studies (see Figure 8), emphasising its value in comparing perceptual aspects of the produced speech to related work. Evaluation methods similar to the MOS include the degradation mean opinion score (DMOS) or the comparative mean opinion score (CMOS); these methods are as popular as the MOS (n = 2). Both the DMOS and the CMOS rate the reconstruction in relation to a reference sample; the former asks the participants to rate the degradation of the reconstruction, whereas the latter asks to rate the identicality. Another commonly employed subjective metric is the AB/ABX test used in 23 of the papers studied. In AB/ABX tests, participants are exposed to the reconstructed audio and the reference audio; afterwards, they must specify which one exhibits a greater degree of a specific attribute (Sisman et al., 2020).

Although subjective evaluation metrics offer a valid perceptual measure of the inherent naturalness of VC results, it is essential to acknowledge their requirement for a substantial number of participants to ensure validity, which can make the process time-consuming. Nevertheless, they have become a standard for evaluating VC results and were employed in nearly all studies analysed.

5 Discussion

In the following section, we summarise our results and, based on our codebook, highlight five specific challenging areas that were discovered in the analysed studies. The challenges will be thoroughly examined, focussing on the methodologies employed to address them in the VC field. Based on these, we explore the potential opportunities in overcoming the challenges and provide recommendations for future research.

5.1 Summary of results

Even though the global voice conversion procedure was similar for most of the studies analysed, the works differed in terms of the underlying disentanglement techniques, vocoders, and loss function used. In general, the work studied in this review can be divided into ***AE-based*** pipelines (n = 77), ***GAN-based*** pipelines (n = 17), ***analysis-synthesis-based*** pipelines (n = 15), and diffusion-based pipelines (n = 7). In general, the AE and most GAN-based systems are similar in their structure using an information bottleneck;

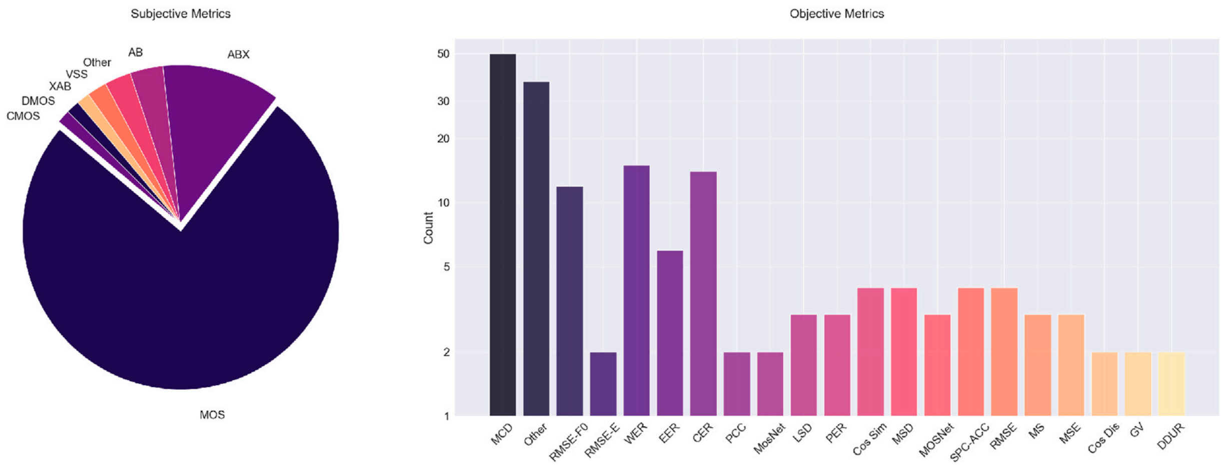

FIGURE 8
Distribution of subjective metrics used (left) and objective metrics employed (right) across the studied work. We see that the MCD is a commonly used objective metric, while the word error rate and character error rate are also popular choices.

however, the GAN-based works combine the analysis and synthesis process into a single generator, whereas the AE pipelines separate the encoder and decoder in order to more easily incorporate conditioning. The diffusion-based pipeline often reports superior results; however, their structures are complex, and their inference time is slow. Papers using these paradigms typically address four different topics: ***conventional voice conversion*** counting zero-shot, many-to-many and general output quality improvements (n = 95), ***style conversion*** counting accent, dialect, and prosody conversion (n = 16), and ***emotion conversion*** (n = 13) and ***voice control*** works focussing on externalising parameter control (n = 7). The works primarily comprise the ***WORLD*** *vocoder* (n = 25), the ***WaveNet*** (n = 24), and the ***HiFi-GAN*** vocoder (21), with the former being prevalent before the 2020s and the latter gaining popularity after its introduction in 2019. Depending on the topic addressed and vocoders used, the works additionally differ in the loss functions included. ***Reconstruction losses*** (n = 77) and ***adversarial loss*** functions (n = 36) are frequently used but are often extended by ***cycle-consistency losses*** (n = 21), constraining the generator, ***Kullback–Leibler divergence*** (n = 17), ensuring that the learned distribution is similar to the true prior distribution, and ***identity-mapping losses*** (n = 13) or ***content losses*** (n = 3) providing contextual robustness.

5.2 Challenging areas in the VC community

Although it is clear that VC is a promising, fast-growing research area, with the proposed models reaching high-quality output in terms of speaker similarity and naturalness, some challenges remain to be overcome. Drawing from our discoveries, we formulate five challenges that are commonly addressed in VC research and offer a synopsis of how the analysed studies endeavour to resolve them.

5.2.1 Architectural challenges

Although GANs, CVAEs, and CAEs have proven high-fidelity results in different voice conversion scenarios, each architecture comes with limitations. GANs come with a theoretical justification of their target distribution matching capabilities, making them useful specifically for approximating the probability distribution of natural speech. However, GANs are widely known to be very difficult to train, and their convergence properties are fragile. Simultaneously, it should be noted that even though many generators in GANs may be able to fool a discriminator, they are yet to fool human ears (Qian et al., 2019). To overcome these challenges, extensions such as CycleGAN have been proposed. However, the inclusion of cycle-consistency losses provides new problems. As CycleGAN does not explicitly model an internal representation, it may not prove sufficient to individually manipulate features such as identity, prosody, or emotion. As mentioned by Sisman et al. (2020), a CycleGAN is simply more suitable for VC between a specific source and target pairs and not for many-to-many conversion. In contrast to GANs, AE and VAE-based structures tend to generate over-smoothed output. Although they provide training stability and easy access to conditioning features, the over-smoothing usually results in poor quality, “buzzy-sounding” speech (Kameoka et al., 2019). Additionally, they are designed to produce acoustic features frame-by frame, making it difficult to learn sequential dependencies in the intermediate acoustic features. Lastly, some A-S structures, as in the works by Nercessian (2021) and Nercessian (2020), provide successful zero-shot conversion and feature manipulation through explicit conditioning signals. However, their oscillator-driven, differentiable vocoders often limit the output quality, and their rather simple encoding structures may be insufficient in complete disentanglement. The degraded output quality may be improved by black-box post-nets for such architectures; however, these will impose higher complexity on the system.

To this end, adversarial AEs/VAEs and A-S structures, as analysed in Section 4.2, emerge as well-documented choices satisfying over-smoothing, training stability, and flexible manipulation. Broadly speaking, they conceptualise VC as a modular bottleneck, wherein individual features can be independently processed and subsequently concatenated to construct a comprehensive information cluster for the decoder. This appears to be a promising methodology. Choi et al. (2021), Wang Y. et al. (2022), and Xie et al. (2022) process the input in four parallel streams, each stream taking care of its own feature. These studies promote both feature disentanglement and provide controllable waveform reconstruction directly in the time domain and a system that can be trained in a stable, end-to-end manner. All three works provide state-of-the-art results.

5.2.2 Disentangling features

Despite the information bottleneck's objective of disentangling speech content and speaker characteristics, a notable amount of prosodic information, such as source pitch, may leak through the bottleneck if it is not added to the latent features. Simultaneously, the naturalness of the converted speech may decrease due to the difference in information contained within the bottleneck features themselves. Theoretically, these problems could be solved by a speaker embedding that contains the target speaker's prosodic information. However, this requires hours of data for each speaker as the speaker encoder would have to learn to reflect such characteristics in its embeddings. As mentioned above, it instead has been proposed to disentangle all three features: speech content, F0, and identity (Qian et al., 2020a). Opposed to the architectures mentioned in the former section, where information is only concatenated in the latent space, this may be done in various ways. Huang et al. (2019), Wang Y. et al. (2022), and Tan et al. (2021) do it in the generator or the vocoder individually as well as in both. In all cases, the F0 conditioning improves the naturalness and similarity of the generated speech. However, even though parts of the system were conditioned on the F0, the amount of F0 information entangled in the content and speaker embeddings was highly dependent on the input itself. More specifically, the amount of entanglement was related to the amount of F0 information that resided in the spectral features of the input. For this reason, most studies found it more efficient to disentangle the content and speaker features from input representations such as the MCCs or MCEPs, rather than the conventional and harmonically rich mel spectrogram. In contrast, Chun et al. (2023) and Hwang et al. (2022) use the wav2vec representation as content-input due to its linguistic-informed but high-level layers.

In general, self-supervised content encoders based on Wav2vec or HuBERT show promise in feature disentanglement, as they have demonstrated the ability to generate valuable linguistic units without relying on transcripts, require no explicit labelling, and operate directly on the waveform. However, not all self-supervised content bottlenecks are free of speaker information. It may especially prove challenging to eliminate speaker-related information in the content embeddings when they are derived directly from HuBERT and wav2vec models without the additional clustering process (Li et al., 2023). To address this challenge, we refer to ContentVec, as proposed by Qian et al. (2022), although it was not indexed in our literature search. ContentVec adds speech disentanglement techniques into a HuBERT-based masked-prediction pipeline in order to obtain linguistic representations free of speaker information. More specifically, they incorporate three different methods into the pipeline: label-generation (teacher), speech representation (student), and prediction. First, they obscure speaker identity in the target labels by converting all utterances to sound like that of the same speaker using a pretrained VC model. Like the work by Choi et al. (2021), they additionally perturb input samples in parallel before they are masked and fed to the speech representation network. Using a contrastive loss on the parallel streams, they thereby compel representations of the same utterance to exhibit similarity regardless of perturbation or speaker variation. Lastly, they condition the predictor on the actual speaker information, ensuring that the student remains uninfluenced by it. Once the speech representation network has been trained in conjunction with the predictor to match generated labels, one can use the student to embed speaker-free linguistic information. In many downstream tasks, including VC, ContentVec outperforms wav2vec and HuBERT-based content representations (Qian et al., 2022). Nonetheless, a drawback of such self-supervised representations is their high computational complexity, rendering them unsuitable for real-time inference or streamable applications. In very recent research, we observe the use of self-supervised models for knowledge distillation. Specifically, a simple CNN-based content encoder can be trained to approximate the distribution of discrete units using clustered HuBERT-based representations as targets Yang et al. (2024). This approach yields a soft unit encoder that outperforms related self-supervised content encoders in terms of speed, disentanglement, and fidelity.

The self-supervised representations may be used to obtain prosodic information. Hwang et al. (2022) use the wav2vec-based content embedding to predict pitch using an external pitch-module, while the speaker embedding is retrieved from a style encoder fed with the source mel spectrogram. Both works report MCD and RMSE-F0 metrics that outperform those of the system explicitly conditioned on pitch. It is imperative to observe that these studies depend on non-interpretable and inefficient self-supervised networks, which may not inherently offer auxiliary pitch control. Nonetheless, note that the input representations wield an impact on the feature disentanglement process, which is why conducting experiments involving diverse inputs for the feature encoders within a system could prove beneficial.

Techniques such as instance normalisation (IN) and VQ can be used to further disentangle the speech features. Zhang Z. et al. (2020) and Huang et al. (2021) introduce adaptive instance normalisation (AdaIN) to adjust the speaker embedding to different styles on a per-instance basis. IN and AdaIN are traditional features in style transfer problems, as their inherent scale and bias parameters allow the speaker modulation to be transformed in a domain-specific manner, limiting it from bleeding into the content embedding (Kaneko et al., 2019b). VQ, on the other hand, is incorporated to confine the leakage of content information into the speaker representation. Wang D. et al. (2021), Wu and Lee (2020), and Chen and Hain (2020) apply VQ on the content embedding, modelling it as a series of discrete codes. The motivation behind employing VQ lies in the observation that the discrete latent codes acquired from VQ-based auto-encoders exhibit a strong correlation with phonemes. A vector quantised content embedding simply is a more condensed representation, providing only the needed information for the decoder. However, mapping continuous values to a set of discrete codebook entries may lead to the loss of fine-grained information. Therefore, it is crucial to have a large codebook size, which, in turn, may impose higher complexity and memory usage.

5.2.3 Approaching mismatch problems

Most VC pipelines suffer from mismatch problems. One such problem is the training-inference mismatch that happens when a VC system is trained using the same utterance from the same speaker, which is the case in most systems. Here, the same input sample is used for content and style/speaker embeddings, making the overall model prone to copying information. This becomes a problem in inference as the model here is presented with different samples, making it produce low output quality. Hwang et al. (2022) tackle the mismatch problem using adversarial style generalisation. More specifically, the style generalisation clusters representations using two different utterances of the same speaker during training. The style encoder is thus trained on utterances and optimised by minimising the difference between their output speaker embeddings. This creates a global style representation for each speaker that is robust in inference scenarios. A different approach to the speaker mismatch problem is to extract multiple speaker embeddings and create averages of these across every speaker. This will generate fixed speaker embeddings independent of the utterances (Tan et al., 2021). Lastly, we have seen the introduction of diffusion models to extent general speaker adaptation quality. Here, data-driven priors are used to improve conversion performance as they regulate the inception of the denoising process. Simultaneously, the diffusion may be conditioned on the speaker information, creating high-quality speech synthesis and high-quality zero-shot capabilities (Choi H.-Y. et al., 2023).

Another issue related to mismatch is the acoustic feature mismatch problem. This becomes apparent when there is a significant disparity between the acoustic representation and the generalisation capabilities of the vocoder used, resulting in reduced output quality. Du and Yao (2023) approach this by discarding the decoder completely, substituting it with a HiFi-GAN-based generator that upsamples the intermediate acoustic representation to match the dimensions of the time domain. In other works, the authors provide an inductive bias for the audio generation, combining neural vocoders with traditional speech/sound production models (Choi et al., 2021). In contrast, Xie et al. (2022) avoid the mismatch problem by introducing information perturbation. As an extension to the above-mentioned feature disentanglement methods, information perturbation aims at perturbing all useless information in the source speech through digital signal processing, thereby limiting the different sub-blocks from learning undesirable attributes (Xie et al., 2022). Specifically, Choi et al. (2021) perturb the audio that is given to the content encoder with pitch shifting, formant shifting, and random frequency shaping, forcing it to adhere only to the linguistics of the input. The input audio fed to the pitch encoder, on the other hand, is only perturbed using the latter two processes, whereas the input to the speaker encoder is unaltered. The process of information perturbation ensures that content and pitch features no longer provide speaker-related information, making the input to the speaker encoder unique. This way of controlling the information flow is reported to be significantly useful as it does not suffer from mismatch problems. Such an approach performs well on CER and SSIM metrics, unlike many other information bottlenecks (Choi et al., 2021). However, it should be noted that such systems still are dependent on the generation model included.

5.2.4 Voice control and interpretability

An inherent constraint in the greater part of the analysed voice conversion models lies in their capacity to only synthesise speech that is either present in the datasets used or defined by a speaker-specific embedding. Manipulating voice in order to create new voice identities or edit specific voice attributes similar to voice transformations remains a challenge for most conversion systems. This constraint is often restricted by the low-level representations that most models compose in their information bottleneck (Choi H.-S. et al., 2023). As mentioned, desirable control features such as pitch, energy or timbre may be entangled either in the latent space or in the traditional speaker-dependent and independent embeddings. Although some models are further conditioned, this choice does not necessarily force the decoder to learn the mapping of different pitch and energy representations. It is trivial to conclude that the more one decomposes a signal into high-level/interpretable representations, the more one can gain access to the controllability and thereby parameterise the model. However, comparatively little research has been devoted to creating speech that sounds like truly novel speakers.

Only six studies explicitly concentrate on voice control. Choi et al. (2021) specifically focussed on this. Using an A-S procedure, they explicitly condition their generation process on separated feature counting: linguistics, speaker embedding, pitch, and energy, which are perturbed using different DSP techniques. Inspired by the source-filter theory, they split the generation process into two: a source and a filter decoder, generating harmonic content and spectral envelopes, respectively. They simply incorporate inductive bias in the model by conditioning each generator on features important for the given generation task. This provides interpretability and formant preserving pitch-shifting capabilities (Choi et al., 2021). The source and filter representations are thereafter summed to represent a mel spectrogram that is fed to the vocoder of choice. Xie et al. (2022) adopt similar perturbation techniques. However, rather than separating the generation process, they feed the feature embeddings directly to one unified waveform generator. Because the generator synthesises speech directly in the time domain, it can easily control different speech attributes. Lastly, Wang Y. et al. (2022) introduce a method for controllable speech representation learning based on disentanglement only. This work adheres to a conventional CAE and A-S structure, wherein feature encoders and latent embeddings are sequentially arranged prior to their input into the decoder. Through additional incorporation of training guidance using reconstruction, content, AIC, and adversarial losses, the authors assert that they achieve a level of disentanglement sufficient for controllability.

Although the aforementioned works use different structures to achieve interpretability and controllability, they share similarities. First, they aim to accomplish disentanglement by controlling the information flow. Second, they incorporate discriminators to tune the output quality and naturalness. Lastly, they all incorporate either perturbation or intermediate features obtained by pretrained wav2vec networks for the linguistic and speaker representations. All works offer convincing results, and these studies should be considered inspirational sources when designing systems that enable explicit voice control in the future.

5.2.5 Real-time constraints

The efficiency of a voice conversion system during inference is bounded by the speed of the generator and vocoder used, the speed of converting the utterance between time and frequency domains, and the speed of the encoders, specifically the content and speaker encoders. As outlined in this study, most pipelines incorporate encoders based on deep convolutional blocks, recurrent neural networks, or large self-supervised models, while rather complex neural vocoders are included to synthesise the acoustic representations. Although this is done to ensure the best output quality, it limits the real-time possibilities of the models. It was also found that only four works of the studies analysed mention the importance of real-time or streamable voice conversion (Baas and Kamper, 2020; Yang H. et al., 2022; Himawan et al., 2022; Tanaka et al., 2023).

The specific focus on enabling live one-shot VC is articulated by Yang H. et al. (2022). Within this framework, each sub-block is carefully designed to facilitate streaming capabilities. This is achieved, in part, through implementing all convolution and self-attention layers as causal while ensuring that all recurrent structures are executed in a unidirectional manner. Lastly, the work adopts a cached sliding-window procedure that processes utterances chunk by chunk, making the pipeline applicable for buffer-based computation. The proposed model achieves a real-time factor of 0.37 on a single CPU, and although the work compromises on network structure, results suggest that the model achieves comparable one-shot VC performance with offline solutions (Yang H. et al., 2022). Similar approaches are taken by Tanaka et al. (2023), where networks are also implemented using causal layers. However, the authors here report degradation due to the “use of causal layers which masks future input information” (Tanaka et al., 2023). To take this into account, they propose knowledge distillation in which a “student” network, implemented as a streamable structure, learns in conjunction with the more complex, non-causal and non-streamable “teacher” network. The work by Tanaka et al. (2023) does not report any real-time factor.

Broadening our perspective, we discover that the approaches taken in the above studies are followed in new and similar work that was not indexed by the included databases. For instance, to adhere to streamable environments, Ning et al. (2023) and Ning et al. (2024) use unidirectional recurrent networks, causal convolution layers, and knowledge distillation in a teacher-student learning approach. Another interesting method is the ability to train traditional non-causal networks and subsequently perform a post-training causal reconfiguration of the trained model, as presented by Caillon and Esling (2022). This technique provides a promising foundation for real-time neural audio synthesis and voice conversion.

Although none of the indexed articles focused on streamability for IoT/edge devices, the newly released work by Yang et al. (2024) manages to perform high-quality VC with ~10 m inference latency on a Pixel 7 smartphone. Looking forward, VC applications are thus undoubtedly expected to operate on the distributed computational continuum (Pujol et al., 2023), where low-end devices can perform inference or training on edge nodes. This integration of IoT/edge intelligence (Taheri et al., 2023) will depend critically on techniques such as bottleneck quantisation for discretisation of latent codes (Abe et al., 1988; Wu et al., 2020), digit quantisation, model quantisation and model pruning (Sudholt et al., 2023), real-time training (Kaspersen et al., 2020), and offloading (Liu and Zhang, 2018). These advancements could make VC more accessible, affordable, and efficient while reducing energy and computational demands. Quantisation, in particular, plays a pivotal role in mitigating resource and energy constraints, although its detailed discussion in edge-based VC applications remains currently limited. Future research should provide practical insights into implementing edge-based VC systems, building upon the innovative approaches identified in current studies.

6 Conclusion and future prospects

VC is a rapidly evolving research area with numerous challenges and several approaches aimed at addressing them. This paper presented a scoping review of 130 papers in the field of VC. The papers were evaluated using a codebook of 14 codes covering research direction, contributions, methods, and deep learning structures employed. We also provided an overview of the most commonly used datasets, sampling rates, and loss functions.

At present, the VC community is focused on mitigating the information bottleneck to facilitate disentangled speech representation learning while simultaneously ensuring high-quality mapping between speech features and intermediate acoustic representations. A degradation of performance and output quality frequently occurs due to entanglement within the embedded feature representations, manifested as leakage between content, speaker, and prosody embeddings. Several techniques are used to retrieve and disentangle such information. For extracting linguistic content, the analysis indicates a consensus favouring the use of ASR-based structures. However, challenges persist, particularly regarding errors inherent in pretrained recognition models, which may lead to mispronunciation in the converted speech output. Despite minor representation, self-supervised models such as the wav2vec model offer promising solutions to this problem, providing time-aligned, speaker-independent linguistic embeddings beneficial for downstream tasks, especially for low-resource languages. Like linguistics, speaker embeddings that are traditionally retrieved using structures from speaker verification research may be derived from self-supervised representations. In contrast, embeddings such as pitch, rhythm, and energy are often integrated using DSP methods to supply prosodic information to the disentanglement process. Including such features proves advantageous as they enable the remaining encoders to focus solely on linguistic and timbre aspects while introducing inductive biases beneficial for synthesis.

Several techniques have been proposed to facilitate the transformation of speech-related features into high-quality outputs. Within this domain, the analysed work can be divided into two primary categories: a) the conversion of features into acoustic representations fed to a pretrained neural vocoder, and b) the use of generative decoders to directly map the features into speech in the time domain. Although the former methodology is susceptible to individual discrepancies and mismatches inherent in the pipeline itself, approaches such as diffusion-based modelling have demonstrated encouraging outcomes in associating speech features with high-fidelity, speaker-adaptive acoustic representations. Conversely, with regard to the latter approach, structures founded on GANs and adversarial learning have exhibited efficacy in generating high-quality audio directly in the time domain. Conditioning such models on pitch contour, pitch embeddings, or excitation signals may further promote output stability and matching of target speaker characteristics.

Although many of the proposed feature disentanglement processes and mapping strategies are reported to improve the output quality, the actual interpretability of the structures remains a challenge. In general, the literature presents a minimal focus on generalisability and voice control, reducing voice conversion to an offline identity conversion problem. This constraint may be overcome by specifically controlling the information flow of the system's bottleneck and by including real-time efficient model structures.

Based on these findings, we summarise the main challenges and provide several recommendations for prospective research.

- **Input representations and feature disentanglement:** It is clear that many possibilities exist for optimising the feature disentanglement in systems using the information bottleneck. However, the methods used highly depend on the use case and implementation device at hand. The feature disentanglement process may be improved by considering consistency learning, information perturbation, the amount of harmonic information present in the input, and the interplay between the content and speaker encoders themselves. Studies providing benchmarks for such claims, measuring disentanglement and the degree to which each method affects the output, do not yet exist.
- **Trustworthy content encoders:** Compared to speaker embeddings, it is still difficult to create embeddings that contain linguistic information only. Although many approaches such as perturbation, ASR-based encoders, PPGs and content loss are provided for successful content embeddings, entanglement and mismatch problems are still prominent. Continuing to develop encoders that can focus solely on linguistics and generalise to new languages is thus of high importance.
- **Analysis-synthesis structures:** Analysis-synthesis structures, closely related to CAEs, are proving useful in stable, high fidelity, and controllable VC. It is thus natural to assume that these architectures are useful in future research.
- **Explicit control and voice design:** Current VC architectures have inherent capabilities of controlling high-level attributes; however, comparatively little research has been devoted to creating speech that sounds like truly novel speakers. VC systems conditioned on speech descriptive characteristics such as age, perceived gender, and tone could broaden the practice of VC.
- **Interpretable structures:** As earlier mentioned, representations from self-supervised architectures such as wav2vec models are proving successful as input to content, speaker, and pitch encoders. Examining such architectures with interpretable deep learning strategies would allow researchers to transfer the self-supervised decision-making mechanisms to less-complex structures, for example, through knowledge distillation strategies.
- **High-fidelity acoustic representations:** When incorporating a neural vocoder, it is imperative to maintain a precise acoustic representation. Various factors influence this, including the level of disentanglement in the speech feature space and the decoder/generator utilised. Diffusion modelling has demonstrated promising results in generating accurately converted acoustic representations in zero-shot scenarios; however, their inference speed heavily relies on the sampling scheme employed. Therefore, developing novel sampling algorithms tailored for rapid synthesis and causal diffusion-based decoding structures presents intriguing avenues for future research.
- **High-fidelity waveform generation:** Even though many neural vocoders are making strides in improving their inference time, the majority still fall short in running real-time. Existing research indicates the potential for combining neural networks with traditional speech/sound production models to achieve efficient and robust performance. Nevertheless, waveform generation methods, such as DSP-informed and oscillator-driven approaches like the neural source-filter system or differentiable WORLD synthesiser, have not demonstrated adequate efficacy in producing high-quality output speech. The prospect of combining these two aspects to develop a low-latency, high-fidelity waveform generator presents an interesting avenue for future exploration.
- **Real-time requirements:** In some cases, latency and real-time efficiency may be needed for the VC application at hand. Student-teacher architectures, model pruning, or post-training causal reconfiguration schemes are interesting additions to existing, well-performing VC systems.

7 Limitations

Although a thorough search, guided by a meticulously crafted keyword list, has been conducted across two pertinent databases, most of the review process has been conducted individually by Author 1, with no forward-citation search on the included studies. However, efforts were made to mitigate potential selection bias through an objective and carefully selected keyword-selection process, as well as to address interpretation bias through ongoing, rigorous coding discussions and meetings involving all authors. Furthermore, no assessment of the quality of the reference list was conducted, as it was determined that the two included databases maintained a specific scientific standard. Even though scoping reviews typically encompass all available evidence regardless of methodological quality (Arksey and O'Malley, 2005), it was still decided to confine the search to full journal articles and paper proceedings. This choice was made to ensure a manageable outcome in terms of evidence.

Finally, we acknowledge the restricted coverage of transformer-based models in this review. Although the review encompasses VC techniques relevant to self-supervision, it maintains a limited focus on LM-related approaches. Although the methodology surrounding transformers and LMs offers promising directions for VC research, the extensive proliferation of LM studies across the wider field of AI has made it impractical to comprehensively incorporate them into this review. Consequently, a deliberate choice was made not to search for LM-related literature. Given the prominence of LMs, we believe their inclusion would require a dedicated review.

Author contributions

AB: conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing–original draft, and writing–review and editing. SS: conceptualization, project administration, supervision, validation, and writing–review and editing. CE: conceptualization, data curation, investigation, methodology, project administration, supervision, resources, validation, and writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Industrial PhD grant from the Innovation Fund of Denmark and the virtual reality and augmented reality production studio KhoraVR and its spin-off, Heka VR.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frsip.2024.1339159/full#supplementary-material

References

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. *ICASSP-88., Int. Conf. Acoust. Speech, Signal Process.* 1, 655–658. doi:10.1109/ICASSP.1988.196671

Al-Radhi, M. S., Csapó, T. G., and Németh, G. (2021). Effects of sinusoidal model on non-parallel voice conversion with adversarial learning. *Appl. Sci.* 11, 7489. doi:10.3390/app11167489

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi:10.1080/1364557032000119616

Baas, M., and Kamper, H. (2020). StarGAN-ZSVC: towards zero-shot voice conversion in low-resource contexts. *Proc. South. Afr. Conf. AI Res. (SACAIR) (Muldersdrift, South Afr.)* 1342, 69–84. doi:10.1007/978-3-030-66151-9_5

Baas, M., and Kamper, H. (2023). "Gan you hear me? reclaiming unconditional speech synthesis from diffusion models," in 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 09-12 January 2023, 906–911.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Advances in neural information processing systems*. Editors H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc.), 33, 12449–12460.

Bonnici, R. S., Benning, M., and Saitis, C. (2022). "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," in 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18-23 July 2022, 1–8.

Brice, S., Saunders, E., and Edwards, B. (2023). Scoping review for a global hearing care framework: matching theory with practice. *Seminars Hear.* 44, 213–231. doi:10.1055/s-0043-1769610

Caillon, A., and Esling, P. (2022). *Streamable neural audio synthesis with non-causal convolutions*. doi:10.48550/arXiv.2204.07064

Cao, Y., Liu, Z., Chen, M., Ma, J., Wang, S., and Xiao, J. (2020). Nonparallel emotional speech conversion using VAE-GAN. *Interspeech 2020 (ISCA)*, 3406–3410. doi:10.21437/Interspeech.2020-1647

Chen, M., and Hain, T. (2020). Unsupervised acoustic unit representation learning for voice conversion using WaveNet auto-encoders. *Proc. Interspeech* 2020, 4866–4870. doi:10.21437/Interspeech.2020-1785

Chen, Y.-N., Liu, L.-J., Hu, Y.-J., Jiang, Y., and Ling, Z.-H. (2022). "Improving recognition-synthesis based any-to-one voice conversion with cyclic training," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 7007–7011.

Cheveigné, A., and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* 111, 1917–1930. doi:10.1121/1.1458024

Choi, H., and Hahn, M. (2021). Sequence-to-Sequence emotional voice conversion with strength control. *IEEE Access* 9, 42674–42687. doi:10.1109/ACCESS.2021.3065460

Choi, H.-S., Lee, J., Kim, W., Lee, J. H., Heo, H., and Lee, K. (2021). "Neural analysis and synthesis: reconstructing speech from self-supervised representations," in *Advances in neural information processing systems (NeurIPS)*. Editors M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc.), 16251–16265. doi:10.48550/arXiv.2110.14513

Choi, H.-S., Yang, J., Lee, J., and Kim, H. (2023a). NANSY++: unified voice synthesis with neural analysis and synthesis. *Eleventh Int. Conf. Learn. Represent.* doi:10.48550/arXiv.2211.09407

Choi, H.-Y., Lee, S.-H., and Lee, S.-W. (2023b). Diff-HierVC: diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *Proc. INTERSPEECH* 2023, 2283–2287. doi:10.21437/Interspeech.2023-817

Chun, C., Lee, Y. H., Lee, G. W., Jeon, M., and Kim, H. K. (2023). "Non-parallel voice conversion using cycle-consistent adversarial networks with self-supervised representations," in 2023 IEEE 20th Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 08-11 January 2023, 931–932.

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., et al. (2014). Scoping reviews: time for clarity in definition, methods, and reporting. *J. Clin. Epidemiol.* 67, 1291–1294. doi:10.1016/j.jclinepi.2014.03.013

Dang, T., Tran, D., Chin, P., and Koishida, K. (2022). "Training robust zero-shot voice conversion models with self-supervised features," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 6557–6561.

Dhar, S., Jana, N. D., and Das, S. (2023). An adaptive-learning-based generative adversarial network for one-to-one voice conversion. *IEEE Trans. Artif. Intell.* 4, 92–106. doi:10.1109/tai.2022.3149858

Ding, Y.-Y., Liu, L.-J., Hu, Y., and Ling, Z.-H. (2022). "A study on low-latency recognition-synthesis-based any-to-one voice conversion," in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 07-10 November 2022, 455–460.

Du, R., and Yao, J. (2023). "High quality and similarity one-shot voice conversion using end-to-end model," in *Proceedings of the 2022 6th international conference on computer science and artificial intelligence* (New York, NY, USA: Association for Computing Machinery), CSAI '22), 284–288. doi:10.1145/3577530.3577575

Du, Z., Sisman, B., Zhou, K., and Li, H. (2022a). Disentanglement of emotional style and speaker identity for expressive voice conversion. In Interspeech 2022. *nil.* doi:10.21437/interspeech.2022-10249doi:10.48550/arXiv.2110.10326

Du, Z., Sisman, B., Zhou, K., and Li, H. (2022b). Disentanglement of emotional style and speaker identity for expressive voice conversion. *Proc. Interspeech* 2022, 2603–2607. doi:10.21437/Interspeech.2022-10249

Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). "Ddsp: differentiable digital signal processing," in *International conference on learning representations*. doi:10.48550/arXiv.2001.04643

Ferro, R., Obin, N., and Roebel, A. (2021). "Cyclegan voice conversion of spectral envelopes using adversarial weights," in *2020 28th European signal processing conference (EUSIPCO)*, 406–410. doi:10.23919/Eusipco47968.2020.9287643

Fu, C., Liu, C., Ishi, C. T., and Ishiguro, H. (2022). "Finding meaning in "wrong responses": the multiple object-awareness paradigm shows that visual awareness is probabilistic," in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 07-10 November 2022, 553–559. doi:10.3758/s13414-021-02398-8

Grant, M., and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Inf. Libr. J.* 26, 91–108. doi:10.1111/j.1471-1842.2009.00848.x

Gu, Y., Zhao, X., Yi, X., and Xiao, J. (2023). "Voice conversion using learnable similarity-guided masked autoencoder," in *Digital forensics and watermarking. Springer nature Switzerland), lecture notes in computer science*. Editors X. Zhao, Z. Tang, P. Comesaña-Alfaro, and A. Piva (Cham), 53–67. doi:10.1007/978-3-031-25115-3_4

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., et al. (2020). Conformer: convolution-augmented transformer for speech recognition. *Proc. INTERSPEECH*, 5036–5040. doi:10.21437/interspeech.2020-3015

He, X., Chen, J., Rizos, G., and Schuller, B. W. (2021). An improved StarGAN for emotional voice conversion: enhancing voice quality and data augmentation. *Proc. Interspeech* 2021, 821–825. doi:10.21437/Interspeech.2021-1253

Himawan, I., Wang, R., Sridharan, S., and Fookes, C. (2022). Jointly trained conversion model with LPCNet for any-to-one voice conversion using speaker-independent linguistic features. *IEEE Access* 10, 134029–134037. doi:10.1109/ACCESS.2022.3226350

Ho, T. V., and Akagi, M. (2021). Cross-lingual voice conversion with controllable speaker individuality using variational autoencoder and star generative adversarial network. *IEEE Access* 9, 47503–47515. Conference Name: IEEE Access. doi:10.1109/ACCESS.2021.3063519

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech Lang. Proc.* 29, 3451–3460. doi:10.1109/TASLP.2021.3122291

Huang, J., Xu, W., Li, Y., Liu, J., Ma, D., and Xiang, W. (2022). "FlowCPCVC: a contrastive predictive coding supervised flow framework for any-to-any voice conversion," in Interspeech 2022 *(ISCA)*, 2558–2562. doi:10.21437/Interspeech.2022-577

Huang, S., Chen, M., Xu, Y., Ke, D., and Hain, T. (2021). WINVC: one-shot voice conversion with weight adaptive instance normalization. In *Pricai 2021: trends in artificial intelligence*, eds. D. N. Pham, T. Theeramunkong, G. Governatori, and F. Liu (Cham: Springer International Publishing), Lecture Notes in Computer Science, 559–573. doi:10.1007/978-3-030-89363-7_42

Huang, W.-C., Luo, H., Hwang, H.-T., Lo, C.-C., Peng, Y.-H., Tsao, Y., et al. (2020a). Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion. *IEEE Trans. Emerg. Top. Comput. Intell.* 4, 468–479. doi:10.1109/TETCI.2020.2977678

Huang, W.-C., Wu, Y.-C., Hayashi, T., and Toda, T. (2020b). "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 06-11 June 2021, 5944–5948.

Huang, W.-C., Wu, Y.-C., Lo, C.-C., Tobing, P. L., Hayashi, T., Kobayashi, K., et al. (2019). Investigation of F0 conditioning and fully convolutional networks in variational autoencoder based voice conversion. *Proc. Interspeech* 2019, 709–713. doi:10.21437/Interspeech.2019-1774

Hwang, I.-S., Lee, S.-H., and Lee, S.-W. (2022). "StyleVC: non-parallel voice conversion with adversarial style generalization," in 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21-25 August 2022, 23–30.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., et al. (2018). "Efficient neural audio synthesis," in *Proceedings of the 35th international conference on machine learning*. Editors J. Dy and A. Krause, 2410–2419. (PMLR), vol. 80 of Proceedings of Machine Learning Research. doi:10.48550/arXiv.1802.08435

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2019). Acvae-vc: non-parallel voice conversion with auxiliary classifier variational autoencoder. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 27, 1432–1443. doi:10.1109/TASLP.2019.2917232

Kaneko, T., and Kameoka, H. (2018). "CycleGAN-VC: non-parallel voice conversion using cycle-consistent adversarial networks," in 2018 26th European Signal Processing Conference (EUSIPCO) (Rome: IEEE), Rome, Italy, 03-07 September 2018, 2100–2104.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019a). "Cyclegan-vc2: improved cyclegan-based non-parallel voice conversion," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6820–6824.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019b). StarGAN-VC2: rethinking conditional methods for StarGAN-based voice conversion. *Proc. Interspeech* 2019, 679–683. doi:10.21437/Interspeech.2019-2236

Kaspersen, E., Górny, D., Erkut, C., and Palamas, G. (2020). "Generative choreographies: the performance dramaturgy of the machine," in *Proc. Intl. Joint conf. Computer vision, Imaging and computer graphics Theory and applications*. doi:10.5220/0008990403190326

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds. *Speech Commun.* 27, 187–207. doi:10.1016/S0167-6393(98)00085-5

Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (2018). "Crepe: a convolutional representation for pitch estimation," in *Proc. Intl. Conf. Acoustics, speech, and signal proc. (ICASSP)*, 161–165. doi:10.1109/icassp.2018.8461329

Kim, K.-W., Park, S.-W., Lee, J., and Joe, M.-C. (2022). "Assem-vc: realistic voice conversion by assembling modern speech synthesis techniques," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 6997–7001.

Lee, Y. K., Kim, H. W., and Park, J. G. (2021). Many-to-Many unsupervised speech conversion from nonparallel corpora. *IEEE Access* 9, 27278–27286. doi:10.1109/ACCESS.2021.3058382

Li, S., Zhang, Q., Li, Y., Li, G., Li, S., and Wang, S. (2023). "Analyzing speaker information in self-supervised models to improve unsupervised speech recognition," in *Proceedings of the 2022 6th international conference on electronic information technology and computer engineering* (New York, NY, USA: Association for Computing Machinery), EITCE '22, 1300–1305. doi:10.1145/3573428.3573659

Li, W., and Wei, T.-J. (2022). "ASGAN-VC: one-shot voice conversion with additional style embedding and generative adversarial networks," in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 07-10 November 2022, 1932–1937.

Lian, J., Zhang, C., Anumanchipalli, G. K., and Yu, D. (2022). Towards improved zero-shot voice conversion with conditional DSVAE. *Proc. Interspeech* 2022, 2598–2602. doi:10.21437/Interspeech.2022-11225

Liang, X., Bie, Z., and Ma, S. (2022). "Pyramid attention CycleGAN for non-parallel voice conversion," in 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 09-12 December 2022, 139–143.

Liu, A. T., wen Yang, S., Chi, P.-H., Hsu, P.-C., and yi Lee, H. (2019). "Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6419–6423.

Liu, F., Wang, H., Peng, R., Zheng, C., and Li, X. (2021). U2-vc: one-shot voice conversion using two-level nested u-structure. *EURASIP J. Audio Speech Music Process.* 2021, 40. doi:10.1186/s13636-021-00226-3

Liu, J., and Zhang, Q. (2018). Offloading schemes in mobile edge computing for ultra-reliable low latency communications. *IEEE Access* 6, 12825–12837. doi:10.1109/access.2018.2800032

Long, Z., Zheng, Y., Yu, M., and Xin, J. (2022). "Enhancing zero-shot many to many voice conversion via self-attention vae with structurally regularized layers," in 2022 5th International Conference on Artificial Intelligence for Industries (AI4I), 59–63.

Lu, W., Xing, X., Xu, X., and Zhang, W. (2021). "Towards unseen speakers zero-shot voice conversion with generative adversarial networks," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14-17 December 2021, 854–858.

Luo, Z., Lin, S., Liu, R., Baba, J., Yoshikawa, Y., and Ishiguro, H. (2023). Decoupling speaker-independent emotions for voice conversion via source-filter networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 31, 11–24. doi:10.1109/TASLP.2022.3190715

Mohammadi, S. H., and Kim, T. (2019). One-shot voice conversion with disentangled representations by leveraging phonetic posteriorgrams. *Interspeech 2019 (ISCA)*, 704–708. doi:10.21437/Interspeech.2019-1798

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* E99, 1877–1884. doi:10.1587/transinf.2015EDP7457

Munn, Z., Peters, M., Stern, C., Tufanaru, C., Mcarthur, A., and Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* 18, 143. doi:10.1186/s12874-018-0611-x

Nercessian, S. (2020). Improved zero-shot voice conversion using explicit conditioning signals. *Interspeech 2020 (ISCA)*, 4711–4715. doi:10.21437/Interspeech.2020-1889

Nercessian, S. (2021). "End-to-End zero-shot voice conversion using a DDSP vocoder," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (New Paltz, NY, USA: IEEE), New Paltz, NY, USA, 17-20 October 2021, 1–5.

Nguyen, B., and Cardinaux, F. (2022). "Nvc-net: end-to-end adversarial voice conversion," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 7012–7016.

Nguyen, T. N., Pham, N.-Q., and Waibel, A. (2022). Accent conversion using pre-trained model and synthesized data from voice conversion. *Interspeech 2022 (ISCA)*, 2583–2587. doi:10.21437/Interspeech.2022-10729

Nikonorov, S., Sisman, B., Zhang, M., and Li, H. (2021). "DeepA: a deep neural analyzer for speech and singing vocoding," in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13-17 December 2021.

Ning, Z., Jiang, Y., Zhu, P., Wang, S., Yao, J., Xie, L., et al. (2024). "DualVC 2: dynamic masked convolution for unified streaming and non-streaming voice conversion," in 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 11106–11110.

Ning, Z., Jiang, Y., Zhu, P., Yao, J., Wang, S., Xie, L., et al. (2023). DualVC: dual-mode voice conversion using intra-model knowledge distillation and hybrid predictive coding. *Proc. INTERSPEECH*, 2063–2067. doi:10.21437/interspeech.2023-1157

Paisa, R., Nilsson, N. C., and Serafin, S. (2023). Tactile displays for auditory augmentation–a scoping review and reflections on music applications for hearing impaired users. *Front. Comput. Sci.* 5. doi:10.3389/fcomp.2023.1085539

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M., and Wei, J. (2022). *Diffusion-based voice conversion with fast maximum likelihood sampling scheme*. doi:10.48550/arXiv.2109.13821

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The kaldi speech recognition toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Pujol, V. C., Donta, P. K., Morichetta, A., Murturi, I., and Dustdar, S. (2023). Edge intelligence-research opportunities for distributed computing continuum systems. *IEEE Internet Comput.* 27, 53–74. doi:10.1109/mic.2023.3284693

Qian, K., Jin, Z., Hasegawa-Johnson, M. A., and Mysore, G. J. (2020a). "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6284–6288.

Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. (2020b). "Unsupervised speech decomposition via triple information bottleneck," in *International conference on machine learning* (PMLR), 7836–7846. doi:10.48550/arXiv.2004.11284

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). "AutoVC: zero-shot voice style transfer with only autoencoder loss," in *Proceedings of the 36th international conference on machine learning*. Editors K. Chaudhuri and R. Salakhutdinov, 5210–5219. (PMLR), vol. 97 of Proceedings of Machine Learning Research. doi:10.48550/arXiv.1905.05879

Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.-I., Cox, D., et al. (2022). "ContentVec: an improved self-supervised speech representation by disentangling speakers," in *Proceedings of the 39th international conference on machine learning*. Editors K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, 18003–18017. (PMLR), vol. 162 of Proceedings of Machine Learning Research. doi:10.48550/arXiv.2204.09224

Reddy, M. K., and Rao, K. S. (2020). DNN-based cross-lingual voice conversion using Bottleneck Features. *Neural Process. Lett.* 51, 2029–2042. doi:10.1007/s11063-019-10149-y

Salinas-Marchant, C., and MacLeod, A. A. N. (2022). Audiovisual speech perception in children: a scoping review. *Speech, Lang. Hear.* 25, 433–449. doi:10.1080/2050571X.2021.1923302

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: unsupervised pre-training for speech recognition. *Proc. INTERSPEECH*, 3465–3469. doi:10.21437/Interspeech.2019-1873

Shi, S., Shao, J., Hao, Y., Du, Y., and Fan, J. (2022). "U-GAT-VC: unsupervised generative attentional networks for non-parallel voice conversion," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 7017–7021.

Singla, Y. K., Shah, J., Chen, C., and Shah, R. R. (2022). "What do audio transformers hear? probing their representations for language delivery 'i&' structure," in 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 910–925.

Sisman, B., Yamagishi, J., King, S., and Li, H. (2020). An overview of voice conversion and its challenges: from statistical modeling to deep learning. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 132–157. doi:10.1109/TASLP.2020.3038524

Stephenson, C., Keskin, G., Thomas, A., and Elibol, O. H. (2019). Semi-supervised voice conversion with amortized variational inference. *Proc. Interspeech* 2019, 729–733. doi:10.21437/Interspeech.2019-1840

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. *Nature* 110, 311–312. doi:10.1038/110311a0

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.* 9, 21–29. doi:10.1109/89.890068

Stylianou, Y. (2009). "Voice transformation: a survey," in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 3585–3588.

Stylianou, Y., Cappe, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* 6, 131–142. doi:10.1109/89.661472

Sudholt, D., Wright, A., Erkut, C., and Valimaki, V. (2023). Pruning deep neural network models of guitar distortion effects. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 31, 256–264. doi:10.1109/taslp.2022.3223257

Sun, L., Li, K., Wang, H., Kang, S., and Meng, H. (2016). Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. *Proc. IEEE Intl. Conf. Multimedia Expo (ICME)*, 1–6. doi:10.1109/ICME.2016.7552917

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Corr. abs/1409*, 3215. doi:10.48550/arXiv.1409.3215

Taheri, J., Dustdar, S., Zomaya, A., and Deng, S. (2023). *Edge intelligence, from theory to practice*. Springer. doi:10.1007/978-3-031-22155-2

Talkin, D., and Kleijn, W. B. (1995). "A robust algoritm for pitch tracking (RAPT)," in *Speech coding and synthesis*. Editor W. B. Kleijn and K. K. Paliwal (San Diego, CA, United States: Elsevier), 495–518.

Tan, Z., Wei, J., Xu, J., He, Y., and Lu, W. (2021). "Zero-shot voice conversion with adjusted speaker embeddings and simple acoustic features," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 06-11 June 2021, 5964–5968.

Tanaka, K., Kameoka, H., Kaneko, T., and Seki, S. (2023). "Distilling sequence-to-sequence voice conversion models for streaming conversion applications," in 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 09-12 January 2023, 1022–1028.

Tang, H., Zhang, X., Wang, J., Cheng, N., and Xiao, J. (2022). "Avqvc: one-shot voice conversion by vector quantization with applying contrastive learning," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022.

Tricco, A., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Levac, D., et al. (2018). Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Ann. Intern. Med.* 169, 467–473. doi:10.7326/M18-0850

Valbret, H., Moulines, E., and Tubach, J. (1992). "Voice transformation using psola technique," in ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, 23-26 March 1992, 145–148.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: a generative model for raw audio. *Proc. 9th ISCA Workshop Speech Synthesis Workshop (SSW 9)* 125. doi:10.48550/arXiv.1609.03499

van Niekerk, B., Carbonneau, M.-A., Zaïdi, J., Baas, M., Seuté, H., and Kamper, H. (2022). "A comparison of discrete and soft speech units for improved voice conversion," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 6562–6566.

Walczyna, T., and Piotrowski, Z. (2023). Overview of voice conversion methods based on deep learning. *Appl. Sci.* 13, 3100. doi:10.3390/app13053100

Wang, D., Deng, L., Yeung, Y. T., Chen, X., Liu, X., and Meng, H. (2021a). VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *Proc. Interspeech* 2021, 1344–1348. doi:10.21437/Interspeech.2021-283

Wang, J., Li, J., Zhao, X., Wu, Z., Kang, S., and Meng, H. (2021b). Adversarially learning disentangled speech representations for robust multi-factor voice conversion. *Proc. Interspeech* 2021, 846–850. doi:10.21437/Interspeech.2021-1990

Wang, Q., Zhang, X., Wang, J., Cheng, N., and Xiao, J. (2022a). "Drvc: a framework of any-to-any voice conversion with self-supervised learning," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022 (IEEE), 3184–3188.

Wang, S., Kostadinov, D., and Borth, D. (2022b). Zero-shot voice conversion via self-supervised prosody representation learning. *Intl. Jt. Conf. Neural Netw. (IJCNN)*, 1–8. doi:10.1109/IJCNN55064.2022.9892405

Wang, Y., Su, J., Finkelstein, A., and Jin, Z. (2022c). "Controllable speech representation learning via voice conversion and AIC loss," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022, 6682–6686.

Wang, Z., Chen, Y., Xie, L., Tian, Q., and Wang, Y. (2023). Lm-Vc: zero-shot voice conversion via speech generation based on language models. *IEEE Signal Process. Lett.* 30, 1157–1161. doi:10.1109/lsp.2023.3308474

Wang, Z., Ge, W., Wang, X., Yang, S., Gan, W., Chen, H., et al. (2020). "Accent and speaker disentanglement in many-to-many voice conversion," in 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, 24-27 January 2021, 1–5.

Wu, D.-Y., Chen, Y.-H., and yi Lee, H. (2020). VQVC+: one-shot voice conversion by vector quantization and u-net architecture. *Interspeech 2020*. doi:10.21437/interspeech.2020-1443

Wu, D.-Y., and Lee, H.-y. (2020). "One-shot voice conversion by vector quantization," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 04-08 May 2020, 7734–7738.

Wu, Y.-C., Hayashi, T., Tobing, P. L., Kobayashi, K., and Toda, T. (2021). Quasi-periodic WaveNet: an autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 1134–1148. doi:10.1109/taslp.2021.3061245

Wu, Z., Virtanen, T., Kinnunen, T., and Chng, E. S. (2013). "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *8th ISCA speech synthesis workshop*, 201–206.

Xie, Q., Yang, S., Lei, Y., Xie, L., and Su, D. (2022). "End-to-end voice conversion with information perturbation," in 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) (IEEE), Singapore, Singapore, 11-14 December 2022, 91–95.

Yamagishi, J., Veaux, C., and MacDonald, K. (2019). *CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit*. version 0.92. doi:10.7488/ds/2645

Yang, H., Deng, L., Yeung, Y. T., Zheng, N., and Xu, Y. (2022a). Streamable speech representation disentanglement and multi-level prosody modeling for live one-shot voice conversion. *Interspeech 2022 (ISCA)*, 2578–2582. doi:10.21437/Interspeech.2022-10277

Yang, J., Zhou, Y., and Huang, H. (2023). Mel-S3R: combining Mel-spectrogram and self-supervised speech representation with VQ-VAE for any-to-any voice conversion. *Speech Commun.* 151, 52–63. doi:10.1016/j.specom.2023.05.004

Yang, S., Tantrawenith, M., Zhuang, H., Wu, Z., Sun, A., Wang, J., et al. (2022b). Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *Proc. INTERSPEECH (arXiv)*, 2553–2557. doi:10.21437/interspeech.2022-571

Yang, Y., Kartynnik, Y., Li, Y., Tang, J., Li, X., Sung, G., et al. (2024). "Streamvc: real-time low-latency voice conversion," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 11016–11020.

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R., Kinnunen, T., et al. (2020). "Intra-lingual semi-parallel and cross-lingual voice conversion," in *Voice conversion challenge 2020*, 80–98.

Zang, X., Xie, F., and Weng, F. (2022). "Foreign accent conversion using concentrated attention," in 2022 IEEE International Conference on Knowledge Graph (ICKG), Orlando, FL, USA, 30 November 2022 - 01 December 2022, 386–391.

Zhang, H., Cai, Z., Qin, X., and Li, M. (2021). "Sig-vc: a speaker information guided zero-shot voice conversion system for both human beings and machines," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23-27 May 2022.

Zhang, J.-X., Ling, Z.-H., and Dai, L.-R. (2020a). Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 28, 540–552. doi:10.1109/TASLP.2019.2960721

Zhang, X., Wang, J., Cheng, N., and Xiao, J. (2023). "Voice conversion with denoising diffusion probabilistic GAN models," in *Advanced data mining and applications: 19th international conference, ADMA 2023, shenyang, China, august 21–23, 2023, proceedings, Part IV* (Berlin, Heidelberg: Springer-Verlag), 154–167.

Zhang, Z., He, B., and Zhang, Z. (2020b). GAZEV: GAN-based zero-shot voice conversion over non-parallel speech corpus. *Proc. Interspeech* 2020, 791–795. doi:10.21437/Interspeech.2020-1710

Zhao, G., Ding, S., and Gutierrez-Osuna, R. (2019a). Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. *Interspeech 2019 (ISCA)*, 2843–2847. doi:10.21437/Interspeech.2019-1778

Zhao, W., Wang, W., Sun, Y., and Tang, T. (2019b). "Singing voice conversion based on wd-gan algorithm," in 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20-22 December 2019, 950–954.

Zhou, K., Sisman, B., and Li, H. (2020). "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19-22 January 2021, 415–422.

Zhou, K., Sisman, B., Liu, R., and Li, H. (2021). "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 06-11 June 2021, 920–924.