A Temporal Envelope-based Speech Reconstruction Approach with EEG Signals during Speech Imagery

Hongde Wu and Fei Chen
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
E-mail: HDWuNg@outlook.com, fchen@sustech.edu.cn Tel: +86-755-88018554

Abstract— **This work studied a brain-computer interface (BCI) system for speech synthesis based on imagined electroencephalography (EEG). The system incorporated a vocoder decomposition layer, a Gaussian process regression (GPR) layer and a vocoder synthesis layer, and was evaluated with speech recordings and imagined EEG signals from a public dataset (i.e., KARAONE). The raw speech signals were decomposed into envelopes in 12 frequency bands. Imagined EEG features were projected to each speech envelope by GPR, then the projected envelopes were used for speech reconstruction. With a cross-subject evaluation scheme, the similarity between the raw and projected envelopes achieved an average normalized covariance of 0.57, and the short-term objective intelligibility measurement between the raw and reconstructed speech yielded an average value of 0.70. Results in this work suggested the potential in developing a BCI-based communication with intelligible speech reconstruction.**

Index Terms speech synthesis, brain computer interface, supervised learning

I. INTRODUCTION

Speech production is one of the most important abilities for human beings. However, there are many people living with speech disabilities around the world (e.g., about 4.0% of U.S. adults had problems in using their voice [1]). In order to help people with speech disabilities, the number of researches in developing speech brain-computer interface (BCI) system has grown [2].

In recent years, machine learning techniques and deep neural networks have been applied to BCIs [e.g., 3-4] and researches shed light on the feasibility of reconstructing an understandable speech from cortex activities. For instance, the speech waveforms of 10 digits were directly reconstructed from the corresponding listening electrocorticography (ECoG) by a fully connected network [5]. Anumanchipalli et al. encoded and decoded the spoken ECoG using recurrent neural networks, and transformed cortex activities into intelligible spoken sentences [6]. Angrick et al. reconstructed speech waveforms using ECoG signals recorded when participants spoke different words [24]. Krishna et al. synthesized the speech features from spoken electroencephalogramy (EEG), as a non-invasive alternative of ECoG [13-14]. Sun et al. reconstructed speech envelopes from imagined EEG [8].

EEG, as a convenient approach in recording neural responses of human brain activities, is emerging in speech-based BCI applications. Zhao et al. created an open database KARAONE and used classification approach for identifying phonological categories in imagined and silent speech [7]. Also, machine learning and traditional features from speech signal processing have made contributions to the imagined EEG based classification problems [10-12]. Saha et al. extracted features of imagined EEG using convolutional neural network [9]. Zhao et al. windowed the imagined EEG signals and calculated statistical features (e.g., mean and spectral entropy) in each window [7]. Temporal envelope (i.e., a slow-changing waveform of speech amplitude variation) carries important perceptual cues for speech synthesis and perception. Early studies showed that temporal envelopes from 4 frequency bands of speech signal could be used to synthesize a highly intelligible speech [15]. This motivated the present work to decode multi-band temporal envelope information from EEG signals for synthesizing an intelligible speech.

Utilizing the KARAONE database, there were many works related to EEG-based speech imagination. Saha et al. had the best result with accuracy of 28.08% in the 11-category classification task (i.e., there were 7 phonemes and 4 words in the KARAONE database) if only the imagined EEG signals were used [9]. Zhao et al. achieved the accuracy of 18.08% in the "vowel-only vs. consonant" classification task, which was slightly lower than the chance level (i.e., 2 phonemes belonging to the vowel-only category and the remaining 9 phonemes and words belonging to the consonant category, with a chance level of 18.18%) [7]. Sun et al. reconstructed the envelopes of original speech from imagined EEG data, but they did not address the cross-subject evaluation (i.e., training set and test set were both from the same subject), and the reconstructed results were not intelligible [8]. These results indicate the difficulty to address imagined EEG based speech classification and reconstruction when imagined EEG signals are used to represent the difference between among speech categories or participants. Recently, Bakhshali et al. obtained an averaged accuracy over 90% in the word-pair classification task (i.e., 6 pairs among /pat/, /pot/, /knew/ and /gnaw/) based on imagined EEG, and they built training models for the classification task for each participant separately [23]. This motivated the present work to conduct a cross-subject speech reconstruction study based on imagined EEG.

The aim of this work was to synthesize speech with the envelope information decoded from the corresponding imagined EEG signals in speech imagery. In this work, EEG data from 4 participants were used for training, and EEG data

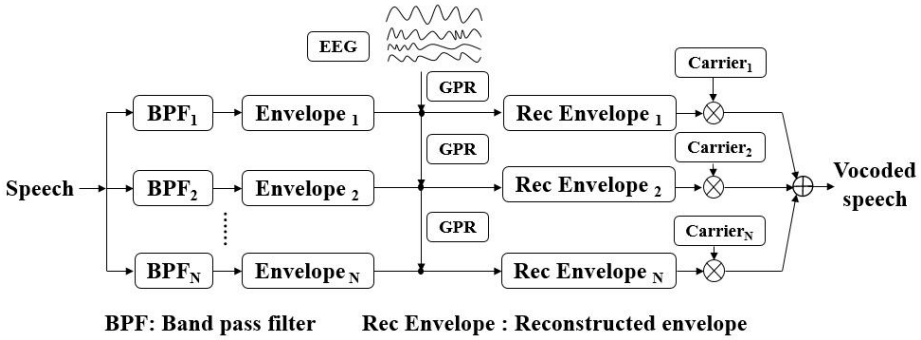

Fig. 1: Overall framework of the vocoder-GPR based speech synthesis approach.

from a new participant were for validation, whereas all participants imagined speech production of words. The speech and EEG signals were taken from KARAONE database [7]. EEG features were extracted in time windows, voice active detection (VAD) was applied to eliminate silent parts in speech recordings, and Gaussian process regression (GPR) was combined with a vocoder model for EEG-speech mapping.

II. METHODS

A. Vocoder based speech synthesis

Vocoder model has been long studied for speech synthesis and understanding the perceptual impacts of temporal envelope [16]. During the original vocoding process, a speech signal is divided into multiple frequency bands by an array of band-pass filters (BPFs), and the temporal envelope from each band is extracted by a wave rectification and a low-pass filter (LPF). The envelope signal is used to modulate the amplitude of a carrier signal (e.g., a pure tone with frequency equal to the center frequency of the band-pass filter corresponding to its frequency band [16]), and all amplitude-modulated carrier signals are summed up to generate the temporal envelope-based vocoded speech, which carries sufficient speech intelligibility information.

B. Gaussian process regression modeling

Figure 1 shows the overall framework of the proposed vocoder-GPR based speech synthesis approach. An important part is to obtain the mapping relationship between the imagined EEG features and speech envelope in each sub-band. Gaussian process, denoted as $G(\cdot)$, is defined by a mean function and a covariance function, as:

$$G(m(x),k(x,y)), \tag{1}$$

where x and y are two different observing sequences, $m(\cdot)$ is the mean function, and $k(\cdot)$ is the covariance function, as:

$$m(x) = E[x], \tag{2}$$

and

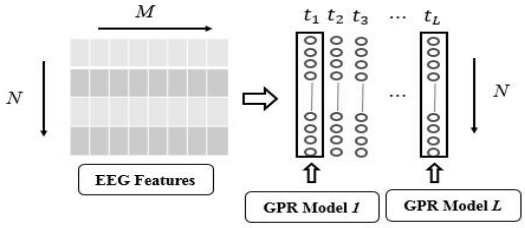

Fig. 2: GPR modeling process for each envelope. The GPR models are trained at each sampling instance t separately and here shows L GPR models. The EEG feature set (N × M) is used for L times.

$$k(x,y) = E[(x - m(x)) \cdot (y - m(y))], \tag{3}$$

where $E[\cdot]$ denotes an expectation operator.

Since the imagination of a speech utterance may not occur accurately at the same time while the imagined EEG signals are recorded [7], we hypothesized that the difference among the EEG signals from different recordings could be represented and learned by a covariance function. Also, we hypothesized that different recordings of the same word followed a Gaussian distribution at each sampling instance. Thus, we implemented a GPR algorithm for speech envelope prediction. Specifically, we predicted the Gaussian distribution at each sampling instance for the speech envelope at each frequency band (see Fig. 1). Figure 2 shows the GPR modeling process for the envelope signal at a specific sub-band in a speech recording.

Here we set EEG features as a and a specific envelope b, as:

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ \vdots & \cdots & \cdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix}, \tag{4}$$

and

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1L} \\ \vdots & \cdots & \cdots & \cdots \\ b_{N1} & b_{N2} & \cdots & b_{NL} \end{bmatrix}, \tag{5}$$

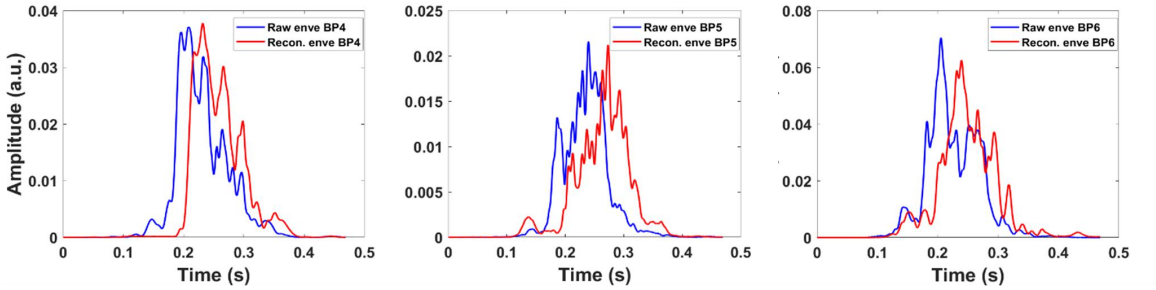

Fig. 3: Reconstructed envelopes (red) and raw envelopes (blue) from band 4 to band 6 of word /pat/.

where N is the total number of EEG (or speech) recordings, M is the dimension of EEG features, and L is the length of envelope, which equals the length of the speech signal.

The envelopes at each sampling instance l (= 1, 2, ⋯, L) follow the multivariate Gaussian distribution, and could be represented as:

$$[b_{1l}\cdots b_{Nl}]^T \sim N_l(\mu,K), \tag{6}$$

where μ is the mean vector consisted of the value of the mean functions, and K is the covariance matrix whose elements K_{ij} is a covariance function value of $k(a_i,a_j)$. Here $k(x,y)$ is the kernel function and we choose a squared exponential function as our kernel, as:

$$k(a_i,a_j) = \alpha^2 \exp\left(-\frac{\|a_i-a_j\|}{2\gamma^2}\right), \tag{7}$$

where α and γ are parameters corresponding to the EEG feature a_i and a_j. We set the hyperparameter ϑ, which consists of $[\alpha,\gamma]$.

Based on the above configuration, from the given EEG feature a_*, we apply the Gaussian likelihood function for the prediction of envelope b_*, as:

$$p(b_*|a_*,a,b,\vartheta,l) = N_l(k_*^T K^{-1} b, \kappa - k_*^T K^{-1} k_*), \tag{8}$$

where

$$k_* = K(a_*,a) = [k(a_*,a_1),\ldots,k(a_*,a_N)], \tag{9}$$

and

$$\kappa = k(a_*,a_*). \tag{10}$$

For achieving the best performance of envelope prediction in Equ. (8), the optimization of hyperparameter ϑ is performed among the training set $[a,b]$ by maximizing the contingent probability $p(a|b,\vartheta)$. We take conjugated gradient descent on the logarithm of $p(b|a,\vartheta)$ for the purpose of optimization, as:

$$log\, p(b|a,\vartheta) = -\frac{N}{2} log 2\pi - \frac{1}{2} b^T K^{-1} b - \frac{1}{2} \log|K|\,. \tag{11}$$

III. Experiments

A. Database

This work selected a public EEG database, i.e., KARAONE [7], for the task of speech reconstruction with EEG signals during speech imagery. This dataset combined 3 modalities (i.e., EEG, face tracking, and audio) during imagined and vocalized phonemic and single-word prompts and 4 female and 8 male participants (27.4 ± 5 years) were included. In their EEG recording experiments, seven phonemic prompts (/iy/, /piy/, /tiy/, /diy, /uw/, /n/ and /m/) and four words (/pat/, /pot/, /knew/ and /gnaw/) were used in repeated experimental trials, and each participant produced 132 trials. Each trial consisted of four stages in the following sequence: (1) a 5-second rest state where participants cleared their mind; (2) a stimulus state where the prompt or word text would appear on the screen and the corresponding utterance was played by speaker; (3) a 5-second imagined speech state where the participant imagined speaking the prompt or word; and (4) a speaking state where the participant spoke the prompt or word aloud. The EEG data were sampled at 1 kHz.

In this work, only EEG signals during speech imagination for words (i.e., /pat/, /pot/, /knew/ and /gnaw/) from the imagined speech state were used for speech reconstruction. We selected 10 channels (i.e., FC6, FT8, C5, CP3, P3, T7, CP5, C3, CP1 and C4) of imagined EEG data since EEG signals from those channels were highly correlated with corresponding speech waveforms [7]. The speech waveforms of the words recorded during the speaking state were selected as the reference to the reconstructed speech waveforms. Since data from 7 participants were with potential problems (i.e., the ground wire was not well connected for 4 participants and 2 participants fell asleep during the experiments; speech recordings from 1 participant were contaminated [7]), we selected the EEG data of the rest 5 participants for imagined speech reconstruction (i.e., participants MM05, MM08, MM09, MM14 and MM15).

B. EEG and speech pre-processing

EEG signals were pre-processed by independent component analysis [17] with software EEGLAB [18]. Bio-signal artifacts like electrocardiography and electromyography were removed. The data were band-pass filtered between 1 Hz and 50 Hz. As processed in [7], EEG signals were segmented by windows. Each window was with a length of 10% of the EEG signal, and there was a 50% of overlap between two adjacent windows, yielding 19 windows for each EEG trial. We

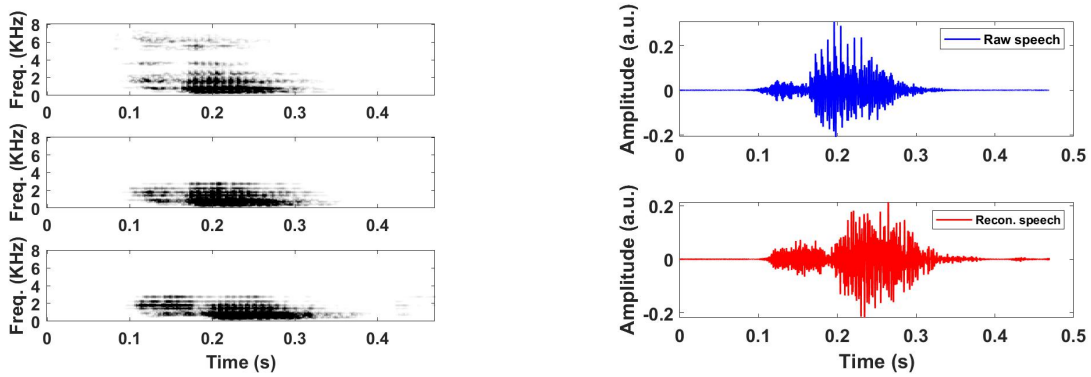

Fig. 4: (Left) Spectrograms of raw speech, vocoded raw speech and vocoded reconstructed speech (from top to bottom) of word /pat/. (Right) raw speech input (blue) and the vocoded reconstructed speech output (red).

calculated the mean value within each window, and computed their first and second order differentials (delta and delta-delta) as EEG features. The reason of this feature selection is that the distribution of EEG data became more Gaussian after mean operation in windows [24], and differential operation carried temporal-changing information. Finally, we calculated the mean value of EEG features among the 10 selected EEG channels [7]. Thus, for each EEG recording, there were 57 (=19 × 3) features or a 1 × 57 feature vector.

Speech recordings were sampled at a rate of 16 kHz. This work applied a Wiener filter to purify the speech from background noise. The contaminative speech and the corresponding EEG signals were deleted. Then we applied VAD to eliminate silent parts and truncated the active speech parts with same length L. Finally, we executed power normalization between speech waveforms.

C. *Model training procedure*

For generating the raw envelopes from a speech recording, this work used a 6th-order Butterworth filter to band-pass filter the input signal into 12 frequency bands between 80 and 3000 Hz according to the cochlear frequency-position mapping function [19]. Through full-wave rectification and low-pass filtering (a 2nd-order Butterworth filter with a cut-off frequency of 200 Hz), the raw envelope waveform was extracted from each frequency band and applied to the GPR model training. With reconstructed envelopes, this work used sinusoidal signal as carrier signal. All amplitude-modulated (using reconstructed envelopes) carrier signals were summed up to generate the vocoded reconstructed speech signal. Finally, the vocoded reconstructed speech signal was adjusted to have the same root-mean-square power as the input speech recording. We trained GPR using the GPML toolbox [20] on Matlab. The main setting of GPR is described in Table I. We trained different models at each sampling instance. The EEG features were used for L times for a single envelope, and there were L × 12 (sub-bands) models for an EEG-speech pair.

The training processes were conducted separately for 4 different words in this experiment (i.e., /pat/, /pot/, /knew/, and /gnaw/) across 5 participants. With the selected 5 participants, we conducted a leave-one validation over the participants and randomly shuffled the training data from 4 participants (i.e., MM05, MM08, MM09 and MM15). In the KARAONE database, there were 12 recordings for each word and each subject, and thus there were 48 (=12 × 4) EEG-speech recordings for each word in the training set. For the testing stage, a new participant (i.e., MM14) was selected for validation with 12 EEG-speech pairs for each word.

Table I: Parameters initialization for GPR setting.

Mean	Covariance	Likelihood	optimization
Null	[0, 0]	-1	Polack-Ribiere

D. *Evaluation*

Early work showed that the normalized covariance metric (NCM) between the raw and processed (i.e., reconstructed in this work) speech signals is highly correlated to the intelligibility of the processed speech [21]. The value of NCM measure ranges from 0 to 1, and a large NCM value indicates a high intelligibility of the processed speech. The computation of the normalized covariance between two envelopes is:

$$r_i = \frac{\sum_t (x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\sum_t (x_i(t) - \mu_i)^2}\sqrt{\sum_t (y_i(t) - \nu_i)^2}}, \quad (12)$$

where $x_i(t)$ is the raw envelope and $y_i(t)$ is the processed (i.e., reconstructed) envelope. μ_i and ν_i are mean values of $x_i(t)$ and $y_i(t)$, respectively. Finally, the normalized covariance of all 12 bands are averaged to give the NCM index of the processed (or reconstructed) speech signal.

Also, the intelligibility of the vocoded reconstructed speech was evaluated with the short-time objective intelligibility (STOI) measure [22]. STOI measures the distortion in spatiotemporal modulation patterns from the noisy or processed speech signal and it takes values from 0 to 1 where a large STOI value indicates a high speech intelligibility. This work calculated STOIs between two signal pairs, i.e., 1) vocoded raw speech (VS) and raw speech (RS), and 2)

Table II: Normalized covariance measures for reconstructed words.

/pat/	/pot/	/knew/	/gnaw/	Mean
0.57±0.25	0.51±0.18	0.60±0.14	0.61±0.12	0.57±0.17

Table III: STOIs for reconstructed words.

	/pat/	/pot/	/knew/	/gnaw/	Mean
VS vs. RS	0.89± 0.02	0.86± 0.03	0.83± 0.04	0.88± 0.03	0.87±0.03
ReconS vs. RS	0.74± 0.15	0.59± 0.19	0.71± 0.19	0.77± 0.16	0.70±0.17

vocoded reconstructed speech (ReconS) and raw speech, represented as 'VS vs. RS' and 'ReconS vs. RS', respectively.

IV. Results

For each word with 12 EEG-speech pairs, 12 NCM measures between reconstructed envelopes (ReconS) and the raw envelopes (RS), and 12 STOI values between vocoded raw speech (VS) or vocoded reconstructed speech (ReconS) and raw speech (RS) were calculated. Results for each word over the selected 5 participants are shown in Table II and Table III, respectively. The mean of NCM measures of the vocoded reconstructed speech from all 4 words is 0.57. The mean of STOI values of the vocoded reconstructed speech from all 4 words is 0.70, while that of the vocoded raw speech is 0.87.

Since there was little previous research synthesizing the intelligible speech synthesis from the EEG recordings during speech imagination, the studies investigating speech synthesis based on listening-based ECoG, spoken ECoG and mimed ECoG (i.e., the ECoG signals when participant was listening, reading aloud and reading silently the words or sentences) were selected for comparison. Similar to the speaking state of the KRAONE experiment, Angrick et al. [24] instructed participants to read aloud the word text on the screen and reconstructed the speech from the spoken ECoG for 6 participants. Between the reconstructed features and vocoded features (i.e., reconstructed sub-band envelopes and raw sub-band envelopes in this work), they obtained averaged NCM of 0.36 for all trials of words. Anumanchipalli et al. [6] synthesized the speech waveform from mimed ECoG for one participant when the participant was reading 58 sentences and obtained averaged NCM of slightly over 0.30. Akbari et al. [5] reconstructed the speech from listening-based ECoG when 5 participants were listening to the utterances of digits (i.e., from zero to nine). They calculated the STOI value between the reconstructed waveforms and speech waveforms participants listened to, and received averaged STOI value of around 0.31. Therefore, results in this work largely outperformed the previous studies in terms of either objective NCM or STOI measurement.

Figure 3 shows examples of envelope reconstruction from band 4 to band 6 of word /pat/. It is seen that the reconstructed (in red) and raw (in blue) envelopes are with high similarity. However, there are distortions at high frequency (i.e., the fluctuating segments) in the reconstructed envelopes, which may reduce the value of NCM measure, as shown in Table II.

The spectrograms of raw speech, vocoded raw speech and vocoded reconstructed speech are shown in Fig. 4. Note that in both Fig. 3 and Fig.4, there is a delay or time shift between the reconstructed and raw envelopes or between the vocoded reconstructed speech and the raw speech, which may partially account for the reduced STOI value of the vocoded reconstructed speech compared with that of the vocoded raw speech (which does not have time shift relative to the raw speech) in Table III.

V. Discussion and Conclusions

The present work studied a vocoder-GPR framework for imagined EEG-speech BCI system. The proposed strategy was based on the perceptual importance of temporal envelope and its ability to synthesize an understandable speech with temporal envelopes from a limited number of frequency bands. Using the KARAONE dataset, speech and imagined EEG signals from 4 subjects were used as training data, and good results of speech reconstruction with validation for a new subject were achieved. This demonstrated the potential of implementing the temporal envelope-based speech reconstruction for a new unknown BCI user.

Regarding to the performance of envelope reconstruction, the normalized covariance measures between the reconstructed envelopes and the raw envelopes were with a standard deviation up to 0.17 (see Table II), which might be partially attributed to the distortions in the reconstructed envelope waveforms, as shown in Fig. 3. Besides, time shift occurred between the reconstructed and raw envelopes. This may be due to the fact that the EEG signals in speech imagination and speech signals were not recorded simultaneously since the EEG signals were from the 5-second imagined speech state and the corresponding speech signals were from the speaking state of the KARAONE experiment.

Note that, unlike the listening-based EEG or spoken-EEG experiments, in the 5-second imagined speech state of the KARAONE experiment, it was unrestricted when the participant began and stopped the imagination. For this reason, the timing of brain evocation could be different among different 5-second imagined EEG signals. Although we applied a covariance function in our model (as mentioned in section 2.2) to consider the difference (including different imagination timing) among EEG signals, there was no operation to align EEG signals with the original speech. Therefore, the output (i.e., reconstructed envelopes or speech) of our method did not guarantee that the timing and duration of the reconstructed speech were absolutely the same as those of the original spoken speech. For addressing this time shift problem, we manually aligned the reconstructed envelopes with the raw envelopes (i.e., calculating the cross-correlation actually), and received an improved NCM value up to 0.85 in a follow-up analysis.

In conclusion, this work studied an imagined EEG-speech BCI system. Spectrogram comparison and objective evaluation with two measures, i.e., normalized covariance and

STOI, showed that the reconstructed envelopes and vocoded reconstructed speech approached to the raw envelopes and vocoded raw speech, respectively. Further efforts to improve the intelligibility of the reconstructed speech could be directed to aligning the reconstructed envelopes relative to the raw envelopes.

ACKNOWLEDGMENTS

This work was supported by the Stable Support Plan Program of Shenzhen Natural Science Fund, the National Natural Science Foundation of China (Grant No. 61971212), and High-level University Fund G02236002 of Southern University of Science and Technology.

REFERENCES

[1] H. J. Hoffman, C. M. Li, K. Losonczy, M. S. Chiu, J. B. Lucas, St. K. O. 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. J. Vidal, "Toward direct brain–computer communication," *Annual Review of Biophysics and Bioengineering*, 2, pp. 157-180, 1973.

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

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

[5] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, 9, article number 874, 2019.

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

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

[8] P. Sun, J. Qin, "Neural networks based EEG-speech models," *arXiv: 1612.05369v*, 2017.

[9] P. Saha, M. Abdul-Mageed, S. Fels, "SPEAK YOUR MIND! Towards imagined speech recognition with hierarchical deep learning," *Conference of the International Speech Communication Association (InterSpeech)*, pp. 141-145, 2019.

[10] B. M. Idrees, O. Farooq, "Vowel classification using wavelet decomposition during speech imagery," *International Conference on Signal Processing & Integrated Networks (SPIN)*, pp. 636-640, 2016.

[11] S. Deng, R. Srinivasan, T. Lappas, M. D'Zmura, "EEG classification of imagined syllable rhythm using hilbert spectrum methods," *Journal of Neural Engineering*, 7(4), pp. e046006, 2010.

[12] E. F. Gonzalez-Castaneda, A. A. Torres-Garcia, C. A. Reyes-Garcia, L. Villasenor-Pineda, "Sonification and textification: Proposing methods for classifying unspoken words from eeg signals," *Biomedical Signal Processing and Control*, 37, pp. 82–91, 2017.

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

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

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

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

[17] G. Gomez-Herrero, W. D. Clercq, H. Anwar, O. Kara, K. Egiazarian, S. V. Huffel, and W. V. Paesschen, "Automatic removal of ocular artifacts in the EEG without an EOG reference channel," *Nordic Signal Processing Symposium (NORSIG)*, pp. 130-133, 2006.

[18] A. Delorme, 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.

[19] D. 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.

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

[21] F. Chen, P. C. Loizou, "Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech," *Journal of the Acoustical Society of America,* 128 (6), pp. 3715-3723, 2010.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, 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.

[23] M. A. Bakhshali, M. Khademi, A. E. Moghadam, S. Moghimi, "EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density," *Biomedical Signal Processing and Control*, 59, 101899, 2020.

[24] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of Neural Engineering*, 16(3), 036019, 2019.