A Method To Map EEG Signals To Spoken Speech Using Gaussian Process Modeling

Hongde Wu, Changjie Pan, Mingtao Li, Fei Chen
University Key Laboratory of Advanced Wireless Communications of Guangdong Province
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen 518055, China
HDWuNg@outlook.com, 11930565@mail.sustech.edu.cn, limt@mail.sustech.edu.cn, fchen@sustech.edu.cn

***Abstract*—This work aimed to map electroencephalography (EEG) signals recorded during speech production to an intelligible speech. Experiments were designed to record EEG and spoken speech signals from normal participants. EEG features were processed with a Gaussian process regression method, and used to estimate multiple temporal amplitude envelopes of a spoken speech signal. The estimated envelopes were further applied to synthesize an intelligible speech signal by using a temporal envelope-based vocoder model. The performance of reconstructing the spoken speech signal was evaluated by the short-term objective intelligibility (STOI) index and the root mean square error (RMSE) between the reconstructed vocoded speech and the original spoken speech. Results showed a small RMSE between two sets of mel-frequency cepstral coefficients, and a STOI measurement up to 0.71. Both measures outperformed results from existing studies with similar tasks, indicating the potential in synthesizing an intelligible spoken speech with EEG signals in brain-computer interface based speech communication.**

Keywords—electroencephalography, speech synthesis, brain computer interface, stochastic analysis

I. Introduction

Speech communication plays an important role in daily activities in human society. Unfortunately, a huge number of people are suffering speech disorders worldwide [1]. To create a new communication approach for people with communication disorder, recent studies investigated the speech-based brain computer interface (BCI) [e.g., 2] and have achieved promising results. For instance, neural cues of speech prosody in receptive [3] and productive [4] speech cortices have been interpreted. Deep neural networks have been applied to BCIs [e.g., 5-6], and studies were able to reconstruct a perceived speech [7] and to reconstruct spectral dynamics of speech from electrocorticography (ECoG) [8]. More recently, speech waveforms of ten digits were directly reconstructed from the corresponding listening ECoG by a fully connected network [9]. Anumanchipalli et al. encoded and decoded the spoken ECoG using recurrent neural network, and transformed cortex activities into articulatory movements and intelligible spoken sentences [10]. Angrick et al. reconstructed speech of different words with 3D convolutional neural network [12]. Since ECoG is an invasive approach of monitoring cortex activities, electroencephalography (EEG) as an alternative has emerged in speech-based BCI applications. Spoken speech and listening utterances were synthesized from the corresponding EEG recordings [e.g., 11, 13]. For the purpose of studying a BCI-based speech communication approach, this work aimed to map EEG signals during speech production to the corresponding speech signal.

Early studies showed that the temporal amplitude envelops of a speech signal could be tracked from ECoG [14]. The temporal amplitude envelopes from 4 frequency bands of a speech signal could be used to synthesize an intelligible speech [15], because temporal envelope (i.e., a slow-changing waveform of speech amplitude variation) carries important perceptual cues for speech perception. Meanwhile, Gaussian process has been proved to be a successful tool in biomedical signal modeling [e.g., 16-17], gaining its high generalization ability from small-sample and non-linear training [18]. In our previous work, Gaussian process regression (GPR) showed the potential in speech reconstruction for English syllable and words using EEG signals during speech imagery in a public dataset [19].

However, there are different mechanisms in speech comprehension among different languages [e.g., 20-21]. As a tone language, Chinese has four tones for each syllable and different tones of Chinese syllables define specific meaning in different contexts. Following the research purpose of this work, this study used EEG signals of spoken speech and Gaussian process regression to decode multi-band temporal envelopes for Chinese speech synthesis.

II. Experiment And Method

A. Data collection

Seven male and four female native Mandarin-Chinese speakers participated in this experiment. All participants were between 20 and 24 years old (mean age 22 years) and with normal abilities in language speech and hearing. EEG signals were recorded by a 64-channel Neuroscan Quick-cap and sampled at a rate of 1000 Hz, where the electrode locations were determined by the 10-20 system [22]. The experiment was approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

This work used 18 Mandarin monosyllables as stimulus materials, which were /a/, /ba/, /bi/, /bu/, /fa/, /fu/, /ji/, /ju/, /la/, /li/, /lu/, /lv/, /ma/, /mi/, /mu/, /yi/, /wu/ and /yu/. These monosyllables were selected to involve different consonant-vowel combinations in Mandarin Chinese, derived from Mandarin Speech Perception Test corpus [23]. Each monosyllable had four Mandarin tones (except the first tone for /mu/ and /lv/) and each tonal monosyllable was randomly repeated for 5 times, resulting in a total of 350 trials for each participant. Data from 4 of the 11 participants were removed because the participants were not concentrated on the tasks or their spoken speech signals were not recorded in the experiment.

There were five stages for each trial in the following sequence: (1) a 3-sec rest state, where participants cleared their mind; (2) a 1.5-sec listening state, where the audio waveform of the tonal monosyllable was played by a loudspeaker; (3) a 2-sec imagined state, where participants imagined to speak the utterance presented; (4) a 2.5-sec intended state, where participants read the monosyllable silently without audible output; and (5) a 2.5-sec speaking state, where participants spoke the monosyllable aloud. The spoken speech signal was recorded at a 16 kHz sampling rate. EEG signals were recorded simultaneously in each stage. In

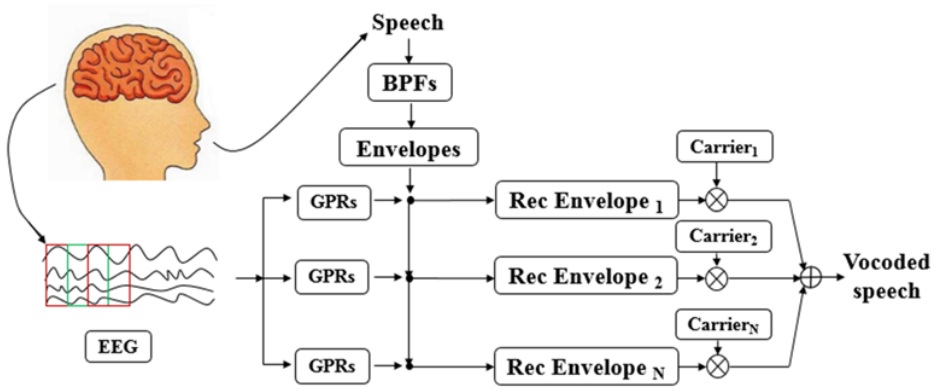

Fig. 2. Overall framework of the vocoder-GPR based speech synthesis approach. Each box on the EEG signal represents a window with length of 10% of the EEG signal, and two neighboring boxes are with 50% overlapping.

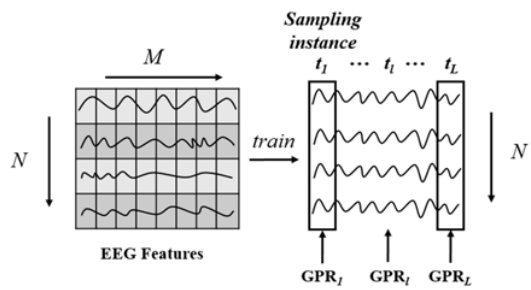

Fig. 1. GPR modeling process for predicting each temporal envelope. GPR models are trained at each sampling instance t separately. There are L GPR models, and the EEG feature set (with size of N × M) is used for L times, where L is the number of time steps in speech sampling. In total, there are L × 24 models in this work.

this work, only the EEG signals from the speaking state were used for the speech reconstruction task for monosyllables.

B. Speech and EEG signal pre-processing

Silent parts in speech recordings were eliminated by voice active detection (VAD) algorithm, as this work focused on the relation between speech information and EEG signals. Also, those parts of EEG signals corresponding to the silent parts of speech were eliminated to maintain the consistency and simultaneity between spoken speech and EEG signals. Finally, different speech waveforms were normalized in power spectrum.

For pre-processing the EEG signals, EEGLAB [24] was applied to remove artifacts (e.g., electrocardiography and electromyography) from the original EEG signals using independent component analysis. Signals between 1 Hz and 30 Hz were remained by band-pass filtering. This work segmented the EEG data with windows with length of 10% of an EEG signal, and with 50% hop between two adjacent windows (see the boxes on the EEG signal in Fig. 1). Thus, there were 19 windows for each channel of EEG signal. For feature extraction, the mean values and their first and second order differentials were calculated as EEG features within each window. By selecting the mean values and differentials by windows, EEG features became closer to Gaussian distribution and carried temporal-changing cues [12]. For EEG channel selection, 10 channels (i.e., FC6, FT8, C5, CP3, P3, T7, CP5, C3, CP1 and C4) were finalized since EEG signals from these channels were reported to have high correlation with the corresponding speech recordings [25]. In order to combine the information from these 10 channels and maintain the high correlation between EEG features and speech recordings, EEG features were averaged among the selected 10 channels, resulting in 57 (=19 × 3) features or a 1 × 57 feature vector for each EEG recording.

C. Multi-band vocoder for speech synthesis

The perceptual contributions of temporal envelope have been revealed by the study of vocoder modeling in speech analysis [15, 26]. The input speech is first processed by a set of band pass filters (BPFs). For the band-pass filtered signal at each frequency band, waveform rectification and low-pass filtering are used to generate the temporal amplitude envelope, which carries important perceptual information [15]. In order to synthesize an intelligible vocoded speech, carrier signals (e.g., pure tone or white noise signal) are modulated by these envelopes and all modulated carrier signals are summed up. When there are more than 4 frequency bands in the vocoding process [15], the vocoded speech is with sufficient intelligibility. In this work, the number of frequency bands is selected as 24, indicating that there are 24 temporal envelops for each speech signal.

D. Gaussian process regression modeling

The proposed approach for speech synthesis is based on the vocoder-GPR framework, as shown in Fig. 1. In order to model the dependencies between the speech envelopes and EEG signals, at each sampling instance, covariance functions are used to represent the difference and dependencies among all EEG signals from all trials. The different speech recordings of the same material (i.e., reading the same monosyllable) follow Gaussian distributions. Thus, for each speech envelope, we model the Gaussian distribution at each sampling instance, as shown in Fig. 2.

Here EEG features are denoted as E and each speech envelope is denoted as S, as:

$$E = \begin{bmatrix} E_1 \\ \vdots \\ E_N \end{bmatrix} = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1M} \\ \vdots & \cdots & \cdots & \vdots \\ E_{N1} & E_{N2} & \cdots & E_{NM} \end{bmatrix}, \quad (1)$$

and

$$S = \begin{bmatrix} S_1 \\ \vdots \\ S_N \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1L} \\ \vdots & \cdots & \cdots & \cdots \\ S_{N1} & S_{N2} & \cdots & S_{NL} \end{bmatrix}, \quad (2)$$

where N is the number of EEG (or speech) trials in the training set, M is the dimension of EEG features, and L is the number of sampling instances (i.e., the length) of the speech envelope, which is also the length of the speech signal. In this work, $N = 280$, $M = 57$, and we took $L = 14400$ (i.e., equal to 0.9 second under 16 kHz sampling rate for speech recording), as the VAD algorithm determined that the active parts for monosyllable speech lasted for around 0.9 second in this experiment.

At each sampling instance l (= 1, 2, ···, L), the speech envelope is modeled as a multivariate Gaussian distribution, and represented as:

$$[S_{1l} \cdots S_{Nl}]^T \sim G(\mu, K), \quad (3)$$

where $G(\cdot)$ denotes the multivariate Gaussian distribution with μ as a vector consisted of the mean values, and K as the covariance matrix with element K_{ij} determined by kernel $k(E_i, E_j)$. Here $k(\cdot)$ is a radial basis function (i.e., RBF kernel), as:

$$k(E_i, E_j) = \alpha^2 \exp\left(-\frac{\|E_i - E_j\|}{2\gamma^2}\right), \quad (4)$$

where α and γ are parameters which determine the hyperparameter $\vartheta = [\alpha, \gamma]$, calculated by all EEG feature vector pairs E_i and E_j.

Given the above mathematic modeling, for a known EEG features E_*, envelope S_* can be predicted by Gaussian likelihood, as:

$$p(S_*|E_*, E, S, \vartheta, l) = N_l(k_*^T K^{-1} S, \kappa - k_*^T K^{-1} k_*), \quad (5)$$

where

$$k_* = K(E_*, E) = [k(E_*, E_1), \ldots, k(E_*, E_N)], \quad (6)$$

and

$$\kappa = k(E_*, E_*). \quad (7)$$

Here the hyperparameter ϑ is optimized by maximizing the conditional probability $p(E|S, \vartheta)$ from the training set $[E, S]$. This work took conjugate gradient descent on the logarithm of $p(S|E, \vartheta)$ for the purpose of optimization, as:

$$log\, p(S|E, \vartheta) = -\frac{N}{2} log 2\pi - \frac{1}{2} S^T K^{-1} S - \frac{1}{2} \log|K|. \quad (8)$$

E. Vocoder-GPR model training

Speech envelopes were generated from the original speech waveform by firstly applying a 6th-order band-pass Butterworth filter between 80 Hz and 6000 Hz, decomposing the speech signal into 24 frequency bands based on frequency-position mapping function [27]. Secondly, full-wave rectification and 2nd-order low-pass filter with a cut-off frequency of 200 Hz were applied to each frequency band to extract the raw speech envelopes, which were used for the GPR model training for envelope reconstruction. After envelopes were reconstructed, all amplitude-modulated carrier signals (i.e., sinusoidal waveform in this work) modulated by reconstructed envelopes were summed up to generate the vocoded speech signal. Finally, the vocoded speech signal was normalized in terms of root-mean-square power, in order to keep the magnitude as the level of the original speech recording. GPR models were trained using the GPML toolbox [18] on Matlab. Table 1 describes the initial parameters of GPR models. Different models were trained at different sampling instance; therefore EEG features were used for L times for each envelope and there were L × 24 (frequency bands) models for a given EEG-speech pair.

The training processes were conducted separately for each participant in this experiment. Using the data from speaking state only, for each participant, 80% (i.e., 280 trials) of the shuffled data served as the training set and the rest 20% (i.e., 70 trials) as the testing set; then a five-fold cross validation was implemented. The speech synthesis was only conducted for 7 participants, since data from 4 of the 11 participants were removed as the participants did not successfully finish the experimental tasks while collecting EEG signals.

TABLE I. PARAMETERS INITIALIZATION FOR GPR SETTING.

Mean	Covariance	Likelihood	Optimization
0	[0, -1]	-1	Polack-Ribiere

TABLE II. COMPARISON FOR SPEECH RECONSTRUCTION WITH [11] AND [12] ON RMSE AND STOI MEASURES.

	RMSE	STOI
Krishna, *et. al* [11]	4.86	N/A
Angrick, *et. al* [12]	N/A	0.33±0.14
Participants 1	1.87	0.62
Participants 2	1.82	0.63
Participants 3	2.39	0.59
Participants 4	2.40	0.62
Participants 5	2.51	0.57
Participants 6	2.67	0.71
Participants 7	2.72	0.67
Average	**2.34±0.36**	**0.63±0.04**

III. RESULTS

For evaluating the performance of synthesizing spoken speech, this work computed the root mean square error (RMSE) between the mel-frequency cepstral coefficients (MFCCs) of the original and reconstructed vocoded speech signals, as:

$$RMSE(x, y) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - y_i)^2}, \quad (9)$$

where x and y are respectively the MFCCs of the original speech signal and the reconstructed vocoded speech signal, and m is the dimension of MFCCs.

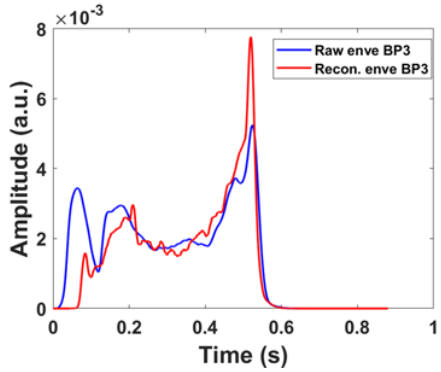

Fig. 3. The reconstructed (red) and raw envelopes (blue) from band 3 of a spoken speech signal.

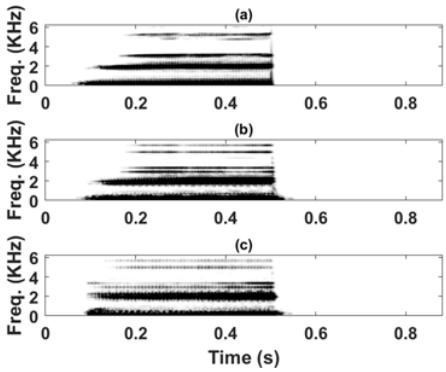

Fig. 4. Spectrograms of (a) the raw speech, (b) the vocoded raw speech and (c) the reconstructed vocoded speech.

Also, the short-time objective intelligibility (STOI) measure [28] between the original and reconstructed vocoded speech signals was calculated. As shown in Table 2, we compared our speech synthesis results with those of Krishna et al. [11] and Angrick et al. [12], both of which reconstructed speech from speaking-state cortex activity.

Krishna et al. [11] reconstructed the MFCCs from the spoken EEG signals and calculated the RMSE between the ground truth and the predicted MFCCs in 13 dimensions. This work extracted MFCC features with the same dimension (i.e., 13) for each participant and each trial for both the original and reconstructed speech, and calculated the averaged RMSE for each participant. The averaged RMSE result in this experiment was 2.34 for 7 participants, with the lowest RMSE value of 1.82 for participant 2. Our results outperformed the results in [11], which had a lowest RMSE value of 4.86.

In the STOI measurement, the experiment in [12] instructed participants to read the words continuously and also recorded the spoken speech continuously, which resulted in a long speech waveform. Therefore, this work also concatenated all the reconstructed speech (i.e., from 70 trials) to generate a long speech waveform for the STOI evaluation and comparison. In this experiment, an averaged STOI value of 0.63 was obtained for 7 participants, with the highest value of 0.71. For comparison, Angrick et al. [12] achieved an overall STOI result of 0.33 for 6 participants.

Figure 3 shows an example of envelope reconstruction from band 3 of a spoken speech signal. It is seen that the reconstructed and raw envelopes are with high similarity. The spectrograms of the raw speech, the vocoded raw speech and the reconstructed vocoded speech are shown in Fig. 4. There is a frequency discontinuity in the high frequency range (between 3 kHz to 4 kHz) when comparing the spectrograms of the raw speech and the reconstructed vocoded speech, which is because the vocoding process decomposes the speech signal into limited (i.e., 24 in this work) frequency bands.

IV. Conclusion

This paper showed that spoken speech signals could be decoded from EEG-based neural recordings during speech production. A GPR based method was used to map EEG signals recorded during speech production onto multiple temporal amplitude envelopes of a speech signal, which were subsequently used to synthesize an intelligible speech signal with an envelope-based vocoder model. The speech synthesis results outperformed those from recent publications [e.g., 11, 12] on objective comparisons with STOI and MFCCs-based RMSE between the original and reconstructed vocoded speech signals.

Acknowledgements

This work was supported by Shenzhen Sustainable Support Program for High-level University (2020), and High-level University Fund of Southern University of Science and Technology.

References

[1] H. J. Hoffman, C. Li, K. Losonczy, M. Chiu, J. Lucas and St. Louis, "Voice, speech, and language disorders in the U.S. population: The 2012 National Health Interview Survey (NHIS)," *Annual Meeting of the Society for Epidemiologic Research*, 648, pp.156, 2014.

[2] J. Vidal, "Toward direct brain–computer communication," *Annual Review of Biophysics and Bioengineering*, 2, pp.157-180, 1973.

[3] C. Tang, L. Hamilton and E. Chang, "Intonational speech prosody encoding in the human auditory cortex," *Science*, 357, pp.797-801, 2017.

[4] B. Dichter, J. Breshears, M. Leonard and E. Chang, "The control of vocal pitch in human laryngeal motor cortex," *Cell*, 174, pp.21-31, 2018.

[5] M. Seonwoo, L. Byunghan and Y. Sungroh, "Deep Learning in Bioinformatics," arXiv: 1603.06430, 2016.

[6] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter and T. Ball, "Deep learning with convolutional neural networks for decoding and visualization of EEG pathology," *Human Brain Mapping*, 38(11), pp.5391-5420, 2017.

[7] B. Pasley, S. David, N. Mesgarani, A. Flinker, S. Shamma, N. Crone, R. Knight and E. Chang, "Reconstructing speech from human auditory cortex," *PLoS Biology*, 10(1), pp.1-13, 2012.

[8] S. Martin, P. Brunner, C. Holdgraf, H. Heinze, N. Crone, J. Rieger, G. Schalk, R. Knight and B. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, 7(14), pp.1-15, 2014.

[9] H. Akbari, B. Khalighinejad, J. Herrero, A. Mehta and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, 9(874), pp.1-12, 2019.

[10] G. Anumanchipalli, J. Chartier and E. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, 568(7758), pp.493-498, 2019.

[11] G. Krishna, C. Tran, Y. Han and M. Carnahan, "Speech synthesis using EEG," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1235-1238, 2020.

[12] M. Angrick, C. Herff, E. Mugler, M. Tate, M. Slutzky, D. Krusienski and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of Neural Engineering*, 16(3), pp.1-10, 2019.

[13] G. Krishna, Y. Han, C. Tran, M. Carnahan and A. Tewfik, "State-of-the-art speech recognition using EEG and towards decoding of speech spectrum from EEG," arXiv: 1908.05743v5, 2020.

[14] J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel and G. Schalk, "The tracking of speech envelope in the human cortex," *PLoS One*, 8(1):e53398, 2013.

[15] R. Shannon, F. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, 270(5234), pp.303-304, 1995.

[16] Y. Yun, H. Kim, S. Shin, J. Lee, A. Deshpande, and C. Kim, "Statistical method for prediction of gait kinematics with Gaussian process regression," *Journal of Biomechanics*, 47(1), pp.186-192, 2014.

[17] C. Glackin, C. Salge, M. Greaves, D. Polani, S. Slavnic, D. Durrant, L. Adrian and M. Zlatko, "Gait trajectory prediction using gaussian process ensembles," *International Conference on Humanoid Robots*, pp.628-633, 2014.

[18] C. Rasmussen and C. Williams, "Gaussian Processes for Machine Learning," *The MIT Press*, ISBN 0-262-18253-X, 2006.

[19] H. Wu and F. Chen, "A temporal envelope-based speech reconstruction approach with EEG signals during speech imagery," *in Proc. of 12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 894-899, 2020.

[20] D. Fogerty and F. Chen, "Vowel spectral contributions to English and Mandarin sentence intelligibility," *in Proc. of 15th Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 499-503, 2014.

[21] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *Journal of the Acoustical Society of America*, 129(5), pp. 3281-3290, 2011.

[22] F. Sharbrough, G. Chatrian, R. Lesser, H. Luders, M. Nuwer and T. Picton, "American electroencephalographic society guidelines for standard electrode position nomenclature," *Journal of Clinical Neurophysiology*, 8(2), pp.200-202, 1991.

[23] Q. Fu, M. Zhu and X. Wang, "Development and validation of the Mandarin speech perception test," *Journal of the Acoustical Society of America*, 129(6), pp.EL267-273, 2011.

[24] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, 134(1), pp.9-21, 2004.

[25] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.992-996, 2015.

[26] D. Xu, L. Wang and F. Chen, "An ERP study on the combined-stimulation advantage in vocoder simulations," *International Conference of the Engineering in Medicine & Biology Society (EMBC)*, pp.2442-2445, 2018.

[27] D. Greenwood, "A cochlear frequency-position function for several species-29 years later," *Journal of the Acoustical Society of America*, 87(6), pp.2592-2605, 1990.

[28] C. Taal, R. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *International Conference on Acoustics Speech & Signal Processing (ICASSP)*, pp.4214-4217, 2010.