frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 18 August 2022
DOI 10.3389/fnins.2022.906616

OPEN ACCESS

EDITED BY
Namkeun Kim,
Sogang University, South Korea

REVIEWED BY
Kyung Myun Lee,
Korea Advanced Institute of Science and Technology, South Korea
Waldo Nogueira,
Hannover Medical School, Germany
Bong Jik Kim,
Chungnam National University Sejong Hospital, South Korea

*CORRESPONDENCE
Jihwan Woo
jhwoo@ulsan.ac.kr

SPECIALTY SECTION
This article was submitted to Neuroprosthetics, a section of the journal Frontiers in Neuroscience

RECEIVED 28 March 2022
ACCEPTED 25 July 2022
PUBLISHED 18 August 2022

CITATION
Na Y, Joo H, Trang LT, Quan LDA and Woo J (2022) Objective speech intelligibility prediction using a deep learning model with continuous speech-evoked cortical auditory responses.
Front. Neurosci. 16:906616.
doi: 10.3389/fnins.2022.906616

Objective speech intelligibility prediction using a deep learning model with continuous speech-evoked cortical auditory responses

Youngmin Na[1], Hyosung Joo[2], Le Thi Trang[2], Luong Do Anh Quan[2] and Jihwan Woo[1,2]*

[1]Department of Biomedical Engineering, University of Ulsan, Ulsan, South Korea, [2]Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea

Auditory prostheses provide an opportunity for rehabilitation of hearing-impaired patients. Speech intelligibility can be used to estimate the extent to which the auditory prosthesis improves the user's speech comprehension. Although behavior-based speech intelligibility is the gold standard, precise evaluation is limited due to its subjectiveness. Here, we used a convolutional neural network to predict speech intelligibility from electroencephalography (EEG). Sixty-four–channel EEGs were recorded from 87 adult participants with normal hearing. Sentences spectrally degraded by a 2-, 3-, 4-, 5-, and 8-channel vocoder were used to set relatively low speech intelligibility conditions. A Korean sentence recognition test was used. The speech intelligibility scores were divided into 41 discrete levels ranging from 0 to 100%, with a step of 2.5%. Three scores, namely 30.0, 37.5, and 40.0%, were not collected. The speech features, i.e., the speech temporal envelope (ENV) and phoneme (PH) onset, were used to extract continuous-speech EEGs for speech intelligibility prediction. The deep learning model was trained by a dataset of event-related potentials (ERP), correlation coefficients between the ERPs and ENVs, between the ERPs and PH onset, or between ERPs and the product of the multiplication of PH and ENV (PHENV). The speech intelligibility prediction accuracies were 97.33% (ERP), 99.42% (ENV), 99.55% (PH), and 99.91% (PHENV). The models were interpreted using the occlusion sensitivity approach. While the ENV models' informative electrodes were located in the occipital area, the informative electrodes of the phoneme models, i.e., PH and PHENV, were based on the occlusion sensitivity map located in the language processing area. Of the models tested, the PHENV model obtained the best speech intelligibility prediction accuracy. This model may promote clinical prediction of speech intelligibility with a comfort speech intelligibility test.

KEYWORDS

speech intelligibility, deep-learning, continuous speech, occlusion sensitivity, EEG

Introduction

Auditory prostheses, such as hearing aids and cochlear implants, provide an excellent opportunity for hearing-impaired patients to rehabilitate their auditory modality. The outcome of auditory prosthesis use depends on the signal processing strategies: modulation of the current pulse train from sound in cochlear implants (CI) (Macherey et al., 2006; Wouters et al., 2015; Nogueira et al., 2019) or reduction of stationary and background noise and customized personal setting of hearing aids (Launer et al., 2016). In addition, the individual's status, such as the insertion depth of the CI electrode, and the experience of cochlear implantation, could also affect the performance of CI (Vandali et al., 2000; Wanna et al., 2014). To evaluate the benefit of auditory prostheses, a behavioral speech intelligibility test is typically conducted using rating scales based on how well the listener comprehends sentences (Kim et al., 2009). In this behavioral test, a listener is asked to repeat or write what they hear in a recognition test. Speech intelligibility is estimated by scoring the number of correctly identified words (Enderby, 1980; Kent et al., 1989; Healy et al., 2015; Lee, 2016). Although the behavioral assessment can be conducted efficiently and quickly, a self-reported approach may be less reliable and less sensitive in evaluating the true hearing capability (Koelewijn et al., 2018). Vocoder simulation has also been used in speech tests to simulate the performance of hearing impairment in normal-hearing listeners (Mehta and Oxenham, 2017).

Event-related potentials (ERPs), in response to word or tone stimuli, have been used to evaluate auditory function objectively. Recently, several studies have shown that electroencephalography (EEG) signals in response to continuous speech stimuli are entrained to speech features: temporal envelope, spectrogram, and phonetics of speech (Scott et al., 2000; Liebenthal et al., 2005; Nourski et al., 2009; Ding and Simon, 2014; O'Sullivan et al., 2015; Crosse et al., 2016; Di Liberto et al., 2018). The speech temporal envelope (ENV), developed using the temporal response function model (TRF), is an effective feature to understand neural responses to continuous speech (Ciccarelli et al., 2019; Nogueira and Dolhopiatenko, 2020, 2022). However, the TRF model is limited in analyzing short (<5 s) responses due to the impact of onset response to a sentence (Crosse et al., 2016, 2021). Therefore, cross-correlation, which measures the similarity between the neural response and the speech sentence, can be more reliable in tracking neural signals in response to short sentences.

It was reported that speech intelligibility affected ENV entrainment (Ding and Simon, 2013; Vanthornhout et al., 2018; Lesenfants et al., 2019; Nogueira and Dolhopiatenko, 2022). Sentence comprehension requires complex hierarchical stages that integrate the phonological and prosodic processes of an acoustic input (Snedeker and Huang, 2009). Vanthornhout et al. (2018) developed a prediction model for the speech reception threshold using the TRF model, which could explain the variance of speech reception. Moreover, Di Liberto et al. (2015) showed that a speech prediction model with a phonetic feature was outperformed by the envelope model. Thus, a combination of ENV and phoneme (PH) onset information can be effective for feature computation. However, predicting a speech intelligibility score from EEG signals to continuous stimuli with a linear input-output model is still a challenge. Recently, deep learning models have been widely used to classify auditory neural outcomes (Ciccarelli et al., 2019; Craik et al., 2019; Roy et al., 2019; Nogueira and Dolhopiatenko, 2020). Ciccarelli et al. (2019), showed that the non-linear model for decoding of auditory attention outperformed the linear model. As a sentence is non-linearly and hierarchically processed in the human brain along the complex auditory pathway, a non-linear model can perform better in predicting speech intelligibility. Thereby, deep learning can be successfully used in a non-linear model to investigate auditory neural processing. Deep learning requires two essential processes for better predictive performance. First, the reduction of attribute noise, which leads to a decrease in overfitting and memorization of noise data, can be achieved by neural tracking with speech features from EEG (Zhu and Wu, 2004; Altaheri et al., 2021; Cherloo et al., 2021; Zhou et al., 2021). Second, data augmentation increases the amount of data and helps to overcome the problem of limited data (Lashgari et al., 2020).

Although accurate classification is achieved through deep learning, it is essential to interpret the results for clinical use. The explainable deep learning models, the gradient-weighted class activation map (Grad-CAM), and the occlusion analysis map have been developed and applied to the classification tasks of an EEG data model (Jonas et al., 2019; Li et al., 2020; Mansour et al., 2020; Uyttenhove et al., 2020; Lombardi et al., 2021). While the Grad-CAM typically highlights the important lesion, the occlusion analysis map tracks multi-focal lesions and thus supports information with higher spatial resolution (Oh et al., 2020; Aminu et al., 2021; Govindarajan and Swaminathan, 2021). Occlusion analysis has been used to discover cortical areas related to movement tasks in EEG classification and identify important regions for image classification (Zeiler and Fergus, 2014; Ieracitano et al., 2021). In this study, we developed a deep learning model to predict speech intelligibility scores with EEG signals to continuous sentences. The typical speech features of ENV and phoneme onset impulse were used. An occlusion sensitivity map was used to select sensitive EEG channels to predict speech intelligibility scores (Esmaeilzadeh et al., 2018; Singh et al., 2020).

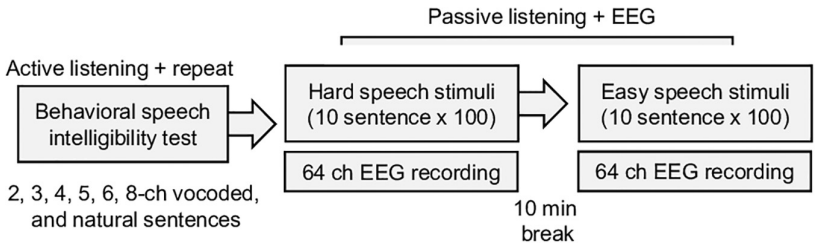

FIGURE 1
Summary of the experimental procedure for the behavior speech intelligibility test and EEG recording. During the behavioral test, vocoded noise, and natural sentence speeches are randomly played, and the participants are asked to repeat the sentences. The electroencephalogram (EEG) responses to the speech stimuli are recorded during the passive listening task.

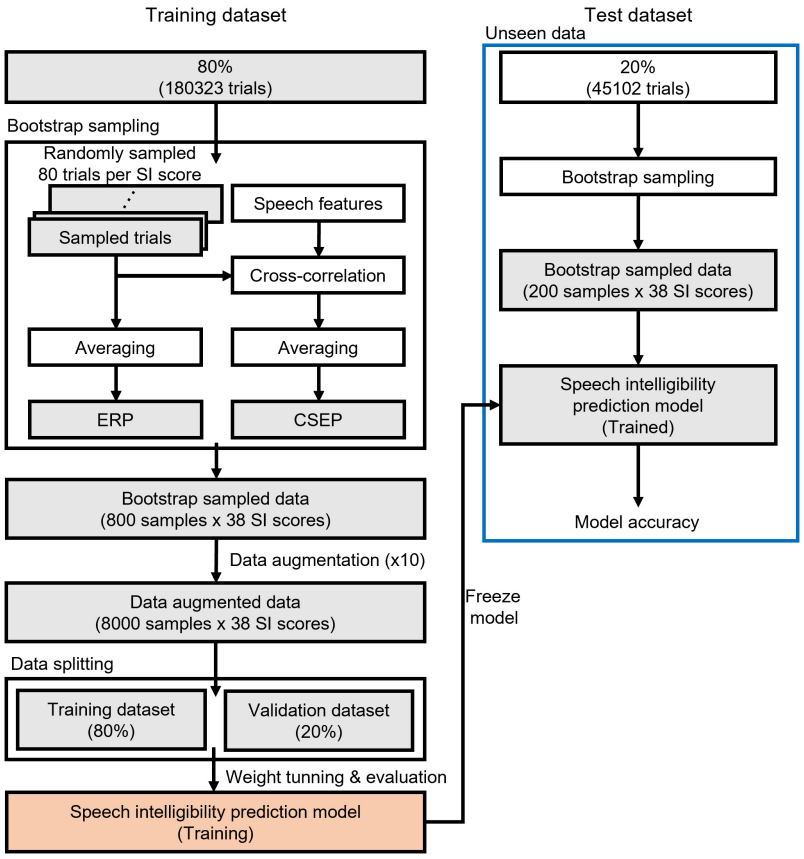

FIGURE 2
Schematic diagram of deep learning training and testing. A training dataset is used to build up a speech intelligibility prediction model and an unseen (test) dataset determines the performance of the model.

Materials and methods

Participants

Eighty-seven participants with normal hearing (44 males and 43 females) participated in this study. They were 20–33 years old (mean = 24.0 and standard deviation = 2.4). All experimental procedures and the written informed consent procedure were reviewed and approved by the Institutional Review Board of the University of Ulsan.

Stimuli

Ten continuous sentences spoken by a male speaker were selected from the Korean standard sentence list for adults

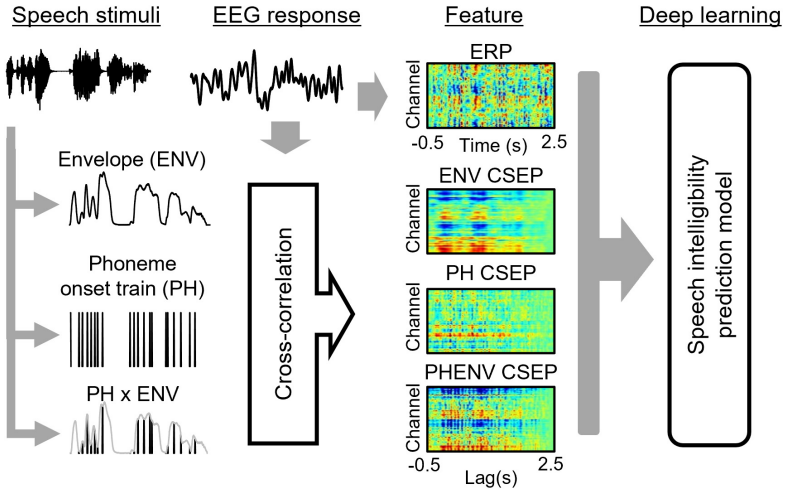

FIGURE 3
The overall scheme of speech intelligibility prediction, including speech stimuli, electroencephalogram data, feature extraction, and model prediction. The speech features of envelope (ENV), phoneme onset train (PH), and ENV ⊗ PH generate the features of ENV continuous-speech evoked potentials (CSEP), PH CSEP, and PHENV CSEP by cross-correlating with the electroencephalogram response (EEG).

(Jang et al., 2008). The duration of each sentence was 1.8 ± 0.2 s, and the number of phonemes in each sentence was 18.6 ± 3.9. The natural (non-vocoded) and noise-vocoded sentences were used in this study. A noise vocoder was used to simulate poor sensitivity with normal-hearing listeners. The vocoder consisted of a logarithmically-spaced filter bank between 200 and 5,000 Hz. Natural sentences are then filtered through the filter bank, which is modulated with a Gaussian white noise and synthesized sequentially (Mehta and Oxenham, 2017). The channel of vocoder parameter was set to 2, 3, 4, 5, and 8 for five noise-vocoded conditions, wherein a lower number of channels generated more spectrally degraded stimuli.

TABLE 1 Deep learning layers and their specifications.

Deep learning layer	Filter size	Kernel dimension (H × W)	Output (H × W × D)
Input			299 × 299 × 3
Conv2D	32	16 × 16	299 × 299 × 32
LeakyReLU			299 × 299 × 32
Conv2D	8	8 × 8	299 × 299 × 8
LeakyReLU			299 × 299 × 8
Maxpooling2D		2 × 2	149 × 149 × 8
Conv2D	8	4 × 4	149 × 149 × 8
LeakyReLU			149 × 149 × 8
Maxpooling2D		2 × 2	148 × 148 × 8
Conv2D	3	3 × 3	148 × 148 × 3
LeakyReLU			148 × 148 × 3
Maxpooling2D		2 × 2	147 × 147 × 3
Batch normalization			147 × 147 × 3
Fully connected		1 × 38	1 × 1 × 38
Softmax			1 × 1 × 38
Classification			38

Behavioral test and electroencephalography

The Korean sentence recognition test was conducted to evaluate the behavioral speech intelligibility score in a soundproof room prior to EEG data acquisition. The test used 10 sentences selected out of 90, and the participant was asked to verbally repeat the sentence which was presented through a loudspeaker (NS-B51, YAMAHA, Hamamatsu, Japan) at a comfortable level of 60 dBA. The behavioral speech test was performed using natural and noise-vocoded sentences. The behavioral speech intelligibility score, which was calculated

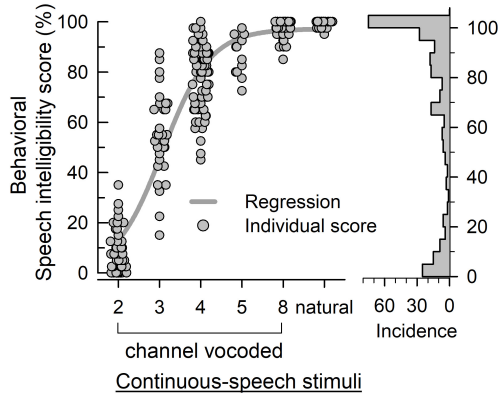

FIGURE 4
The behavioral speech intelligibility score (left panel) of individuals in response to vocoded and natural continuous-speech stimuli. The bars (right panel) indicate the incidence of each score.

using the number of correctly repeated words out of 40 target words, ranged from 0 to 100, with a step of 2.5.

EEG data were recorded using a 64-channel system (Biosemi Active 2, Netherlands) in a soundproof room. The natural and vocoded sentences were randomly played by a loudspeaker (NS-B51, YAMAHA, Hamamatsu, Japan) 1 m away from the participants. The inter-stimulus interval (between sentences) was set to 3 s, and each sentence was repeated 100 times. Difficult tasks had precedence over easy tasks to minimize learning throughout the tasks (**Figure 1**). During this passive experiment, a participant could watch a silent video with subtitles on an LCD monitor and could rest for 10 min between sessions. The raw EEG data were downsampled to 256 Hz for computational efficiency and preprocessed using the EEGLAB toolbox (Delorme and Makeig, 2004). The down-sampled EEG data were re-referenced using average referencing and band-pass (1–57 Hz) filtered by a Hamming windowed sinc FIR filter (Widmann, 2006). The typical eye-movement related artifact was rejected using the extended infomax independent component analysis and manually inspected correction. The EEG data were epoched in the intervals −0.5 to 2.5 s, relative to stimulus onset.

Speech features: Envelope, phoneme, and envelope and phoneme

The PH onset impulse train and the ENV of the natural sentences were used as speech (stimuli) features. All PH onsets in the sentences were automatically identified (Yoon and Kang, 2013) using Praat software (University of Amsterdam, Netherlands) and manually confirmed. The number of phonemes in each sentence ranged from 17 to 22 (mean: 18.6, standard deviation: 3.9). The PH onset impulse train consisted of a sequence of unit impulses at the onset time of the phoneme. The ENVs were computed using a full-wave rectifier and a low-pass filter (30 Hz cutoff). The cutoff frequency of 30 Hz was chosen to obtain a sufficient amplitude envelope of EEG data (Souza and Rosen, 2009; Roberts et al., 2011). Using these aforementioned values, the product of the multiplication of the PH and ENV (PHENV) was calculated.

TABLE 2 Results of behavioral speech intelligibility scores with natural and noise vocoded sentences.

Sentence type		Behavioral score (%)	
		Mean	*SD*
Noise vocoded	2 Channel	7.5	8.08
	3 Channel	55.2	17.03
	4 Channel	77.3	12.63
	5 Channel	86.4	8.01
	8 Channel	97.9	3.44
	Natural sentence	99.6	1.01

SD, standard deviation.

TABLE 3 Comparison of the performance of deep learning models using event-related potentials (ERP), stimuli envelopes (ENV), phonemes (PH), and phoneme-envelopes (PHENV).

	Deep learning using			
	ERP	ENV	PH	PHENV
Accuracy	97.33%	99.42%	99.55%	99.91%

Deep learning for speech intelligibility prediction

The EEG data were randomly split into a training set (80% of the original data set) and an unseen test set (20%), as depicted in **Figure 2**. An ERP was computed by averaging 80 EEG data epochs. A bootstrap sampling procedure was employed to generate ERPs and continuous speech-evoked potentials (CSEPs), evenly distributed across the range of speech intelligibility scores. The deep learning features of the CSEPs were computed by averaging the cross-correlation coefficients between the EEG data epochs and speech (stimuli) features. As a result, 800 ERPs and CSEPs from the training set, and 200 ERPs and CSEPs from the test set, were taken.

To enlarge the number of training datasets and guarantee that they reached 8,000, the training datasets were augmented with one of three approaches: Gaussian noise, temporal cutout, or sensor dropout. One of the approaches was randomly selected for each augmentation (Wang et al., 2018). The augmentation techniques used the best parameters obtained in a previous

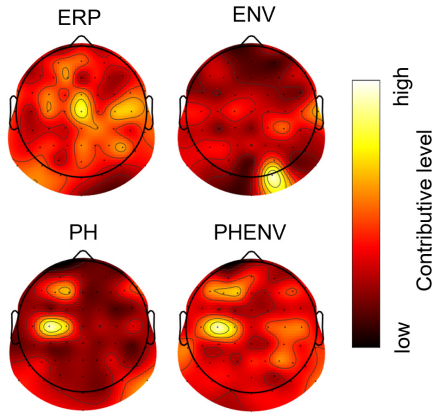

FIGURE 5
The topographic map of occlusion sensitivity visualizing the important brain regions for classification. Dark to bright red color denotes relatively low to high contributive levels for prediction.

TABLE 4 Summary of regions of significant contribution for deep learning. The ten most significant EEG channels and their regions, to predict speech intelligibility, are selected using occlusion sensitivity analysis.

Deep learning using	Regions of significant contribution for deep learning	
	EEG channels	Brain regions
ERP	Cz, C6, C4, FCz, P6, F1, CP2, C3, Pz, FC3	Central, frontal, parietal
ENV	O2, T8, PO4, CP4, C6, C1, T7, CP2, FT8, P5	Occipital, temporal, parietal
PH	C3, C1, F1, P10, F3, F8, P9, C5, F6, PO8	Central, frontal, parietal
PHENV	C3, C1, F1, F3, F5, P4, C5, TP7, C6, C4	Central, frontal, parietal

EEG, electroencephalography; ERP, event-related potential; ENV, stimuli envelope; PH, phoneme; PHENV, phoneme-envelope.

study (Cheng et al., 2020). Gaussian noise was added to the signal, and the ratio of noise to signal was 0.6. The temporal cutout was a random temporal window replaced with Gaussian noise, and the duration of the temporal window was 0.625 s (about 20% of the 3 s recording period). The sensor dropout was a random subset of sensors replaced with zeros, and the number of dropping sensors was 12 (about 20% of the 64 electrodes). The validation datasets (20% of the augmented training datasets) were randomly selected.

Figure 3 shows the overall schematic representation of speech feature, feature extraction, and speech intelligibility classification. Four deep learning models to predict the behavioral speech intelligibility scores were trained using the ERPs, envelope-based CSEPs, phoneme-based CSEPs, and phoneme-envelope-based CSEPs. The ERP and CSEP at each channel were plotted against time after sentence onset, as seen in Figure 3. The color in each panel indicated the amplitude of the ERP and CSEP. Each panel was resized to 299 × 299 from the original size of 64 × 768 for computational efficiency, to build up the model with small kernels and numbers of layers, and then used for deep learning (Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Tan and Le, 2019). The deep learning architecture consisted of four convolutional layers which were fully connected. Max pooling, leakyReLU, and batch normalization layers were employed in the convolution process. See Table 1 for more details about the deep learning architecture. The Adam optimizer was used for training the deep learning models (Kingma and Ba, 2014). The initial learning rates of the optimizer, batch size, and epoch value were set to 0.001, 64, and 5, respectively. Training data were shuffled before training to avoid any bias and overfitting. Finally, four deep learning models were evaluated by computing the classification accuracy, using the unseen test set as follows:

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP).$$

where *TP*, *TN*, *FP*, and *FN* denote true positive, true negative, false positive, and false negative, respectively. The occlusion sensitivity maps showed which channels contributed more to classifying the speech intelligibility score. Compromising spatial resolution and computational efficiency, the map used a 5 × 5 occluding mask and stride.

Results

Figure 4 plots the behavioral speech intelligibility scores in response to natural and noise-vocoded (2, 3, 4, 5, and 8 channel) sentences. Although the scores are not evenly distributed, it covered overall score ranges except 30.0, 37.5, and 40%. Table 2 shows the statistical summary of the behavioral scores.

Table 3 summarizes the performance of the deep learning models. The predictive accuracies were 97.33% (ERP), 99.42% (ENV), 99.55% (PH), and 99.91% (PHENV). Compared to the probabilistic chance level of 2.63%, the four deep learning models achieved comparable performance on predicting the speech intelligibility score. The deep learning model with the feature based on the PHENV yielded the highest accuracy of 99.91%.

Figure 5 shows the topographical map of occlusion sensitivity computed from the four deep learning models. The color indicates the level of contribution to a classification decision at each channel. Here, the dominant contribution was observed in the occipital region for the ENV-based model, whereas the dominancy was spread over the central, frontal, and parietal brain regions for the ERP, the PH-based model, and the PHENV-based model. Table 4 summarizes the 10 most sensitive EEG channels and the corresponding brain regions for deep learning to predict speech intelligibility scores.

Conclusion and discussion

In this study, we developed a deep learning model to predict speech intelligibility scores using continuous speech-evoked EEG signals. The cross-correlation coefficients between typical speech features (PH and ENV) and EEG responses to speech were implemented as a feature for deep learning and the model achieved the highest classification accuracy of 99.91%. The topographic map illustrating the frontal, central, and parietal regions provided important information for the classification.

Several studies have employed a linear model (i.e., TRF) to predict individual speech intelligibility from EEG responses to overlapped sentences or long story (14 min) stimuli

(Lesenfants et al., 2019; Muncke et al., 2022). One issue with stimulus-driven EEG signals is the speech onset response, which is greater in magnitude than the overall neural activity. For this reason, Crosse et al. (2021) reported that the TRF model may not be a feasible approach to apply in EEG signals in response to short-duration (<5 s) stimuli (Crosse et al., 2021). It is therefore essential to consider the methodological approach to model building in response to speech and continuous stimuli. It should also be noted that the TRF model requires regularization coefficient tuning to avoid overfitting, which makes use of more computational resources and is more complex than cross-correlation. Furthermore, a deep learning model with the cross-correlation coefficient can leverage a non-linear feature to predict the non-linear property of speech intelligibility (Accou et al., 2021).

Subjects participated in the passive listening condition during the electrophysiological data collection in this study. Passive listening provides less experimental fatigue than active listening and can be performed by young children (Roy, 2018; O'Neill et al., 2019). Several studies on selective attention decoding and cortical tracking to long story stimuli have employed the active listening paradigm to keep subjects attentive (Vanthornhout et al., 2018; Lesenfants et al., 2019; Accou et al., 2021; Nogueira and Dolhopiatenko, 2022). Although these participants were asked regarding the stimuli during the experiment for active listening, it may be difficult to ensure a stable attentive level throughout the entire task. In particular, Kong et al. (2014) reported that neural responses from active and passive listeners were similar in quiet conditions, whereas the differences of cross-correlation function were observed in competing speaker conditions. Thus, attention should be considered when predicting speech intelligibility in a selective listening condition.

The occlusion sensitivity enabled the decision of deep learning interpretability (Zeiler and Fergus, 2014; Ieracitano et al., 2021). Here, occlusion sensitivity explained that neural activity from the central and left frontal region made the most important contribution to speech understanding. The topographic map of occlusion sensitivity in PH and PHENV cases showed that the language dominant region (typically F3 within the middle frontal gyrus and TP7 within the middle temporal gyrus) was highly involved in speech intelligibility processes (Scrivener and Reader, 2022). The results are comparable with the findings of neuroimaging studies, specifically that of the sentence-processing network, including the middle frontal and middle temporal gyri (Peelle et al., 2004, 2010; Fiebach et al., 2005; Smirnov et al., 2014). Also, it supports the middle temporal gyrus and the supramarginal gyrus involvement in syntactic and phonological processing (Friederici, 2011). Deep learning with PH and PHENV could be reasonably explainable and interpretable by occlusion sensitivity.

This study has several limitations for clinical implementation. The deep learning model was developed using data from a limited group. The noise-vocoder was used to simulate hearing impairment with normal hearing listeners. Since no data from cochlear implant and hearing aid users were accessed, the model should be sufficiently validated with data of hearing-impaired individuals. In addition, although the group-level deep learning model was developed and tested in this study, it was still challenging to optimize the model with individual-level features due to inter-subject variability (Cheng et al., 2020; Accou et al., 2021). Further investigation of subject-specific models is necessary for the clinical prediction of speech intelligibility. These are the key issues for future studies. We also plan to improve the model by incorporating source EEG data rather than 64-channel EEG data and optimizing the channels based on the occlusion sensitivity map.

Data availability statement

The data used to support the findings of this study are available from the corresponding author upon request.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of the University of Ulsan. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YN and JW designed the experiment, developed the model, and examined the results. YN, LT, HJ, and LQ collected data and performed data preprocessing. All authors were involved in preparing the manuscript.

Funding

This work was supported by grants from the National Research Foundation of Korea (NRF-2020R1A2C2003319).

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

References

Accou, B., Monesi, M. J., Hamme, H. V., and Francart, T. (2021). Predicting speech intelligibility from EEG in a non-linear classification paradigm. *J. Neural Eng.* 18:066008. doi: 10.1088/1741-2552/ac33e9

Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S. U., Altuwaijri, G. A., Abdul, W., et al. (2021). Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Comput. Appl.* 1–42. doi: 10.1007/s00521-021-06352-5

Aminu, M., Ahmad, N. A., and Noor, M. H. M. (2021). Covid-19 detection via deep neural network and occlusion sensitivity maps. *Alexandria Eng. J.* 60, 4829–4855. doi: 10.1016/j.aej.2021.03.052

Cheng, J. Y., Goh, H., Dogrusoz, K., Tuzel, O., and Azemi, E. (2020). Subject-aware contrastive learning for biosignals. *arXiv* [Preprint]. doi: 10.48550/arXiv.2007.04871

Cherloo, M. N., Amiri, H. K., and Daliri, M. R. (2021). Ensemble Regularized Common Spatio-Spectral Pattern (ensemble RCSSP) model for motor imagery-based EEG signal classification. *Comput. Biol. Med.* 135:104546. doi: 10.1016/j.compbiomed.2021.104546

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., et al. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Sci. Rep.* 9:11538. doi: 10.1038/s41598-019-47795-0

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *J. Neural Eng.* 16:031001. doi: 10.1088/1741-2552/ab0ab5

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10:604. doi: 10.3389/fnhum.2016.00604

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., and Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Front. Neurosci.* 15:705621. doi: 10.3389/fnins.2021.705621

Delorme, A., and Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Di Liberto, G. M., Lalor, E. C., and Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage* 166, 247–258. doi: 10.1016/j.neuroimage.2017.10.066

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Ding, N., and Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* 33, 5728–5735. doi: 10.1523/JNEUROSCI.5297-12.2013

Ding, N., and Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Front. Hum. Neurosci.* 8:311. doi: 10.3389/fnhum.2014.00311

Enderby, P. (1980). Frenchay dysarthria assessment. *Int. J. Lang. Commun. Disord.* 15, 165–173. doi: 10.3109/13682828009112541

Esmaeilzadeh, S., Belivanis, D. I., Pohl, K. M., and Adeli, E. (2018). "End-to-end Alzheimer's disease diagnosis and biomarker identification," in *Machine Learning in Medical Imaging. MLMI 2018 Lecture Notes in Computer Science*, eds Y. Shi, H. I. Suk, M. Liu, (Cham: Springer)11046, 337–345. doi: 10.1007/978-3-030-00919-9_39

Fiebach, C. J., Schlesewsky, M., Lohmann, G., Von Cramon, D. Y., and Friederici, A. D. (2005). Revisiting the role of Broca's area in sentence processing: Syntactic integration versus syntactic working memory. *Hum. Brain Mapp.* 24, 79–91. doi: 10.1002/hbm.20070

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiol. Rev.* 91, 1357–1392. doi: 10.1152/physrev.00006.2011

Govindarajan, S., and Swaminathan, R. (2021). Differentiation of COVID-19 conditions in planar chest radiographs using optimized convolutional neural networks. *Appl. Intell.* 51, 2764–2775. doi: 10.1007/s10489-020-01941-8

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas: IEEE)770–778. doi: 10.1109/CVPR.2016.90

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (2015). An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *J. Acoust. Soc. Am.* 138, 1660–1669. doi: 10.1121/1.4929493

Ieracitano, C., Mammone, N., Hussain, A., and Morabito, F. C. (2021). A novel explainable machine learning approach for EEG-based brain-computer interface systems. *Neural Comput. Appl.* 34, 11347–11360. doi: 10.1007/S00521-020-05624-W

Jang, H., Lee, J., Lim, D., Lee, K., Jeon, A., and Jung, E. (2008). Development of Korean standard sentence lists for sentence recognition tests. *Audiol* 4, 161–177. doi: 10.21848/audiol.2008.4.2.161

Jonas, S., Rossetti, A. O., Oddo, M., Jenni, S., Favaro, P., and Zubler, F. (2019). EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Hum. Brain Mapp.* 40, 4606–4617. doi: 10.1002/hbm.24724

Kent, R. D., Weismer, G., Kent, J. F., and Rosenbek, J. C. (1989). Toward Phonetic Intelligibility Testing in Dysarthria. *J. Speech Hear. Disord.* 54, 482–499. doi: 10.1044/jshd.5404.482

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.* 126, 1486–1494. doi: 10.1121/1.3184603

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv* [Preprint]. doi: 10.48550/arXiv.1412.6980

Koelewijn, T., Zekveld, A. A., Lunner, T., and Kramer, S. E. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hear. Res.* 367, 106–112. doi: 10.1016/j.heares.2018.07.011

Kong, Y. Y., Mullangi, A., and Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hear. Res.* 316, 73–81. doi: 10.1016/j.heares.2014.07.009

Lashgari, E., Liang, D., and Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *J. Neurosci. Methods* 346, 108885. doi: 10.1016/j.jneumeth.2020.108885

Launer, S., Zakis, J. A., and Moore, B. C. J. (2016). "Hearing Aid Signal Processing," in *Hearing Aids*, eds G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper (Cham: Springer International Publishing), 93-130. doi: 10.1007/978-3-319-33036-5_4.

Lee, J. (2016). Standardization of Korean speech audiometry. *Audiol. Speech Res.* 12, S7–S9. doi: 10.21848/asr.2016.12.S1.S7

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., and Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations. *Hear. Res.* 380, 1–9. doi: 10.1016/j.heares.2019.05.006

Li, Y., Yang, H., Li, J., Chen, D., and Du, M. (2020). EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM. *Neurocomputing* 415, 225–233. doi: 10.1016/j.neucom.2020.07.072

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Lombardi, A., Tavares, J. M. R. S., and Tangaro, S. (2021). Editorial: Explainable artificial intelligence (XAI) in systems neuroscience. *Front. Syst. Neurosci.* 15:766980. doi: 10.3389/fnsys.2021.766980

Macherey, O., Van Wieringen, A., Carlyon, R. P., Deeks, J. M., and Wouters, J. (2006). Asymmetric pulses in cochlear implants: Effects of pulse shape, polarity, and rate. *J. Assoc. Res. Otolaryngol.* 7, 253–266. doi: 10.1007/s10162-006-0040-0

Mansour, M., Khnaisser, F., and Partamian, H. (2020). An explainable model for EEG seizure detection based on connectivity features. *arXiv* [Preprint]. doi: 10.48550/arXiv.2009.12566

Mehta, A. H., and Oxenham, A. J. (2017). Vocoder simulations explain complex pitch perception limitations experienced by cochlear implant users. *J. Assoc. Res. Otolaryngol.* 18, 789–802. doi: 10.1007/s10162-017-0632-x

Muncke, J., Kuruvila, I., and Hoppe, U. (2022). Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise. *Front. Neurosci.* 835:876421. doi: 10.3389/fnins.2022.876421

Nogueira, W., Cosatti, G., Schierholz, I., Egger, M., Mirkovic, B., and Büchner, A. (2019). Toward decoding selective attention from single-trial EEG data in cochlear implant users. *IEEE Trans. Biomed. Eng.* 67, 38–49. doi: 10.1109/TBME.2019.2907638

Nogueira, W., and Dolhopiatenko, H. (2020). "Towards decoding selective attention from single-trial EEG data in cochlear implant users based on deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2020*, (Spain: IEEE), 8708–8712. doi: 10.1109/TBME.2019.2907638

Nogueira, W., and Dolhopiatenko, H. (2022). Predicting speech intelligibility from a selective attention decoding paradigm in cochlear implant users. *J. Neural Eng.* 19:026037. doi: 10.1088/1741-2552/ac599f

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *J. Neurosci.* 29, 15564–15574. doi: 10.1523/JNEUROSCI.3065-09.2009

Oh, Y., Park, S., and Ye, J. C. (2020). Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* 39, 2688–2700. doi: 10.1109/TMI.2020.2993291

O'Neill, E. R., Kreft, H. A., and Oxenham, A. J. (2019). Cognitive factors contribute to speech perception in cochlear-implant users and age-matched normal-hearing listeners under vocoded conditions. *J. Acoust. Soc. Am.* 146:195. doi: 10.1121/1.5116009

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Peelle, J. E., McMillan, C., Moore, P., Grossman, M., and Wingfield, A. (2004). Dissociable patterns of brain activity during comprehension of rapid and syntactically complex speech: Evidence from fMRI. *Brain Lang.* 91, 315–325. doi: 10.1016/j.bandl.2004.05.007

Peelle, J. E., Troiani, V., Wingfield, A., and Grossman, M. (2010). Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity. *Cereb. Cortex* 20, 773–782. doi: 10.1093/cercor/bhp142

Roberts, B., Summers, R. J., and Bailey, P. J. (2011). The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes. *Proc. R. Soc. B Biol. Sci.* 278, 1595–1600.

Roy, R. A. (2018). Auditory working memory: A comparison study in adults with normal hearing and mild to moderate hearing loss. *Glob. J. Otolaryngol.* 13, 1–14. doi: 10.19080/GJO.2018.13.555862

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *J. Neural Eng.* 16:051001. doi: 10.1088/1741-2552/ab260c

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123, 2400–2406. doi: 10.1093/brain/123.12.2400

Scrivener, C. L., and Reader, A. T. (2022). Variability of EEG electrode positions and their underlying brain regions: Visualizing gel artifacts from a simultaneous EEG-fMRI dataset. *Brain Behav.* 12:e2476. doi: 10.1002/brb3.2476

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. doi: 10.48550/arXiv.1409.1556

Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *J. Imaging* 6:52. doi: 10.3390/jimaging6060052

Smirnov, D., Glerean, E., Lahnakoski, J. M., Salmi, J., Jääskeläinen, I. P., Sams, M., et al. (2014). Fronto-parietal network supports context-dependent speech comprehension. *Neuropsychologia* 63, 293–303. doi: 10.1016/j.neuropsychologia.2014.09.007

Snedeker, J., and Huang, Y. T. (2009). "Sentence processing," in *The Cambridge handbook of child language*, ed. L. Bavin (Cambridge: Cambridge University Press), 321–337.

Souza, P., and Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. *J. Acoust. Soc. Am.* 126, 792–805. doi: 10.1121/1.3158835

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition*, 27–30 June 2016, Las Vegas, NV, 1–9. doi: 10.1109/CVPR.2015.7298594

Tan, M., and Le, Q. (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks,"in *Proceedings of the 36th International Conference on Machine Learning* (Long Beach: IEEE) 97, 6105–6114.

Uyttenhove, T., Maes, A., Van Steenkiste, T., Deschrijver, D., and Dhaene, T. (2020). "Interpretable epilepsy detection in routine, interictal EEG data using deep learning," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, eds E. Alsentzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, (Belgium: Ghent University)136, 355–366.

Vandali, A. E., Whitford, L. A., Plant, K. L., and Clark, G. M. (2000). Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system. *Ear Hear.* 21, 608–624. doi: 10.1097/00003446-200012000-00008

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *J. Assoc. Res. Otolaryngol.* 19, 181–191. doi: 10.1101/246660

Wang, F., Zhong, S., Peng, J., Jiang, J., and Liu, Y. (2018). "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks," in *International Conference on Multimedia Modeling*, eds A. Elgammal, T. H. Chalidabhongse, S. Aramvith, Y-S. Ho, K. Schoeffmann, C. W. Ngo, N. E. O'Connor, and M. Gabbouj,(Cham: Springer), 82–93. doi: 10.1155/2021/2520394

Wanna, G. B., Noble, J. H., Carlson, M. L., Gifford, R. H., Dietrich, M. S., Haynes, D. S., et al. (2014). Impact of electrode design and surgical approach on scalar location and cochlear implant outcomes. *Laryngoscope* 124, S1–S7. doi: 10.1002/lary.24728

Widmann, A. (2006). *Firfilt EEGLAB Plugin, Version 1.5. 1*. Leipzig: Leipzig University.

Wouters, J., McDermott, H. J., and Francart, T. (2015). Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Process Mag.* 32, 67–80.

Yoon, T., and Kang, Y. (2013). *The Korean phonetic aligner program suite.* Available online at: http://korean.utsc.utoronto.ca/kpa/ (accessed August 1, 2022).

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer), doi: 10.1007/978-3-319-10590-1_53

Zhou, Y., Huang, S., Xu, Z., Wang, P., Wu, X., and Zhang, D. (2021). *Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review*. Cham: IEEE.

Zhu, X., and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev.* 22, 177–210. doi: 10.1007/s10462-004-0751-8