Computers in Biology and Medicine 182 (2024) 109090

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Speech synthesis from three-axis accelerometer signals using conformer-based deep neural network

Jinuk Kwon [a], Jihun Hwang [a], Jee Eun Sung [b], Chang-Hwan Im [a,c,d,e,*]

[a] Department of Electronic Engineering, Hanyang University, Seoul, South Korea
[b] Department of Communication Disorders, Ewha Womans University, Seoul, South Korea
[c] Department of Biomedical Engineering, Hanyang University, Seoul, South Korea
[d] Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[e] Department of HY-KIST Bio-Convergence, Hanyang University, Seoul, South Korea

ARTICLE INFO

Keywords:
Spoken speech synthesis
Three-axis accelerometer
Deep neural network
Conformer
Silent speech interface

ABSTRACT

Silent speech interfaces (SSIs) have emerged as innovative non-acoustic communication methods, and our previous study demonstrated the significant potential of three-axis accelerometer-based SSIs to identify silently spoken words with high classification accuracy. The developed accelerometer-based SSI with only four accelerometers and a small training dataset outperformed a conventional surface electromyography (sEMG)-based SSI. In this study, motivated by the promising initial results, we investigated the feasibility of synthesizing spoken speech from three-axis accelerometer signals. This exploration aimed to assess the potential of accelerometer-based SSIs for practical silent communication applications. Nineteen healthy individuals participated in our experiments. Five accelerometers were attached to the face to acquire speech-related facial movements while the participants read 270 Korean sentences aloud. For the speech synthesis, we used a convolution-augmented Transformer (Conformer)-based deep neural network model to convert the accelerometer signals into a Mel spectrogram, from which an audio waveform was synthesized using HiFi-GAN. To evaluate the quality of the generated Mel spectrograms, ten-fold cross-validation was performed, and the Mel cepstral distortion (MCD) was chosen as the evaluation metric. As a result, an average MCD of 5.03 $\pm$ 0.65 was achieved using four optimized accelerometers based on our previous study. Furthermore, the quality of generated Mel spectrograms was significantly enhanced by adding one more accelerometer attached under the chin, achieving an average MCD of 4.86 $\pm$ 0.65 ($p < 0.001$, Wilcoxon signed-rank test). Although an objective comparison is difficult, these results surpass those obtained using conventional SSIs based on sEMG, electromagnetic articulography, and electropalatography with the fewest sensors and a similar or smaller number of sentences to train the model. Our proposed approach will contribute to the widespread adoption of accelerometer-based SSIs, leveraging the advantages of accelerometers like low power consumption, invulnerability to physiological artifacts, and high portability.

1. Introduction

Silent speech interfaces (SSIs) are emerging alternative communication methods that do not depend on vocalized speech, which is the most natural and widespread form of communication in human society. An SSI enables users to convey their intentions when audible speech is limited or unavailable [1]. This technology can be used in a wide range of applications across diverse fields. For example, an SSI can assist patients suffering from speech impairments due to traumatic injuries, laryngectomy, and neurodegeneration by restoring their ability to interact with external environments and significantly improving their quality of life [2]. An SSI can also be employed to communicate in noisy environments where verbal communication may be hindered by ambient noise [3] and in noise-sensitive environments where quietness is necessary [4]. Additionally, the utilization of SSIs could be invaluable for security-sensitive tasks, including military operations [5] and the maintenance of user privacy in public places [6]. Owing to their promising potential as novel modes of communication, SSIs have

* Corresponding author. Department of Biomedical Engineering, Hanyang University, Seoul, South Korea.
E-mail addresses: kowm2000@hanyang.ac.kr (J. Kwon), fhffoddl0616@hanyang.ac.kr (J. Hwang), jeesung@ewha.ac.kr (J.E. Sung), ich@hanyang.ac.kr (C.-H. Im).

https://doi.org/10.1016/j.compbiomed.2024.109090
Received 22 May 2024; Received in revised form 23 August 2024; Accepted 29 August 2024
Available online 3 September 2024

recently generated considerable interest.

The successful implementation of SSIs, which interpret nonaudible speech-related activities and convert them into audible speech and text, crucially depends on the utilization of suitable sensing techniques. These techniques must effectively capture speech-related information embedded in non-acoustic activities. The most representative modalities for SSIs include imaging techniques, such as video cameras [7] and ultrasound imaging [8]. These techniques can directly measure the movement of articulators, such as the lips, jaw, tongue, and vocal tract of a body, using high-resolution two- and three-dimensional images, allowing the implementation of high-performance SSIs. Magnetic articulography-based modalities, including permanent magnet articulography [9] and electromagnetic articulography (EMA) [10], are also key sensing methods for SSIs. These modalities track the movements of speech articulators by detecting changes in the magnetic field using magnetic sensors attached to the articulators. Recently, the use of electrophysiological signals such as muscular activity (i.e., surface electromyography (sEMG)) [11] and brain activity (i.e., electroencephalography (EEG) and electrocorticography) [12] to implement SSIs have been extensively investigated. In addition, studies have investigated the feasibility of various other modalities such as magnetic resonance imaging [13], electro-optical somatography [14], electropalatography (EPG) [15], and ultra-wideband impulse radar [16] for implementing SSIs.

However, these conventional modalities have practical limitations in real-world scenarios because of their bulkiness, stationarity, high cost, low signal quality, invasiveness, and lack of portability [15–19]. Although sEMG has advantages over other modalities, including lightweight sensors, affordability, and noninvasiveness, its practical application for long-term use is impeded by factors such as bodily fluids and muscle fatigue, which can alter the characteristics of sEMG signals [20]. To address these limitations, our previous study [21] implemented a novel SSI using three-axis accelerometers to measure the physical movements of articulators. The results demonstrated that only four accelerometers attached to the facial areas near the mouth could effectively measure speech-related information for classifying 40 words spoken silently. The SSI achieved a high classification accuracy of 95.58 %, surpassing that of a conventional sEMG-based SSI using six sEMG electrodes (accuracy = 89.68 %, $p < 0.0005$). Furthermore, the developed accelerometer-based SSI achieved similar or higher performance than other conventional modalities with a small amount of training data. Our initial results demonstrated that accelerometers have excellent potential as new modalities for implementing practical SSIs. As a next step, investigating the feasibility of synthesizing spoken speech solely from accelerometer signals is necessary. This is because the successful synthesis of spoken speech can demonstrate the potential for synthesizing silently spoken (or mimed) speech, which is the ultimate goal of SSIs, i.e., empowering users to express their intention rather than merely commanding a limited set of words.

Regarding conventional modalities, various studies have demonstrated the feasibility of synthesizing speech from data recorded while people read given sentences aloud. For example, Akbari et al. [22] reconstructed speech by recognizing lip movements from facial video recordings. They developed a deep neural network-based approach to extract speech-related features and generate an auditory spectrogram, which was then converted into an audio waveform using NSLtools [23]. Kimura et al. [24] developed SottoVoce, which synthesizes audio signals from tongue movements recorded using an ultrasound probe placed under the jaw. SottoVoce consists of two deep neural networks. The first neural network extracts an acoustic feature, i.e., a Mel-scale spectrum, from a sequence of ultrasound images, and the second neural network enhances the quality of the generated acoustic feature. The sequence of the generated Mel-scale spectrum is then converted into voice using a Griffin–Lim vocoder [25]. Taguchi and Kaburagi [26] developed a bidirectional long short-term memory (bi-LSTM)-based articulatory-to-speech conversion method. This method converts the movement trajectory of the lip and tongue obtained using EMA sensors into speech feature parameters such as Mel cepstral parameters, fundamental frequencies, and voiced/unvoiced flags. Speech is produced using a speech synthesizer called WORLD Morise et al., 2016 [27]. Janke and Diener [28] introduced a speech synthesis technique for sEMG of articulatory muscles. This technique converts features extracted from sEMG into Mel frequency cepstral coefficients and fundamental frequencies and then generates speech sounds using a Mel log spectrum approximation vocoder [29].

Based on previous studies, most speech synthesis technologies included two main components: 1) an acoustic feature generator that transforms source data into acoustic features and 2) a vocoder that synthesizes audio waveforms from these acoustic features. Neural vocoders have rapidly progressed with the application of deep generative models, including the autoregressive model [30], flow-based generative model [31], and generative adversarial network (GAN) [32]. Specifically, HiFi-GAN [33], a state-of-the-art GAN-based neural vocoder, synthesizes high-fidelity audio waveforms in real-time from a Mel spectrogram on a single GPU. In this study, we employed HiFi-GAN as a vocoder without further development and mainly focused on designing an optimal acoustic feature generator to convert accelerometer signals into a Mel spectrogram. For acoustic feature generators, extracting appropriate features is crucial for establishing the relationships between the source data and the Mel spectrogram. Recent advancements in deep learning have demonstrated that deep neural networks have exceptional capability for automatic feature extraction [34]. Convolutional neural networks (CNNs) exhibit remarkable proficiency in capturing local information, whereas recurrent neural networks (RNNs), such as LSTM and gated recurrent units, can effectively deal with temporal and sequential information [35]. In particular, bi-LSTM, an extended version of LSTM, can process sequence data more effectively by simultaneously considering both past and future information [36]. Recently, Transformer [37], a self-attention-based neural network, was proposed to address the long-sequence dependency problems inherent in conventional RNNs. Owing to its excellent performance, Transformer has become the model of choice for sequential decoding tasks, particularly in natural language processing. Furthermore, various Transformer variants have been developed and extended to a wide range of tasks, including vision-related tasks [38], genomic sequence analysis [39], and SSIs [40].

In this study, we investigated, for the first time, the possibility of speech synthesis using accelerometer-based SSIs. Five accelerometers were attached to the facial surface to acquire speech-related movements while participants spoke sentences aloud. We followed the conventional two-step approach involving the acoustic feature generator and vocoder. A new acoustic feature generator based on a deep neural network was proposed to convert recorded accelerometer signals into a Mel spectrogram. The developed deep neural network consists of convolution-augmented Transformer (Conformer) blocks [41], leveraging both local and global features, along with bi-LSTM to process information from both past and future contexts. The recent GAN-based neural vocoder, HiFi-GAN, was then employed to synthesize the audio signal from the generated Mel-spectrogram. Nineteen healthy individuals participated in the experiments and collectively read 270 sentences; the quality of the generated Mel spectrograms were evaluated in a subject-dependent manner using ten-fold cross-validation, with Mel cepstral distortion (MCD), the most commonly used metric in biosignal-based speech synthesis studies, as the evaluation metric.

The major contributions of this paper are as follows.

- This paper is the first study to investigate the feasibility of direct speech synthesis using three-axis accelerometer signals attached to the facial surface.
- We propose a new deep neural network architecture based on Conformer to convert accelerometer signals into the Mel

spectrograms of recorded sounds, measured simultaneously with the accelerometer signals as people spoke sentences aloud.
- The synthesized audio signals are highly intelligible, suggesting that three-axis accelerometers have great potential for silent speech interfaces.

The remainder of the paper is organized as follows: Section 2 describes the experimental paradigm, data recording, signals processing, and the method for converting three-axis accelerometer signals into corresponding Mel spectrograms. Section 3 presents the experimental results. Section 4 and Section 5 provide the discussion and conclusion of the research, respectively.

2. Methods

2.1. Participants

Nineteen native-Korean individuals (12 males and 7 females, aged 22.5 ± 2.09 years) participated in the experiments. None of the participants reported a serious history of neurological, psychiatric, or other severe diseases that could have influenced the experimental results. Before the experiments, details of the experiments were provided to all the participants, and written informed consent was obtained from each participant. The study and its experimental protocol were approved by the Institutional Review Board (IRB) of Hanyang University, Republic of Korea (IRB No. HYU-2019-11-007-7) according to the Declaration of Helsinki.

2.2. Experiment paradigm

In this study, 270 Korean sentences were used considering the overall experiment duration, which was limited to approximately 1 h, excluding the setup time, to minimize participant fatigue and psychological resistance due to the lengthy experiment time. Among them, 180 sentences were composed of words that covered as many Korean phonetic combinations as possible, whereas the remaining 90 sentences were extracted from the Korean Single Speaker Speech (KSS) dataset [42]. There are no identical ones among the 270 sentences, and the sentences and their English translations are listed in the Supplementary Material. The participants were seated in a comfortable armchair 70 cm away from an LCD monitor and completed nine sessions of the experiments, each comprising 30 trials of instruction and task periods, with sufficient breaks provided between sessions. During the instruction period, the participants were encouraged to take short breaks to prevent potential muscle fatigue, and the target sentence to be read was displayed on the screen to familiarize them. When the participants were ready to proceed, they were instructed to close their mouths and press the spacebar on the keyboard in front of them. During the task period, a beep sound was presented as a trigger for the participants to begin reading the designated sentence aloud. To ensure the acquisition of high-quality data, the participants were instructed to press the spacebar to complete the trial or to press "x" if they misread or stuttered, allowing them to read the sentence again. The timing sequence of a single trial is illustrated in Fig. 1. Accelerometer and voice signals from the 270 sentences (30 sentences × 9 sessions) were collected synchronously for each participant. The visual instructions and sound stimuli were presented using E-prime 3.0 (Psychology Software Tools, Sharpsburg, PA, USA).

2.3. Data recording and signal processing

2.3.1. Audio signals

The audio signals were recorded using a condenser USB microphone (VS-100, LAILTONE, South Korea) located under the LCD monitor (70 cm away from the participants) at a sampling rate of 44,100 Hz. The raw audio signals were segmented from 0.5 s relative to the task onset time to the end of the reading sentence. Here, data from 0 s to 0.5 s were excluded to prevent the beep sound from being recorded in the audio signals. The segmented audio signals were downsampled to 22,050 Hz and subsequently normalized between −0.95 and 0.95 for the application of HiFi-GAN (https://github.com/jik876/hifi-gan). The background noise in the normalized audio signals was reduced using the Noisereduce library (https://github.com/timsainb/noisereduce), with a decreased proportion of 0.95. Log-Mel spectrograms of the preprocessed audio signals were obtained using the TorchAudio library [43] with an nFFT of 1,024, 80 Mel filter banks, a maximum frequency of 8000 Hz, and the center option set false, which are the same parameters used in HiFi-GAN. The window and hop lengths were set to 40 and 20 ms, respectively, to align with the accelerometer signals, which had a sampling rate of 50 Hz.

2.3.2. Three-axis accelerometers

Five inertial microelectromechanical system chips (MPU-9250, Invensense, San Jose, CA, USA) were used to record the three-axis accelerometer signals at a sampling rate of 50 Hz. The locations of the four sensors were established based on the positions that achieved the highest classification accuracy in our previous study [21]. These were channel #1 next to the philtrum, channel #2 next to the lip corner, channel #3 under the lip and vertically aligned under channel #1, and channel #4 at the center of the base of the mandible. Another sensor (channel #5) was attached beneath the chin, which is a location commonly used in sEMG-based SSI studies [44,45]. The locations of the sensors are shown in Fig. 2.

The raw accelerometer signals were segmented into epochs from 0.5 s relative to the task onset time to the time when the participant pressed the spacebar, to align with the audio signals for each trial. The segmented accelerometer signals were then normalized using z-score normalization along the time axis, as we did in our previous study, to increase the training speed and enhance performance[21,46–48]. No

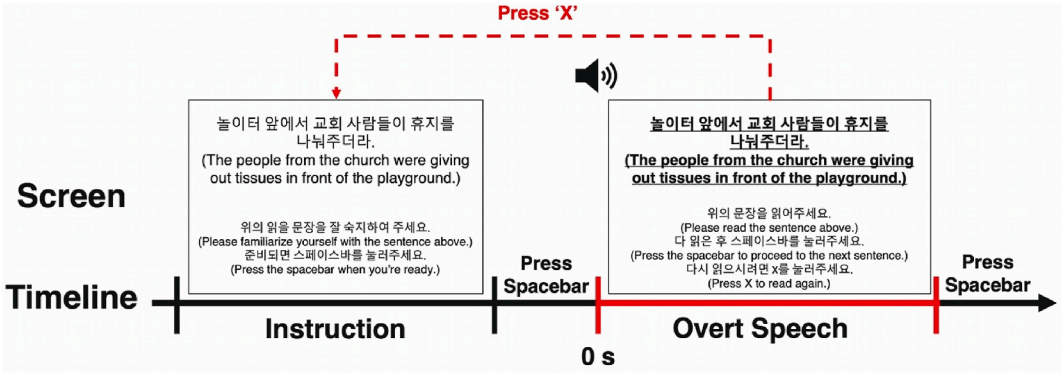

Fig. 1. Timing sequence of a single trial. The text inside the parentheses is the English translation of the corresponding Korean text presented to the participants.

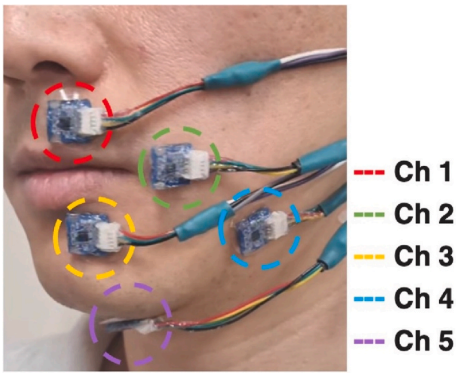

Fig. 2. Locations of three-axis accelerometers.

additional preprocessing and calibration were applied to the normalized accelerometer signals.

2.4. Accelerometer-to-Speech

2.4.1. Acoustic feature generator for accelerometer signals

In this study, we developed a Conformer-based acoustic feature generator to convert accelerometer signals into Mel spectrograms. Conformer [41] has an advanced Transformer architecture that combines Transformer and CNN models. Accordingly, it compensates for the limited fine-grained local feature extraction and high computational cost of the Transformer as well as the limited capability of a CNN to capture global contexts. Combining the Transformer and CNN models enables the Conformer model to capture both local and global dependencies in sequence data while maintaining computational efficiency. This has made the Conformer model a de facto model for various speech-processing tasks such as automatic speech recognition [49,50].

The developed acoustic feature generator comprises a linear layer, four sequential Conformer blocks, and two bi-LSTM layers, followed by a linear layer, as shown in Fig. 3. The first linear layer, consisting of 2048 hidden units, processes the accelerometer signals, whose dimensions are number of time samples × number of channels (i.e., 3 axes × 5 channels = 15), resulting in an output with dimensions of number of time samples × 2048. Note that using a consistent number of time samples is necessary; however, the data lengths of the different sentences were different in our case. To address this issue, we set the number of time samples to 128 to train the model and employed a random starting point for each sentence in every epoch to ensure that the entire sentence could be utilized in the training process. In the acoustic feature generator, all the Conformer blocks have an identical architecture consisting of four sequentially stacked modules comprising two feedforward modules with multi-head self-attention and convolution modules in between, followed by layer normalization. The detailed architecture of each module is shown in Fig. 4. The hyperparameters of the encoder dimension, number of attention heads, convolution kernel size, and expansion factor of the feedforward modules were set to 2,048, 32, 31, and 3, respectively. Other hyperparameters were set to the same values as those used in the default Conformer model [41]. The two bi-LSTM layers also had the same parameter configurations of 2048 hidden units per direction, and the last linear layer comprised 80 hidden units, resulting in an output dimension of the number of time samples × 80 (the number of Mel filter banks).

The acoustic feature generator was trained to minimize the mean squared error between the log-Mel spectrograms derived from the preprocessed audio signals and those generated using the acoustic feature generator for 1000 epochs with a batch size of four. The AdamW [51] optimizer was employed with cosine annealing and a warmup restart scheduler [52]. The initial and maximum learning rates were set to 0.00001 and 0.0001, respectively. The cycle and warm-up steps were

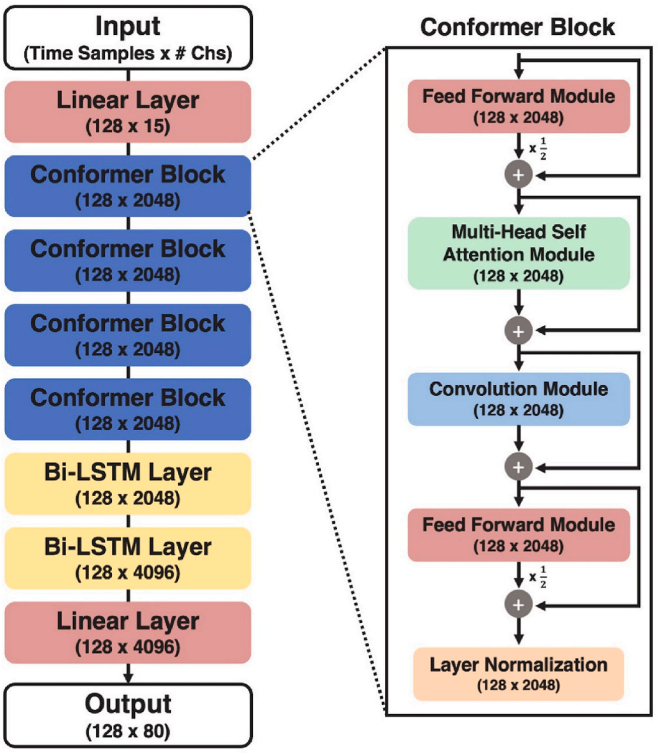

Fig. 3. Architecture of developed acoustic feature generator. Dimensions listed under each network represent input dimensions for each layer, block, and module.

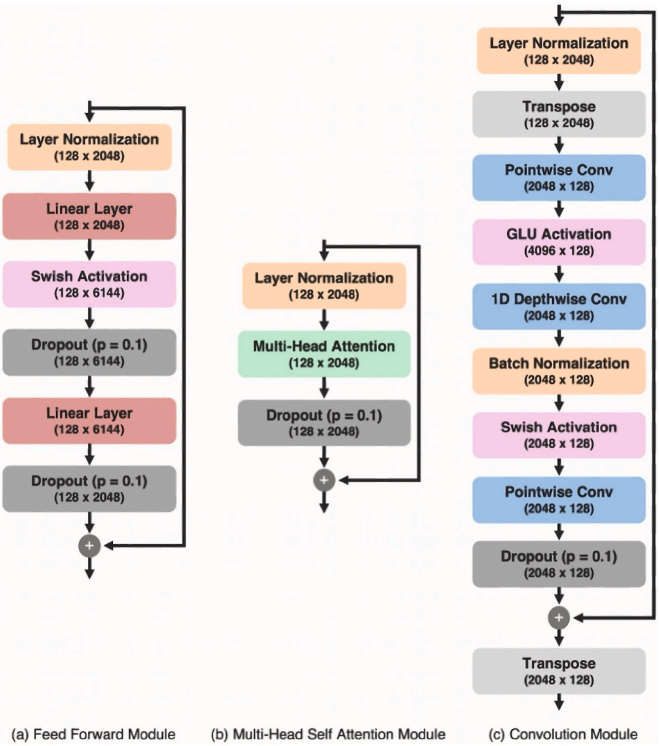

Fig. 4. Architecture of each module comprising Conformer blocks. Dimensions listed under each layer represent input dimensions.

500 and 100, respectively, and the rate of decrease in the maximum learning rate per cycle was 0.9. All the hyperparameters were empirically determined using data from the first two folds of Participant #1. All implementation and training processes were performed using PyTorch [53], and an NVIDIA RTX 2080Ti GPU was used to accelerate the training process. The overall flowchart of the developed system is shown in Fig. 5.

2.5. Performance evaluation

The performance of the developed SSI system was evaluated in a subject-dependent manner, meaning that an individual model was created for each participant. A ten-fold cross-validation strategy was employed, where the dataset was divided into ten sub-datasets, each comprising 27 sentences. Nine of these sub-datasets were used to train the model, while the remaining subset was used to evaluate the quality of the generated Mel spectrogram. This procedure was repeated ten times to assess the quality of all the synthesized sentences, with each model trained exclusively on data from the corresponding participant. Here, it is noteworthy that none of the 270 sentences used in this study were identical, indicating that all the sentences in the test set of each fold were entirely not included in the training dataset. The MCD, calculated using the "melcd" function in the nnmnkwii library (https://github.com/r9y9/nnmnkwii), was used as the evaluation metric.

3. Results

3.1. Mel spectrogram generated using three-axis accelerometer signals

The ground truth and generated Mel spectrograms from Participants #4, #9, and #15 are shown in Fig. 6. Here, the "Ground Truth" refers to the Mel spectrograms derived from the preprocessed audio signals and "Generated" denotes those converted from the three-axis accelerometer signals using the developed acoustic feature generator. Note that all the Mel spectrograms generated from the three-axis accelerometer signals were selected from the test set of the ten-fold cross-validation. The exemplar sentences shown in Fig. 6 are randomly selected from each participant. As shown in the figure, the generated Mel spectrograms closely match the original Mel spectrograms. MCDs were evaluated for all the sentences to assess the quality of the generated Mel spectrograms and then averaged across the 270 sentences for each participant. The mean MCDs with standard deviations for each participant are listed in Table 1. The grand average MCD for all the participants is 4.86 ± 0.65, which is denoted as "Mean" in the table.

3.2. Speech synthesis using generated mel spectrograms

In this study, HiFi-GAN, a GAN-based neural vocoder, was employed to convert the Mel spectrograms generated from the accelerometer signals into audio signals. The authors of HiFi-GAN provide a pre-trained model, trained with the LJSpeech dataset (https://keithito.com/LJ-Speech-Dataset/), that enables high-speed speech generation; however, Mel spectrogram parameters used in the pre-trained model differ from those in our study. Therefore, in this study, we modified the hyperparameters of HiFi-GAN V1. The original window size of 1024 and hop size of 256 were changed to 882 (40 ms) and 441 (20 ms), respectively. The nFFT and segment size, initially set to 1024 and 8,192, were adjusted to 882 and 14,112, respectively. Furthermore, the upsampling rates and kernel sizes were changed from [8, 8, 2, 2] to [7, 7, 3, 3] and from [16, 16, 4, 4] to [13, 13, 5, 5], respectively. The modified HiFi-GAN was trained using the same LJSpeech dataset as the pre-trained model for 1000 epochs. Examples of the synthesized audio signals are available at https://jkwon0331.github.io/Acc2Speech/.

3.3. Effect of additional channel

In this study, channel #5 was newly employed, in addition to the four channels that showed the highest classification accuracy in our previous SSI study. To further investigate the effect of adding channel #5, the MCDs were evaluated using both four- and five-channel configurations. In Fig. 7, the white and gray bars represent the averaged MCDs for the four- and five-channel configurations, respectively. The error bars indicate the standard deviation. The grand averaged MCDs for the four- and five-channel configurations over all the participants are 5.03 ± 0.65 and 4.86 ± 0.65, respectively, denoted by "Mean" in the figure. Here, a lower MCD implies better speech synthesis performance. Because the Kolmogorov–Smirnov test indicated that the normality criterion was not satisfied, the Wilcoxon signed-rank test was performed for statistical analyses. The results showed a significant statistical difference between the averaged MCDs for the four- and five-channel configurations ($p < 0.001$), indicating that the addition of the fifth channel improved the quality of speech synthesis.

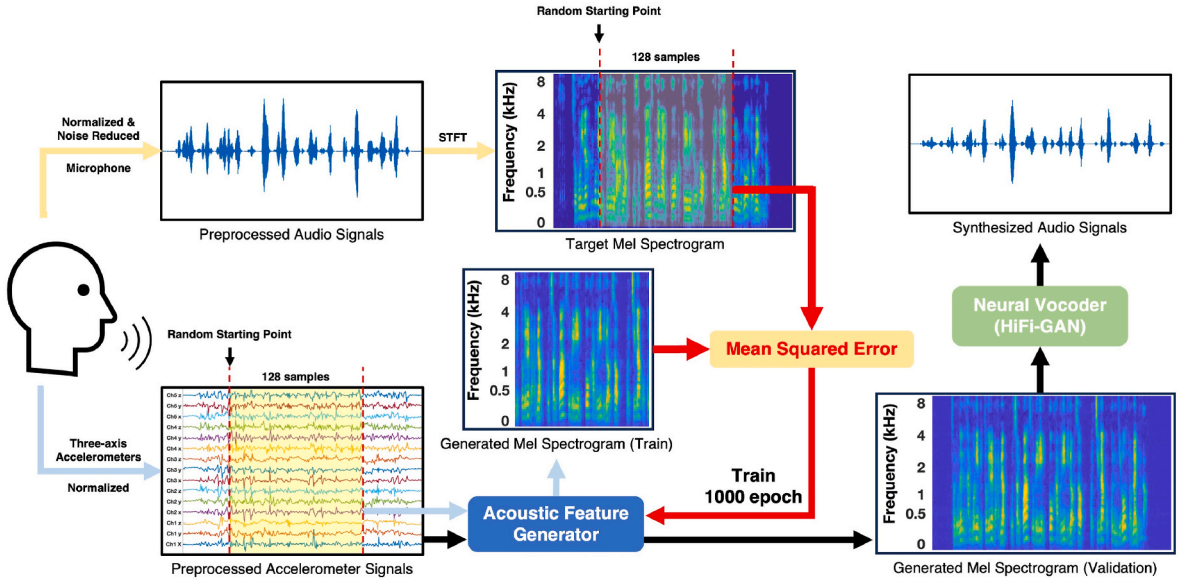

Fig. 5. A detailed Schematic of the pipeline for the proposed system. Both the target Mel spectrogram and preprocessed accelerometer signals share an identical random starting point. After the training phase, only the pipeline indicated by the black line is utilized for validation.

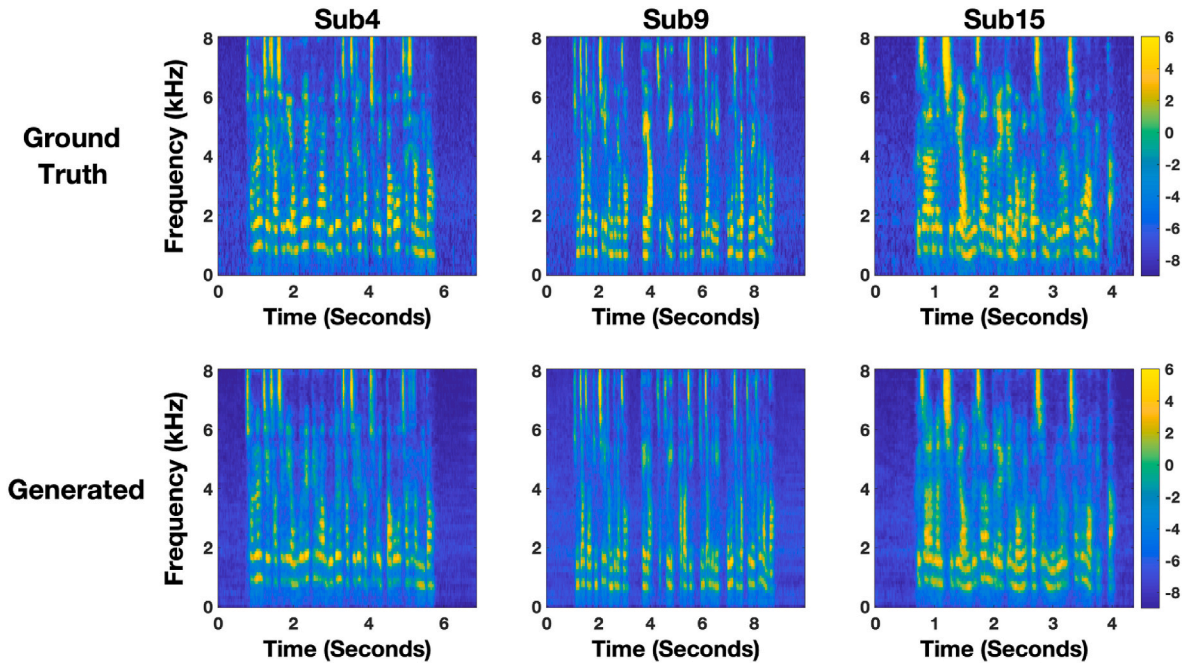

Fig. 6. Exemplar Mel spectrograms for three participants. "Ground Truth" denotes Mel spectrograms derived from preprocessed audio signals and "Generated" indicates those generated from three-axis accelerometer signals. Color bars indicate decibel scale.

Table 1
Averaged MCDs and standard deviation (std) over 270 sentences.

Participant	1	2	3	4	5	6	7	8	9	10
Average	4.32	5.65	5.42	5.05	5.88	3.18	4.52	5.41	4.43	4.26
std	0.70	0.78	0.60	0.79	0.80	0.51	0.72	0.82	0.61	0.63
Participant	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	Mean
Average	4.34	5.15	5.58	4.96	5.52	4.58	4.71	4.63	4.81	4.86
std	0.71	0.76	0.80	0.66	0.84	0.97	0.80	0.62	0.91	0.65

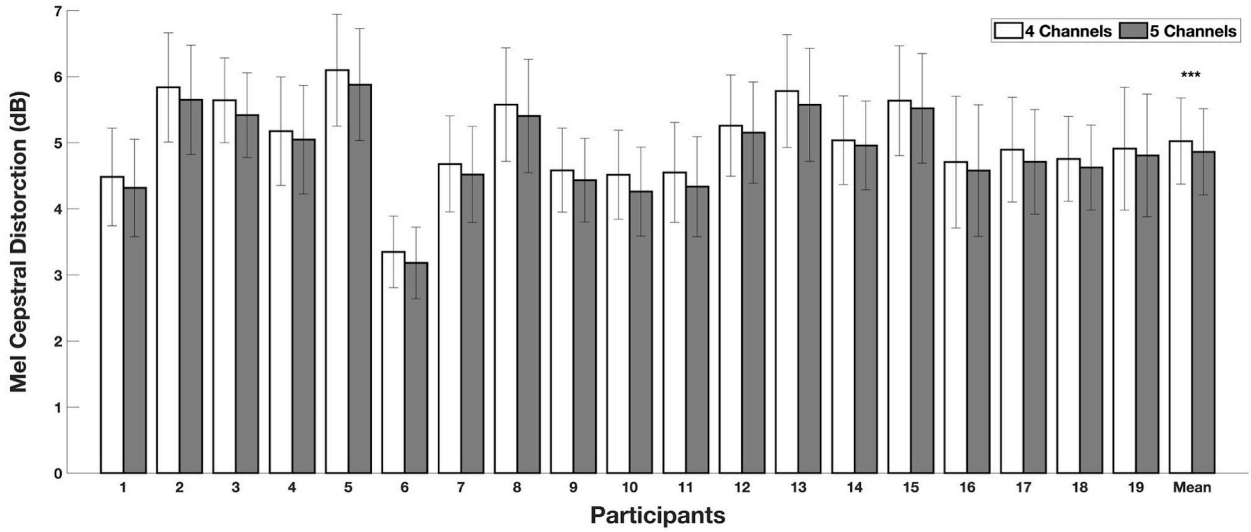

Fig. 7. Averaged Mel cepstral distortion (MCD), evaluated over all 270 sentences for both four- and five-channel configurations, represented as white and gray bars, respectively. Grand averaged MCDs for both configurations over all participants are 5.03 ± 0.65 and 4.86 ± 0.65, respectively, denoted by "Mean." Error bars indicate standard deviations. Here, *** represents $p < 0.001$ (Wilcoxon signed-rank test).

3.4. Comparison with conventional modalities

We compared the averaged MCD of our developed system with those obtained in previous speech synthesis studies that utilized other types of signals, including sEMG, EMA, and EPG, which are the most representative modalities for SSI because of their portability with small sensors

and wearables. Table 2 presents the results of the comparison. Notably, the developed method achieves the lowest averaged MCD even with the fewest sensors and similar or smaller amounts of training data. Additionally, our study involves the largest group of participants, demonstrating the generalizability of our results. Although this cannot be considered an objective evaluation, the results demonstrate that three-axis accelerometers are sufficiently effective in capturing speech-related information, enabling the generation of a Mel spectrogram that closely resembles the original Mel spectrogram. The result suggests that three-axis accelerometer-based SSIs have considerable potential for practical applications.

4. Discussion

In this study, we investigated the feasibility of synthesizing spoken speech using three-axis accelerometer signals for the first time. Five accelerometers were attached to the face of each participant to capture speech-related information while they read 270 Korean sentences aloud. A Conformer-based deep neural network was proposed as an acoustic feature generator to decode the accelerometer signals and reconstruct the corresponding Mel spectrogram, which was then converted to the audio signal using HiFi-GAN neural vocoder. The quality of the generated Mel spectrograms were evaluated using the MCD metric, employing a ten-fold cross-validation strategy. All 270 Korean sentences were tested using the proposed model whose training data did not include the test data of each fold. As a result, a grand average MCD of 4.86 ± 0.65 was achieved, and synthesized audio waveforms using a modified HiFi-GAN can be accessed from our website (https://jkwon0331.github.io/Acc2Speech/). The generated speech was sufficiently intelligible for any native Korean speaker to identify the meaning of the spoken sentences. Interestingly, the synthesized speech closely resembled the ground-truth speech even though HiFi-GAN was trained on an English sentence dataset. Moreover, the current study identified an additional beneficial location for three-axis accelerometer-based SSIs, which can significantly improve the quality of the generated Mel spectrograms. To the best of our knowledge, this is the first demonstration of synthesizing intelligible speech using three-axis accelerometer signals. Our study represents a necessary and significant step forward in the implementation of accelerometer-based SSIs for practical applications.

Although our study showed promising results, some limitations should be addressed in future studies. First, a major limitation is that accelerometer signals cannot capture tongue movements, which adversely affects the quality of both generated Mel spectrograms and synthesized speeches. This issue is particularly critical because many words involve similar or identical movements of the mouth, to which the accelerometers were attached in this study. Therefore, accurately capturing tongue movements is essential for realizing an accurate and reliable communication system. We expect that the use of large language models such as Chat-GPT can enhance the performance of accelerometer-based SSIs after transferring the synthesized speech to text. In addition, the concurrent use of other biosignals, such as sEMG, can be considered to enhance the performance of speech synthesis to measure tongue-related information during speech. Second, the channel configuration used in this study has not yet been optimized. Given that accelerometer-based speech research is a relatively new field, no studies have specifically focused on determining the optimal channel configuration. In this study, we employed an additional accelerometer attached under the chin with the four-channel configuration suggested in our previous study, and this additional accelerometer significantly contributed to improving the quality of the generated Mel spectrogram. In addition, we believe that there are still considerable areas to explore, including the neck, which may provide additional information on vocal cord movements. Therefore, we plan to investigate optimized channel configurations for accelerometer-based SSIs, potentially by employing high-density sensors as utilized in previous sEMG-based SSI studies [57,58].

Thanks to recent advancements in deep learning, the quality of generated Mel spectrograms can be enhanced. First, feature extraction and selection methods can be employed. In the current study, we used only the normalized accelerometer signals without additional feature extraction. However, recent studies have demonstrated that exploiting appropriate features can significantly improve the performance of deep learning models [59,60]. Therefore, it is necessary to explore feature extraction and selection methods for accelerometer-based speech interfaces in future studies. Furthermore, the latest hybrid deep learning architecture can be employed by combining the proposed network with a parallel inception concept to deal with multiple features [59] or cosine similarity [61]. Second, applying the GAN [62] architecture can notably improve the generative results. GAN consists of two competing networks: a discriminator, to discern whether the input is real or generated, and a generator, to create realistic data to deceive the discriminator. The fundamental concept of GANs is that competitive interaction between the generator and discriminator can produce high-quality realistic data. If a large dataset consisting of original and synthesized Mel spectrograms acquired from multiple subjects could be collected in future studies, the overall performance of accelerometer-based SSIs might be greatly improved by training GAN models using both spectrograms. This is an interesting topic that should be pursued in future studies. Additionally, the performance of deep neural networks is significantly influenced by the amount of training data, with larger datasets typically leading to higher performance [47,63]. In this context, the utilization of recent deep learning-based data augmentation techniques [64,65] can effectively increase the training dataset without needing additional training sessions and can serve as a potential approach to enhance the quality of generated Mel spectrograms. Similarly, subject-based transfer learning with fine-tuning [59,66,67] is considered a promising method for improving the performance of deep neural networks by leveraging data from different participants. Therefore, investigating appropriate data augmentation and transfer learning methods for accelerometer-based SSIs could be valuable for future research. Ultimately, our final goal is to implement subject-independent SSIs without individual training data from new users to achieve the practical use of SSIs as novel communication modes. Since our results applied to cases where the training and evaluation data came from the same participants, the method may not perform well or generalize effectively on completely new users. Therefore, it is necessary to pursue further studies to utilize the proposed accelerometer-based SSIs in a subject-independent manner. This issue seems to be challenging because training the deep learning model with data from 18 participants to test a single participant's data takes about a month under our computational environment. This implies that the subject-independent SSIs cannot be implemented on portable devices such as tablets or smartphones. Therefore, further studies needs to be conducted to develop a new deep learning model appropriate for implementing subject-independent SSIs,

Table 2
Comparison with previous studies.

Authors	# Participants	Modality (# Sensors)	Dataset	Training Data	Average MCD
Diener et al. [54]	8 speakers	sEMG (40 Sensors)	290 Utterances	250 Utterances	7.68
Chen et al. [55]	3 speakers	EMA (9 Sensors)	354 Utterances	304 Utterances	7.176
Chen et al. [56]	1 speaker	EPG (124 Sensors)	320 Utterances	222 Utterances	5.173
Ours	19 speakers	Accelerometer (5 Sensors)	270 Utterances	243 Utterances	4.86

which reduces the computational burden while maintaining the overall performance.

Building on the findings of this study, our next research will focus on investigating the feasibility of reconstructing silent speech using accelerometer signals, which would be more challenging owing to the absence of audio signals. We believe that this challenge can be addressed by employing a length regulator-based approach that utilizes both spoken and silent speech datasets, as demonstrated in a previous EMG-based SSI study [45]. Another major research topic is the implementation of wireless and wearable accelerometer-based SSIs for practical applications. To this end, we can utilize tattoo-like flexible sensors [68] that have been successfully applied in various fields to reduce potential discomfort in real-world scenarios. This advancement can significantly enhance the practicality of accelerometer-based SSIs. Furthermore, considering stretchable sensors, it is worthwhile to investigate the feasibility of hybrid SSIs that integrate our developed accelerometers with sEMG sensors to improve the performance of SSIs. Several stretchable sEMG sensors have already been introduced [69,70], which can be readily combined with three-axis accelerometers. Since both sensors measure different underlying signals, their combination can exploit the complementary characteristics of the two modalities, potentially leading to a substantial improvement in the overall SSI performance.

5. Conclusion

In this study, we demonstrated for the first time the feasibility of spoken speech synthesis from three-axis accelerometer signals. We employed five accelerometers attached to the face to capture speech-related information and proposed a new Conformer-based acoustic feature generator to convert the recorded signals into a Mel spectrogram. The grand average MCD of the developed method was 4.86 ± 0.65 across 19 participants, significantly surpassing the performances of conventional modalities such as EMA, EPG, and sEMG. Moreover, the developed approach does not require any prior knowledge-based preprocessing of accelerometer signals: it can be easily realized with simple z-score normalization. The reconstructed Mel spectrograms were then fed to a HiFi-GAN neural vocoder to synthesize the audio waveforms. Interestingly, the synthesized audio samples were similar to the original voices and intelligible enough, even though the neural vocoder was trained using sentences in different languages. Our findings can be considered a significant milestone in demonstrating the high potential of accelerometer-based SSIs.

CRediT authorship contribution statement

Jinuk Kwon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Jihun Hwang:** Data curation. **Jee Eun Sung:** Methodology. **Chang-Hwan Im:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MIST) (RS-2020-II201373 and RS-2024-00336673), in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2024-00348497), and in part by the Alchemist Brain to X (B2X) Project under Grant 20012355 funded by the Ministry of Trade, Industry, and Energy (MOTIE), South Korea. J.E. Sung was supported by the NRF grant funded by MSIT (2022R1A2C2005062).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109090.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg, Silent speech interfaces, Speech Commun. 52 (4) (2010) 270–287, https://doi.org/10.1016/j.specom.2009.08.002.

[2] G. Meltzner, J. Heaton, Y. Deng, G. Luca, S. Roy, J. Kline, G. Luca, S. Roy, J. Kline, Silent speech recognition as an alternative communication device for persons with laryngectomy, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (12) (2017) 2386–2398, https://doi.org/10.1109/TASLP.2017.2740000.

[3] H.-S. Cha, W.-D. Chang, C.-H. Im, Deep-learning-based real-time silent speech recognition using facial eelectromyogram recorded around eyes for hands-free interfacing in a virtual reality environment, Virtual Real. 26 (3) (2022) 1047–1057, https://doi.org/10.1007/s10055-021-00616-0.

[4] L. Pandey, A.S. Arif, LipType: a silent speech recognizer augmented with an independent repair model, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), Yokohama, Japan, 2021, pp. 1–19, https://doi.org/10.1145/3411764.3445565.

[5] Y. Deng, G. Colby, J.T. Heaton, G.S. Meltzner, Signal processing advances for the MUTE sEMG-based silent speech recognition system, in: IEEE Military Communications Conference (MILCOM), IEEE, Orlando, FL, USA, 2012, pp. 1–6, https://doi.org/10.1109/MILCOM.2012.6415781.

[6] L. Pandey, K. Hasan, A.S. Arif, Acceptability of speech and silent speech input methods in private and public, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), Yokohama, Japan, 2021, pp. 1–13, https://doi.org/10.1145/3411764.3445430.

[7] K. Sun, C. Yu, W. Shi, L. Liu, Y. Shi, Lip-interact: improving mobile device interaction with silent speech commands, in: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, Association for Computing Machinery (ACM), Berlin, Germany, 2018, pp. 581–593, https://doi.org/10.1145/3242587.3242599.

[8] G. Gosztolya, P. Á, L. Tóth, T. Grósz, A. Markó, T.G. Csapó, Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, Budapest, Hungary, 2019, pp. 1–8, https://doi.org/10.1109/IJCNN.2019.8852153.

[9] J.A. Gonzalez, L.A. Cheah, J.M. Gilbert, J. Bai, S.R. Ell, P.D. Green, R.K. Moore, A silent speech system based on permanent magnet articulography and direct synthesis, Comput. Speech Lang 39 (2016) 67–87, https://doi.org/10.1016/j.csl.2016.02.002.

[10] B. Cao, N. Sebkhi, T. Mau, O.T. Inan, J. Wang, Permanent magnetic articulograph (PMA) vs electromagnetic articulograph (EMA) in articulation-to-speech synthesis for silent speech interface, in: Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 17–23, https://doi.org/10.18653/v1/W19-1703.

[11] G.S. Meltzner, J.T. Heaton, Y. Deng, G. De Luca, S.H. Roy, J.C. Kline, Development of sEMG sensors and algorithms for silent speech recognition, J. Neural. Eng. 15 (4) (2018) 046031, https://doi.org/10.1088/1741-2552/aac965.

[12] C. Herff, T. Schultz, Automatic speech recognition from neural signals: a focused review, Front. Neurosci. 10 (2016) 429, https://doi.org/10.3389/fnins.2016.00429.

[13] Y. Yu, A.H. Shandiz, L. Tóth, Reconstructing speech from real-time articulatory MRI using neural vocoders, in: 2021 29th European Signal Processing Conference (EUSIPCO), Virtual: IEEE, 2021, pp. 945–949, https://doi.org/10.23919/EUSIPCO54536.2021.9616153.

[14] S. Stone, P. Birkholz, Silent-speech command word recognition using electro-optical stomatography, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), International Speech Communication Association (ISCA), San Francisco, CA, USA, 2016, pp. 2350–2351.

[15] T. Schultz, M. Wand, T. Hueber, D.J. Krusienski, C. Herff, J.S. Brumberg, Biosignal-based spoken communication: a survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (12) (2017) 2257–2271, https://doi.org/10.1109/TASLP.2017.2752365.

[16] D. Ferreira, S. Silva, F. Curado, A. Teixeira, Exploring silent speech interfaces based on frequency-modulated continuous-wave radar, Sensors 22 (2) (2022) 649, https://doi.org/10.3390/s22020649.

[17] P. Birkholz, S. Stone, K. Wolf, D. Plettemeier, Non-invasive silent phoneme recognition using microwave signals, IEEE/ACM Transactions on Audio, Speech, Language Processing 26 (12) (2018) 2404–2411, https://doi.org/10.1109/TASLP.2018.2865609.

[18] M.R. Sobhani, H.E. Ozum, G.G. Yaralioglu, A.S. Ergun, A. Bozkurt, Portable low cost ultrasound imaging system, in: IEEE International Ultrasonics Symposium (IUS), IEEE, Tours, France, 2016, pp. 1–4, https://doi.org/10.1109/ULTSYM.2016.7728837.

[19] R. Wang, Z. Fang, J. Gu, Y. Guo, S. Zhou, Y. Wang, C. Chang, J. Yu, High-resolution image reconstruction for portable ultrasound imaging devices, EURASIP J. Appl. Signal Process. 2019 (1) (2019) 56, https://doi.org/10.1186/s13634-019-0649-x.
[20] J. He, D. Zhang, N. Jiang, X. Sheng, D. Farina, X. Zhu, User adaptation in long-term, open-loop myoelectric training: implications for EMG pattern recognition in prosthesis control, J. Neural. Eng. 12 (4) (2015) 046005, https://doi.org/10.1088/1741-2560/12/4/046005.
[21] J. Kwon, H. Nam, Y. Chae, S. Lee, I.Y. Kim, C.-H. Im, Novel three-Axis accelerometer-based silent speech interface using deep neural network, Eng. Appl. Artif. Intell. 120 (2023) 105909, https://doi.org/10.1016/j.engappai.2023.105909.
[22] H. Akbari, H. Arora, L. Cao, N. Mesgarani, Lip2Audspec: speech reconstruction from silent lip movements video, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, Alberta, Canada, 2018, pp. 2516–2520, https://doi.org/10.1109/ICASSP.2018.8461856.
[23] T. Chi, P. Ru, S.A. Shamma, Multiresolution spectrotemporal analysis of complex sounds, J. Acoust. Soc. Am. 118 (2) (2005) 887–906, https://doi.org/10.1121/1.1945807.
[24] N. Kimura, M. Kono, J. Rekimoto, SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), Glasgow, Scotland, UK, 2019, p. 146, https://doi.org/10.1145/3290605.3300376.
[25] S. Stone, P. Birkholz, Cross-speaker silent-speech command word recognition using electro-optical stomatography, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Barcelona: IEEE, 2020, pp. 7849–7853, https://doi.org/10.1109/ICASSP40776.2020.9053447.
[26] F. Taguchi, T. Kaburagi, Articulatory-to-speech conversion using Bi-directional long short-term memory, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), International Speech Communication Association (ISCA), Hyderabad, India, 2018, pp. 2499–2503, https://doi.org/10.21437/Interspeech.2018-999.
[27] M. Morise, F. Yokomori, K. Ozawa, WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE Trans. Info Syst. 99 (7) (2016) 1877–1884, https://doi.org/10.1587/transinf.2015EDP7457.
[28] M. Janke, L. Diener, EMG-to-Speech: direct generation of speech from facial electromyographic signals, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (12) (2017) 2375–2385, https://doi.org/10.1109/TASLP.2017.2738568.
[29] S. Imai, Cepstral analysis synthesis on the Mel frequency scale, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Boston, Massachusetts, USA, 1983, pp. 93–96, https://doi.org/10.1109/ICASSP.1983.1172250.
[30] A.v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: a generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016). https://arxiv.org/abs/1609.03499.
[31] R. Prenger, R. Valle, B. Catanzaro, Waveglow: a flow-based generative network for speech synthesis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, United Kingdom, 2019, pp. 3617–3621, https://doi.org/10.1109/ICASSP.2019.8683143.
[32] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W.Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, A.C. Courville, Melgan: generative adversarial networks for conditional waveform synthesis, in: Advances in Neural Information Processing Systems (NIPS), 2019. Vancouver, Canada.
[33] J. Kong, J. Kim, J. Bae, Hifi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis, in: Advances in Neural Information Processing Systems (NIPS), Virtual, 2020, pp. 17022–17033.
[34] M. Kim, B. Cao, T. Mau, J. Wang, Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (12) (2017) 2323–2336, https://doi.org/10.1109/TASLP.2017.2758999.
[35] U. Asgher, K. Khalil, M.J. Khan, R. Ahmad, S.I. Butt, Y. Ayaz, N. Naseer, S. Nazir, Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface, Front. Neurosci. 14 (2020) 584, https://doi.org/10.3389/fnins.2020.00584.
[36] Y. Bin, Y. Yang, F. Shen, N. Xie, H.T. Shen, X. Li, Describing video with attention-based bidirectional LSTM, IEEE Trans. Cybern. 49 (7) (2019) 263–2641, https://doi.org/10.1109/TCYB.2018.2831447.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NIPS), 2017. Long Beach, CA, USA.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16 x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, https://arxiv.org/abs/2010.11929, 2020.
[39] J. Clauwaert, W. Waegeman, Novel transformer networks for improved sequence labeling in genomics, IEEE ACM Trans. Comput. Biol. Bioinf 19 (1) (2022) 97–106, https://doi.org/10.1109/TCBB.2020.3035021.
[40] R. Song, X. Zhang, X. Chen, X. Chen, X. Chen, S. Yang, E. Yin, Decoding silent speech from high-density surface electromyographic data using transformer, Biomed. Signal Process Control 80 (2023) 104298, https://doi.org/10.1016/j.bspc.2022.104298.
[41] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, Conformer: convolution-augmented transformer for speech recognition, arXiv preprint arXiv:2005.08100, https://arxiv.org/abs/2005.08100, 2020.
[42] K. Park, KSS dataset: Korean single speaker speech dataset. https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset, 2018.
[43] Y.Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E.Z. Yang, J. Lian, J. Hwang, J. Chen, P. Goldsborough, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, Torchaudio: building blocks for audio and speech processing, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, Alberta, Canada, 2022, pp. 6982–6986, https://doi.org/10.1109/ICASSP43922.2022.9747236.
[44] A. Abdullah, K. Chemmangat, A computationally efficient sEMG-based silent speech interface using channel reduction and decision tree-based classification, Procedia Comput. Sci. 171 (2020) 120–129, https://doi.org/10.1016/j.procs.2020.04.013.
[45] H. Li, H. Lin, Y. Wang, H. Wang, M. Zhang, H. Gao, Q. Ai, Z. Luo, G. Li, Sequence-to-Sequence voice reconstruction for silent speech in a tonal language, Brain Sci. 12 (7) (2022) 818, https://doi.org/10.3390/brainsci12070818.
[46] Z. Guo, P. Liu, J. Yang, Y. Hu, Multivariate time series classification based on MCNN-LSTMS network, in: Proceedings of the 2020 12th International Conference on Machine Learning and Computing (ICMLC), Association for Computing Machinery (ACM), Shenzhen, China, 2020, pp. 510–517, https://doi.org/10.1145/3383972.3384013.
[47] J. Kwon, C.-H. Im, Subject-independent functional near-infrared spectroscopy-based brain–computer interfaces based on convolutional neural networks, Front. Hum. Neurosci. 15 (2021), https://doi.org/10.3389/fnhum.2021.646915.
[48] A. Shoeibi, D. Sadeghi, P. Moridian, N. Ghassemi, J. Heras, R. Alizadehsani, A. Khadem, Y. Kong, S. Nahavandi, Y.-D. Zhang, Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models, Front. Neuroinf. 15 (2021), https://doi.org/10.3389/fninf.2021.777977.
[49] Y. Yang, P. Wang, D. Wang, A conformer-based acoustic model for robust automatic speech recognition, arXiv preprint arXiv:2203.00725 (2022). https://arxiv.org/abs/2203.00725.
[50] M. Burchi, V. Vielzeuf, Efficient conformer: progressive downsampling and grouped attention for automatic speech recognition, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Cartagena, Colombia, 2021, pp. 8–15, https://doi.org/10.1109/ASRU51503.2021.9687874.
[51] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017). https://arxiv.org/abs/1711.05101.
[52] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016). https://arxiv.org/abs/1608.03983.
[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: Advances in Neural Information Processing Systems (NIPS), 2017. Long Beach, CA, USA.
[54] L. Diener, M.R. Vishkasougheh, T. Schultz, CSL-EMG_Array: an open access corpus for EMG-to-speech conversion, in: Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), International Speech Communication Association (ISCA), Shanghai, China, 2020, pp. 3745–3749, https://doi.org/10.21437/Interspeech.2020-2859.
[55] Y.W. Chen, K.H. Hung, S.Y. Chuang, J. Sherman, W.C. Huang, X. Lu, Y. Tsao, EMA2S: an end-to-end multimodal articulatory-to-speech system, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, Daegu, Korea, 2021, pp. 1–5, https://doi.org/10.1109/ISCAS51556.2021.9401485.
[56] L.C. Chen, P.H. Chen, R.T.H. Tsai, Y. Tsao, EPG2S: speech generation and speech enhancement based on electropalatography and audio signals using multimodal learning, IEEE Signal Process. Lett. 29 (2022) 2582–2586, https://doi.org/10.1109/LSP.2022.3184636.
[57] M. Zhu, H. Zhang, X. Wang, X. Wang, Z. Yang, C. Wang, O.W. Samuel, S. Chen, G. Li, Towards optimizing electrode configurations for silent speech recognition based on high-density surface electromyography, J. Neural. Eng. 18 (1) (2021) 016005, https://doi.org/10.1088/1741-2552/abca14.
[58] J. Zhuang, M. Zhu, X. Wang, D. Wang, Z. Yang, X. Wang, L. Qi, S. Chen, G. Li, Comparison of contributions between facial and neck muscles for speech recognition using high-density surface electromyography, in: 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), IEEE, Tianjin, China, 2019, pp. 1–5, https://doi.org/10.1109/CIVEMSA45640.2019.9071636.
[59] J. Wu, Y. Zhang, L. Xie, Y. Yan, X. Zhang, S. Liu, X. An, E. Yin, D. Ming, A novel silent speech recognition approach based on parallel inception convolutional neural network and Mel frequency spectral coefficient, Front. Neurorob. 16 (2022), https://doi.org/10.3389/fnbot.2022.971446.
[60] K. Chola Raja, S. Kannimuthu, Deep learning-based feature selection and prediction system for autism spectrum disorder using a hybrid meta-heuristics approach, J. Intell. Fuzzy Syst. 45 (1) (2023) 797–807, https://doi.org/10.3233/JIFS-223694.
[61] S. Pragadeeswaran, S. Kannimuthu, Cosine deep convolutional neural network for Parkinson's disease detection and severity level classification using hand drawing spiral image in IoT platform, Biomed. Signal Process Control 94 (2024) 106220, https://doi.org/10.1016/j.bspc.2024.106220.
[62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014. Montreal, Quebec.
[63] J. Kwon, C.-H. Im, Novel signal-to-signal translation method based on StarGAN to generate artificial EEG for SSVEP-based brain-computer interfaces, Expert Syst. Appl. 203 (2022) 117574, https://doi.org/10.1016/j.eswa.2022.117574.

[64] F. Fahimi, S. Dosen, K.K. Ang, N. Mrachacz-Kersting, C. Guan, Generative adversarial networks-based data augmentation for brain-computer interface, IEEE Transact. Neural Networks Learn. Syst. 3 (9) (2020) 4039–4051, https://doi.org/10.1109/TNNLS.2020.3016666.

[65] Y. Luo, B.-L. Lu, EEG data augmentation for emotion recognition using a conditional wasserstein GAN, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Honolulu, Hawaii, USA, 2018, pp. 2535–2538, https://doi.org/10.1109/EMBC.2018.8512865.

[66] C.M. Wong, Z. Wang, A.C. Rosa, C.P. Chen, T.-P. Jung, Y. Hu, F. Wan, Transferring subject-specific knowledge across stimulus frequencies in SSVEP-based BCIs, IEEE Trans. Autom. Sci. Eng. 18 (2) (2021) 552–563, https://doi.org/10.1109/TASE.2021.3054741.

[67] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2021) 43–76, https://doi.org/10.1109/JPROC.2020.3004555.

[68] K. Zhao, Y. Zhao, R. Qian, C. Ye, Recent progress on tattoo-like electronics: from materials and structural designs to versatile applications, Chem. Eng. J. 477 (2023) 147109, https://doi.org/10.1016/j.cej.2023.147109.

[69] W. Dong, H. Zhang, H. Liu, T. Chen, L. Sun, A super-flexible and high-sensitive epidermal sEMG electrode patch for silent speech recognition, in: 2019 IEEE 32nd International Conference on Micro Electromechanical Systems (MEMS), IEEE, Seoul, Korea, 2019, pp. 565–568, https://doi.org/10.1109/MEMSYS.2019.8870672.

[70] H. Liu, W. Dong, Y. Li, F. Li, J. Geng, M. Zhu, T. Chen, H. Zhang, L. Sun, C. Lee, An epidermal sEMG tattoo-like patch as a new human–machine interface for patients with loss of voice, Microsystems & Nanoengineering 6 (1) (2020) 1–13, https://doi.org/10.1038/s41378-019-0127-5.