EMGSE: ACOUSTIC/EMG FUSION FOR MULTIMODAL SPEECH ENHANCEMENT

Kuan-Chen Wang[1] *Kai-Chun Liu*[2] *Hsin-Min Wang*[3] *Yu Tsao*[2]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan

R10942076@ntu.edu.tw, t22302856@citi.sinica.edu.tw, whmat@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw

ABSTRACT

Multimodal learning has been proven to be an effective method to improve speech enhancement (SE) performance, especially in challenging situations such as low signal-to-noise ratios, speech noise, or unseen noise types. In previous studies, several types of auxiliary data have been used to construct multimodal SE systems, such as lip images, electropalatography, or electromagnetic midsagittal articulography. In this paper, we propose a novel EMGSE framework for multimodal SE, which integrates audio and facial electromyography (EMG) signals. Facial EMG is a biological signal containing articulatory movement information, which can be measured in a non-invasive way. Experimental results show that the proposed EMGSE system can achieve better performance than the audio-only SE system. The benefits of fusing EMG signals with acoustic signals for SE are notable under challenging circumstances. Furthermore, this study reveals that cheek EMG is sufficient for SE.

Index Terms— Non-invasive, multimodal, electromyography, speech enhancement, deep neural network

1. INTRODUCTION

It is inevitable that speech signals will be contaminated by certain types of ambient noise in daily life. To recover clean speech signals from noisy signals, speech enhancement (SE) is applied to improve the speech quality and intelligibility. SE is critical for various speech-related applications and increases their robustness in real-world environments, such as automatic speech recognition (ASR) [1, 2], speaker recognition [3, 4], hearing aids [5] and cochlear implants [6]. Many methods have been developed to conduct SE, including the spectral subtraction [7], Wiener filtering [8], minimum mean square error (MMSE) estimator [9] and Kalman filtering [10]. Recently, neural network (NN) based methods have been widely applied in this research area owing to the outstanding nonlinear mapping capability of an NN. Different types of deep learning models have been applied to SE, such as fully connected neural networks [11, 12], convolutional neural networks (CNNs) [13], a fully convolutional network (FCN) [14, 15], a recurrent neural network (RNN) [16] and long short-term memory (LSTM) models [17]. Some approaches combine multiple types of models to better extract spatial and temporal information [18]. Moreover, several advanced network structures or designs have provided further improvements in SE. Notable examples are generative adversarial networks [19] and attention-based networks [20]. These approaches are verified to outperform conventional SE algorithms in terms of their performance and robustness.

Although NN-based methods have achieved a significant improvement and have become mainstream in SE, some challenging conditions, such as low signal-to-noise ratios (SNRs) or speech noise corruption, can still compromise the effectiveness of an audio-only SE system. To deal with such challenges, numerous studies have adopted auxiliary data in SE. These data offer information on articulatory motion in different aspects without contamination by air background noise. For instance, lip images [21], bone-conducted microphone signals [22], electropalatography (EPG) [23], and electromagnetic midsagittal articulography (EMMA) [24] have all been verified to be feasible for constructing multimodal SE systems. However, these data have certain disadvantages. For example, the quality of the lip images may be affected by background scenes and quick head movements. Some types of articulatory data (e.g., EPG or EMMA) that install sensors within the mouth may cause user discomfort. To address these disadvantages, we adopted facial electromyography (EMG) approach for the proposed SE system. Facial EMG measures the activation potentials of the human articulatory muscles by attaching single or array electrodes to specific places such as the cheek and chin. EMG sensors are more comfortable for users because they can be attached to the surface of the skin and record the signals noninvasively.

In this paper, we propose a novel multimodal SE system that fuses facial EMG and audio signals for SE, namely EMGSE. Experiments were conducted on the data of an open-access corpus, CSL-EMG_Array [25]. A baseline system with audio-only SE was applied for comparison with EMGSE, and the performance was evaluated using objective metrics. Experiment results show that the performance of the EMGSE system surpasses that of the audio-only SE system. Furthermore, the EMGSE is more robust under difficult conditions, including low SNRs and speech noise. Therefore, the EMGSE can be applied in scenarios where noisy speech signals are still available and clear communication is necessary for certain purposes [24], such as security guards or field agents performing duties in the site of a famous sports event.

The remainder of this paper is organized as follows. Section 2 introduces related studies conducted in this area. Section 3 presents the architecture of the proposed EMGSE system. Section 4 presents the experimental details and results used to illustrate the performance of the proposed method. Finally, Section 5 provides some concluding remarks regarding this research.

2. RELATED WORK

2.1. Neural-network-based SE

Currently, NN-based methods dominate the research field of SE owing to the powerful nonlinear mapping capability of deep learning models. In general, these methods can be categorized into two types: masking-based SE and mapping-based SE. Masking-based SE sup-

presses the noise components by estimating a mask and applying it to corrupted speech [26]; whereas mapping-based SE directly estimates clean speech from noisy speech. According to the input and output data type of mapping-based SE, two types of SE methods can be derived: waveform-mapping- and spectral-mapping-based SE. Several multimodal SE systems have achieved good performance when using the spectral-mapping-based method [23, 24]. Thus, the proposed EMGSE also applies a spectral-mapping-based SE method, and its network is constructed with rectified linear units (RELU), fully connected (FC) layers, and bidirectional long short-term memory (BLSTM). The properties of the BLSTM model are briefly described in the next paragraph.

BLSTM has been proven to be a powerful model for sequential learning tasks. It can utilize both the past and future contexts of the data using two LSTM hidden layers. One of the layers processes the data sequence forward, and the other processes it backward. In addition, LSTM can cope with the gradient vanishing and gradient exploding problems. Owing to its capabilities and advantages, BLSTM has been applied in many speech-related applications, such as speech recognition [1], speaker recognition [27] and SE [17, 20, 24, 23]. The following equations show the operation of the LSTM network used in EMGSE:

$$i_t = \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}\, h_{t-1} + b_{hi}\right), \quad (1)$$
$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}\, h_{t-1} + b_{hf}\right), \quad (2)$$
$$g_t = \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}\, h_{t-1} + b_{hg}\right), \quad (3)$$
$$o_t = \sigma\left(W_{io}x_t + b_{io} + W_{ho}\, h_{t-1} + b_{ho}\right), \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad (5)$$
$$h_t = o_t \odot \tanh(c_t), \quad (6)$$

where x is the input; i, f, o, and g are the input, forget, output, and cell gates, respectively; b is the bias vector; c is the cell state vector; and h_t and h_{t-1} are the hidden state vectors at times t and $t-1$, respectively. In addition, $\odot$ denotes the Hadamard product (element-wise product), and σ is the sigmoid activation function.

2.2. Multimodal SE

Multimodal SE incorporates different articulatory data with noisy speech signals to improve the robustness of SE systems. Various data types have been used as auxiliary information for multimodal SE systems, including visual cues [15, 21], bone-conducted microphone recorded waveforms [22], EMMA [24] and EPG [23]. Moreover, different fusion strategies were used in these studies. In [24], unilateral concatenating exhibited the best performance, whereas in [23], early fusion performed better than late fusion on SE. However, the late fusion strategy has been proven to be more effective under many different scenarios [21, 22]. Each data type is input into the network separately and subsequently merged into a single vector after encoding. In this study, the proposed EMGSE utilizes a late fusion strategy to combine EMG signals and a noisy audio spectrogram for SE.

2.3. Speech-related tasks and open access corpus of facial EMG

EMG signals can indicate muscle activity. Such signals can be measured by electrodes attached to the skin without any invasive sensors or complex installations. Many studies have demonstrated the feasibility of using facial EMG in speech-related applications, such as speech recognition [28, 29] and generation [30, 31]. Thus, this study applies EMG signals to the proposed multimodal SE system as auxiliary data. The experimental data used in this study are the

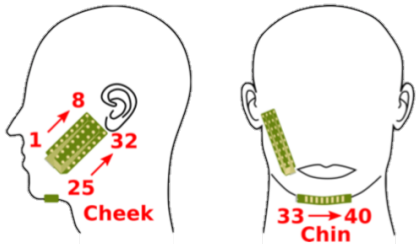

Fig. 1. Placement of EMG array electrodes in CSL-EMG_Array corpus [25]. There are 40 channels of EMG signals recorded during the speech.

CSL-EMG_Array corpus provided by Diener et al [25]. This corpus was developed mainly for EMG-to-speech studies. In this corpus, audio and EMG signals were recorded simultaneously, and the sampling rates were 16 kHz and 2048 Hz, respectively. The EMG signals were recorded with two array electrodes placed at the cheek and chin, as shown in Fig. 1 [25]. Furthermore, 5 cross-row channels were excluded in this study, and therefore only 35 EMG channels were used for SE. There were 12 sessions in the corpus, and each session involved 7 recording blocks to reflect the reality of the speech-conversion scenario. We only used the data of block 1 (containing audio and EMG data for 340 English utterances) in audible sessions from 8 speakers, as the main research topic focuses on SE.

3. PROPOSED METHODS

In this section, we describe the details of the proposed EMGSE. The feature extraction of EMG and audio is shown in Subsection 3.1. The overall network structure and fusion method of the EMGSE are illustrated in Subsection 3.2.

3.1. Feature extraction

A feature extraction process is required to extract muscle movement information because raw EMG signals are too noisy to be used directly. Some studies have shown that the time domain (TD) feature set of EMG is suitable for speech recognition and generation [25, 30, 31]. Therefore, we referred to the TD15 feature set [25] for our EMGSE. Initially, high-pass and low pass third-order Butterworth filters with a cutoff frequency of 134 Hz were employed to separate the raw EMG data into high- and low-frequency parts. A Blackman window of 32 ms length and 8 ms shift was then applied to extract the TD feature set from each frame. The feature set is calculated as follows (from left to right: mean and power of low-frequency part and absolute-value mean, power and zero-crossing rate of high-frequency part):

$$TD(x) = [\frac{1}{n}\sum_{k=1}^{n} x_{low}[k], \frac{1}{n}\sum_{k=1}^{n}(x_{low}[k])^2, \frac{1}{n}\sum_{k=1}^{n}|x_{high}[k]|, \frac{1}{n}\sum_{k=1}^{n}(x_{high}[k])^2, ZCR(x_{high})], \quad (7)$$

where x is the EMG data of a frame, and x_{low} and x_{high} are the low- and high-frequency parts of x, respectively.

To obtain the context information, 15 frames in the past and future were stacked to form a TD15 vector (with 31 frames) of an EMG channel. Then we stack TD15 vectors from 35 EMG channels to form the input vector of EMGSE with dimensions of 35 (channels)

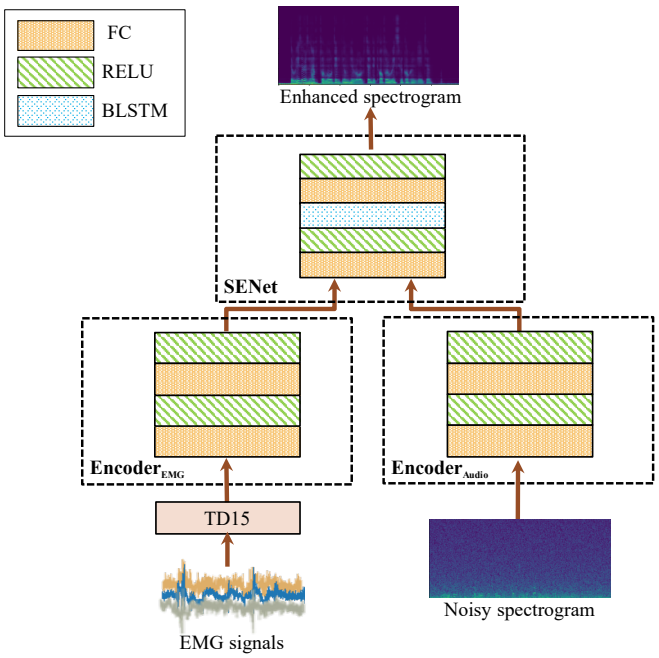

Fig. 2. Architecture of EMGSE

$\times 31$ (frames) $\times 5$ (features) $= 5,425$. Finally, we applied min-max normalization to the feature vector and made its value range from 0 to 1.

For the acoustic signal, we applied STFT to transform the noisy audio signals into the spectrogram. We also used a Blackman window of 32 ms length and 8 ms shift in STFT to match the condition of the EMG TD features. Finally, log1p and min-max normalization were applied to the spectrogram.

3.2. EMGSE network structure

Fig. 2 shows the overall network structure of EMGSE. Initially, the EMG and audio encoders compress both feature vectors into 100 dimensions. Both encoders are constructed through 2 FC layers (with 200 and 100 dimensions, respectively) using RELU. A dropout of 0.5 is added in all layers of EMG encoders to compensate for the relatively small EMG dataset. The SE network (SENet) is then further applied to transform the latent vector into a clean audio spectrogram. Two encoded vectors are initially fused into a latent vector with dimensions of 200 through an FC layer using RELU. The remaining parts of SENet are a BLSTM layer with 2 hidden layers (output dimensions of 500) and an output FC layer (257 dimensions) with RELU. Finally, we combine the output spectrogram with the phase component of the corresponding noisy audio and conduct an inverse STFT to reconstruct the enhanced audio waveform. This process can be expressed as follows:

$$v_{EMG}[n] = \text{Encoder}_{EMG}\{x^{EMG}[n]\}, \quad (8)$$

$$v_{Audio}[n] = \text{Encoder}_{Audio}\{x^{Audio}[n]\}, \quad (9)$$

$$z[n] = \text{SENet}\{v_{EMG}[n], v_{Audio}[n]\}, \quad (10)$$

where $x^{EMG}[n]$ and $x^{Audio}[n]$ are the EMG feature vector and noisy spectrogram at time n, respectively; $v_{EMG}[n]$ and $v_{Audio}[n]$ are the encoded vectors; and $z[n]$ is the enhanced spectrogram at time n.

4. EXPERIMENTS

4.1. Experiment setup

As mentioned in Section 2, this study uses the corpus CSL-EMG_Array data in block 1 from 8 speakers (340 utterances per speaker) to validate EMGSE. A total of 340 utterances were separated into 280, 20, and 40 utterances for training, validation, and testing, respectively. For the training and validation sets, we applied 100 types of nonspeech noise [32] to generate noisy audio data. Each utterance is corrupted with five randomly selected types of noise at five SNRs (-10, -5, 0, 5, and 10 dB). For the test set, we added 18 unseen noise types (car noise, engine noise, pink noise, white noise, two types of street noises, six kinds of background Chinese speakers, and six kinds of English speakers) to clean utterances at four SNRs (-11, -4, -1, and 4 dB) to cause mismatch conditions. We evaluated the performance of SE using two evaluation criteria: perceptual evaluation of speech quality (PESQ) [33] and short-time objective intelligibility (STOI) [34]. The PESQ score ranges from 0.5 to 4.5, and the STOI score ranges from 0 to 1. Higher PESQ and STOI scores represent better speech quality and intelligibility, respectively.

4.2. Implementation details

The proposed EMGSE used the L1 loss and Adam optimizer to update the weights. The learning rate was set to 0.0001. To avoid an overfitting, we stopped training as the validation loss stopped dropping after 15 epochs and saved the network parameters with the least validation loss. We built a baseline system, i.e., SE with audio only (SE(A)) for comparison with EMGSE. SE(A) has an identical structure to EMGSE, except that only acoustic signals are used for the input data.

4.3. Results and discussion

Fig. 3 shows the overall performance of the two SE systems for the eight speakers. It can be seen that EMGSE achieves higher PESQ and STOI scores than SE(A) for all speakers. With facial EMG, PESQ scores can increase from 0.1 to 0.3, and STOI scores can increase by approximately 0.03 to 0.1 from case to case. To further examine the effect of facial EMG on the SE, Table 1 presents the average performance for eight speakers under four SNRs (-11, -4 , -1, and 4 dB). The results show that EMGSE can perform particularly better than SE(A) for a low SNR (-11 dB). The PESQ and STOI scores increased by 0.225 and 0.097, respectively. Furthermore, Table 2 demonstrates the average performance of the two systems for different noise types. We can observe that the performance of EMGSE outperforms that of SE(A), particularly in speech noise types. Overall, the improvement with facial EMG increases as the noise conditions become tougher, and the analysis results validate that facial EMG is beneficial to multimodal SE[1].

To study the improvement produced in the proposed system more thoroughly, we inspect the latent space and the output vector of the fusion layer using 200 dimensions in EMGSE. Fig. 4 shows the latent spaces in audio-input-only (clean speech and noisy speech with an SNR level of -11 dB) and the EMG-audio-input case. Comparing the latent spaces of the audio-input-only cases, we can see that patterns appearing in the speech occurrence are much clearer in the clean-speech-input case. After the EMG data were applied, some patterns became more protruding. To evaluate the improved quality of the latent space, we plot the differences between the latent

[1]Noise data used in testing and the demonstration of proposed EMGSE can be viewed in https://eric-wang135.github.io/EMGSE/

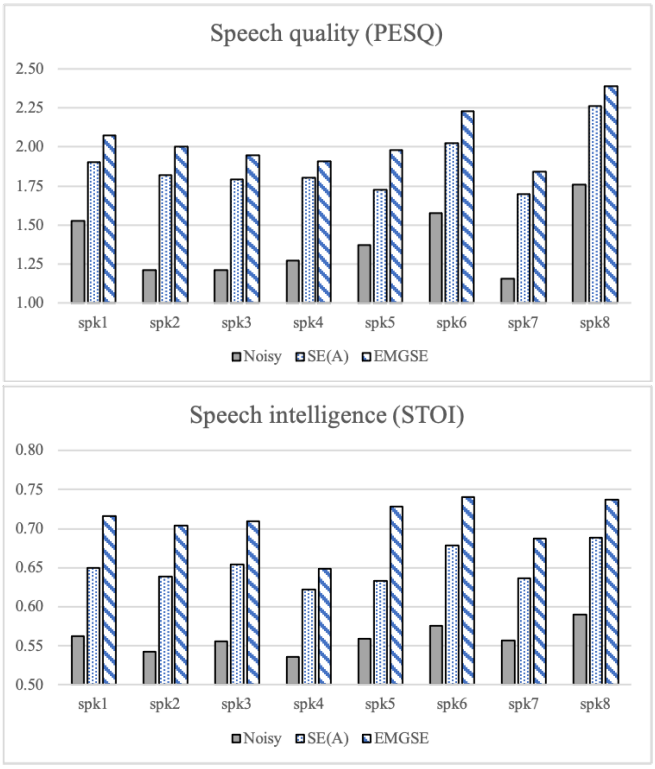

Fig. 3. Performance of EMGSE and audio-only SE system (SE(A)) and noisy test data evaluated based on PESQ and STOI scores. Data on 8 speakers were tested, and the application of facial EMG improved the SE performance in every case.

Table 1. Evaluation results on SE(A) and EMGSE under different SNRs.

	Noisy		SE(A)		EMGSE	
	PESQ	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**
-11dB	0.923	0.394	1.197	0.446	**1.452**	**0.553**
-6dB	1.138	0.481	1.608	0.579	1.829	0.658
-1dB	1.448	0.579	2.03	0.697	2.207	0.751
4dB	1.722	0.663	2.333	0.764	2.476	0.801
Avg.	1.308	0.529	1.792	0.621	1.991	0.691

Table 2. Evaluation results on SE(A) and EMGSE for different noise types.

	Noisy		SE(A)		EMGSE	
	PESQ	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**
Chinese	1.277	0.504	1.695	0.591	**1.928**	**0.669**
English	1.290	0.514	1.677	0.584	**1.935**	**0.676**
Car	1.692	0.675	2.297	0.777	2.314	0.781
Engine	1.286	0.55	1.985	0.686	2.094	0.729
Pink	1.281	0.552	1.943	0.658	2.077	0.713
White	1.303	0.585	2.072	0.691	2.223	0.737
Street	1.404	0.564	2.028	0.704	2.129	0.743

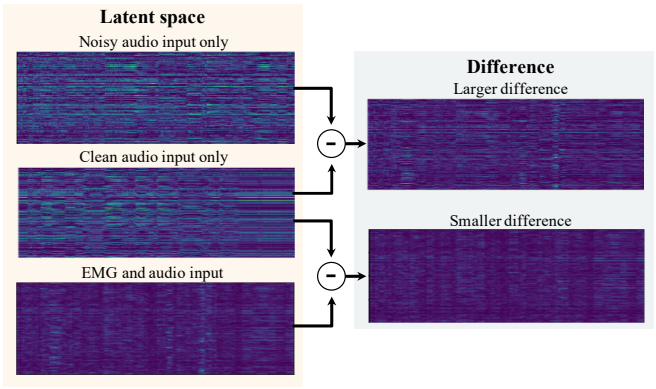

Fig. 4. The latent spaces of EMGSE with different input data conditions and their difference. The latent space of EMGSE in clean-audio-input-only case can be viewed as a reference for better SE results. When facial EMG is combined with noisy audio input, the difference between the latent space and the reference minimizes, demonstrating the benefits of facial EMG.

Table 3. Performance of SE(A) and EMGSE with only 28 and 35 channels. A reduction in the number of channels seems to have little influence.

	SE(A)	**EMGSE**	**EMGSE$_{\text{cheek}}$**
PESQ	1.879	2.046	2.039
STOI	0.65	0.709	0.711

space of a clean-audio input and those with and without an EMG input. It can be observed that the difference minimizes after the EMG is adopted. These results again verified the effectiveness of facial EMG on the SE.

We explored the SE performance using only EMG sensors placed on the cheek (28 EMG channels). Table 3 shows the average performance of SE(A) and EMGSE with 28 and 35 channels for the 8 speaker. To our surprise, a reduction in the number of channels had little effect on the performance. The PESQ score only decreased by approximately 0.007 and the STOI score remained almost unchanged. Therefore, it can be deduced that the cheek EMG is sufficient for SE. Reducing the number of channels produces some advantages. We can decrease the computational effort and further improve the efficiency of the SE. Moreover, users may feel more comfortable if fewer sensors are needed. These benefits increase the practicability of the EMGSE system under a real-world scenario.

5. CONCLUSION

In this study, facial EMG was used as auxiliary data, and a non-invasive and speaker-dependent multimodal SE system, EMGSE, was proposed. To the best of our knowledge, this is the first study applying EMG to SE. Experimental results show that fusing EMG signals with acoustic signals can improve SE performance, especially under challenging circumstances, such as a low signal-to-noise ratio (SNR) and speech noise contamination. In addition, cheek EMG has been shown to be sufficient for SE, increasing the practicability of EMGSE. In the future, we plan to use EMG channel feature selection methods and EMG denoising algorithms as preprocessing to further enhance the effect of facial EMG in EMGSE.

6. REFERENCES

[1] Ashutosh Pandey, Chunxi Liu, Yun Wang, and Yatharth Saraf, "Dual application of speech enhancement for automatic speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 223–228.

[2] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7009–7013.

[3] Samia Abd El-Moneim, MA Nassar, Moawad I Dessouky, Nabil A Ismail, Adel S El-Fishawy, and Fathi E Abd El-Samie, "Text-independent speaker recognition using lstm-rnn and speech enhancement," *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 24013–24028, 2020.

[4] Yanpei Shi, Qiang Huang, and Thomas Hain, "Robust speaker recognition using speech enhancement and attention model," *arXiv preprint arXiv:2001.05031*, 2020.

[5] Gyuseok Park, Woohyeong Cho, Kyu-Sung Kim, and Sangmin Lee, "Speech enhancement for hearing aids with deep learning on environmental noises," *Applied Sciences*, vol. 10, no. 17, pp. 6077, 2020.

[6] Tobias Goehring, Federico Bolner, Jessica JM Monaghan, Bas Van Dijk, Andrzej Zarowski, and Stefan Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hearing research*, vol. 344, pp. 183–194, 2017.

[7] S Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1979, vol. 4, pp. 200–203.

[8] Pascal Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE, 1996, vol. 2, pp. 629–632.

[9] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[10] K Paliwal and Anjan Basu, "A speech enhancement method based on kalman filtering," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1987, vol. 12, pp. 177–180.

[11] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, vol. 2013, pp. 436–440.

[12] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[13] Szu-Wei Fu, Yu Tsao, and Xugang Lu, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Interspeech*, 2016, pp. 3768–3772.

[14] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 006–012.

[15] Rung-Yu Tseng, Tao-Wei Wang, Szu-Wei Fu, Chia-Ying Lee, and Yu Tsao, "A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[16] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152.

[17] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[18] Tsun-An Hsieh, Hsin-Min Wang, Xugang Lu, and Yu Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020.

[19] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[20] Chiang-Jen Peng, Yun-Ju Chan, Cheng Yu, Syu-Siang Wang, Yu Tsao, and Tai-Shih Chi, "Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario," *arXiv preprint arXiv:2101.02550*, 2021.

[21] Shang-Yi Chuang, Hsin-Min Wang, and Yu Tsao, "Improved lite audio-visual speech enhancement," *arXiv preprint arXiv:2008.13222*, 2020.

[22] Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[23] R.T.H. Tsai P.H. Chen and Y. Tsao, "Multimodal electropalatography-audio speech enhancement," in *submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*, 2021.

[24] Yu-Wen Chen, Kuo-Hsuan Hung, Shang-Yi Chuang, Jonathan Sherman, Xugang Lu, and Yu Tsao, "A study of incorporating articulatory movement information in speech enhancement," *arXiv preprint arXiv:2011.01691*, 2020.

[25] Lorenz Diener, Mehrdad Roustay Vishkasougheh, and Tanja Schultz, "Csl-emg array: An open access corpus for emg-to-speech conversion," *Proc. Interspeech 2020*, pp. 3745–3749, 2020.

[26] Nasir Saleem and Muhammad Irfan Khattak, "Deep neural networks for speech enhancement in complex-noisy environments.," *IJIMAI*, vol. 6, no. 1, pp. 84–90, 2020.

[27] Shaofei Xue and Zhijie Yan, "Improving latency-controlled blstm acoustic models for online speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5340–5344.

[28] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel, "Towards continuous speech recognition using surface electromyography," in *Ninth International Conference on Spoken Language Processing*, 2006.

[29] Erik J Scheme, Bernard Hudgins, and Phillip A Parker, "Myoelectric signal classification for phoneme-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 4, pp. 694–699, 2007.

[30] Matthias Janke and Lorenz Diener, "Emg-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[31] Lorenz Diener and Tanja Schultz, "Investigating objective intelligibility in real-time emg-to-speech conversion.," in *INTERSPEECH*, 2018, pp. 3162–3166.

[32] G. Hu, "100 nonspeech environmental sounds," 2004.

[33] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[34] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.