A STUDY OF INCORPORATING ARTICULATORY MOVEMENT INFORMATION IN SPEECH ENHANCEMENT

Yu-Wen Chen[1], *Kuo-Hsuan Hung*[1], *Shang-Yi Chuang*[1], *Jonathan Sherman*[1], *Xugang Lu*[2], *and Yu Tsao*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]National Institute of Information and Communications Technology, Japan

ABSTRACT

Although deep-learning algorithms have made great advances on speech enhancement (SE), SE performance is still limited against highly challenging conditions, such as unseen noise types or very low signal-to-noise ratios (SNRs). Given that the mechanisms of vocal articulation are robust or even unaffected by changes in the auditory environment, we propose a novel multimodal audio-articulatory-movement SE model (AAMSE) to improve performance in such challenging conditions. We combine articulatory movement features and audio data for both waveform-mapping-based and spectral-mapping-based SE systems with three fusion strategies. Experimental results confirm that by combining the modalities, AAMSE notably improves the SE performance in both speech quality and intelligibility compared to the audio-only SE baselines. Furthermore, AAMSE shows robust results under very low SNRs and unseen noise type conditions.

***Index Terms*—** articulatory movement, neural network, multimodal learning, speech enhancement

1. INTRODUCTION

Speech enhancement (SE) aims to improve speech quality and intelligibility by reducing noise components within distorted speech signals. SE has been commonly used as a pre-processing step of various speech-related applications, such as automatic speech recognition (ASR) [1, 2, 3], speaker recognition [4], and hearing aids [5, 6]. Recently, neural-network (NN)-based SE methods have come to dominant the research field. The deep denoising autoencoder [7, 8, 9], fully connected neural network [10, 11, 12], convolutional neural network [13, 14, 15], and long short-term memory model [16, 17, 18] are well-known SE methods that adopt NN models as the core architecture.

NN-based SE methods usually only use audio signals as the input, but the contingent weak point is that the SE performance decreases drastically when encountering unseen noise types or very low signal-to-noise-ratio (SNR) conditions. To address this issue, [19, 20] have proposed audio-visual multimodal SE systems. However, visual data introduces more limitation - only the external vocal tract (lips) are considered, greater storage and processing capacities are required, and unseen video conditions (capture quality/lighting, obstructions, facial angles, sudden movements, etc.) will limit performance similar to unseen audio - the same weakness it attempted to improve. On the other hand, articulatory movements have been confirmed to provide useful and complementary information to acoustic signals, and can be used to synthesize speech signals [21, 22].

In this study, we use Electromagnetic Midsagittal Articulography (EMMA) data as our articulatory movement features. The technology captures articulatory movements by using an electromagnetic field to induce currents in sensors which are placed on articulators, such as tongues or lips. In previous works, Wei *et al.* [23] studied the articulators' contribution during a speech. Hiroya *et al.* [24] used an HMM-based speech production model to estimate articulatory movements from speech acoustics. However, to our best knowledge, no one has tried to use articulatory movements as additional features in SE systems.

We test audio-articulatory-movement SE (AAMSE) models with three fusion strategies in both waveform-mapping-based and spectral-mapping-based SE systems. Experimental results show that the proposed AAMSE models outperform the baseline audio-only SE models, and still achieve high intelligibility even in low signal-to-noise ratio (SNR) levels.

The rest of the paper is organized as follows. Section 2 introduces the related works. The proposed articulatory movement features and the AAMSE frameworks are presented in Section 3. Experimental details and results are given in Section 4 to demonstrate the performance of the proposed approaches. Section 5 concludes our work.

2. RELATED WORKS

We implement AAMSE on one waveform-mapping-based and two spectral-mapping-based SE systems. Fully convolutional neural networks (FCN) have been confirmed to be an effective waveform-mapping-based SE model [14]. In this study, we integrate the articulatory movements in the time domain with this model. We also adopt two spectral-mapping-based models: the time delay neural network

(TDNN) [25] and bi-directional long short-term memory networks (BLSTM). The two models both consider the temporal relation within speech signals. TDNN is a fully-connected feed-forward neural network that has been proven to be powerful in handling temporal dependencies, and the BLSTM network considers both forward and backward sequences of inputs. Compared with regular feed-forward neural networks, LSTMs have feedback connections, so the BLSTM can extend attention over arbitrary time intervals and is suitable to process time series data such as speech signals and articulatory movements.

For the waveform-mapping-based systems, SE directly processes speech waveforms. For the spectral-mapping-based systems, short-time Fourier transform (STFT) and inverse STFT are applied to transform speech between waveforms and spectral features, where only the magnitude components are enhanced, while the phase components are borrowed from the original noisy speech.

3. PROPOSED AAMSE

In this section, we first explain the EMMA signals, which are used as our articulatory movement data. Then, we will introduce the proposed AAMSE system with three fusion strategies.

3.1. Characteristics of the articulatory movement data

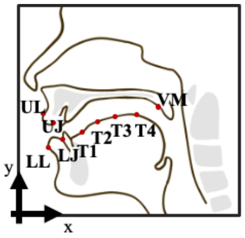

Fig. 1. Positions of the EMMA sensors.

In this study, we use EMMA (collected by NTT, Tokyo, Japan) as the articulatory movement data. The sensor coils of EMMA are placed at the upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (T1), tongue blade (T2), tongue dorsum (T3), tongue rear (T4), and velum (VM) as shown in Fig. 1. EMMA records the Cartesian coordinates of each sensor points at a sampling rate of 250 Hz. Fig. 2 shows the speech spectrograms and the EMMA signals of two speakers speaking the same utterance. Both the spectrograms and EMMA signals present similar patterns, indicating that these signals are highly-related to pronunciations.

3.2. Three fusion strategies

The goal of SE is to convert a noisy speech signal s into an enhanced speech signal $\hat{x}$ that is close to the clean speech

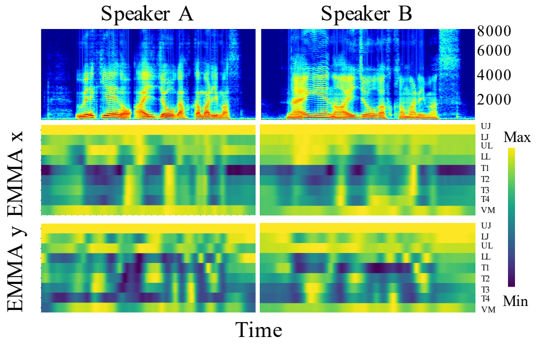

Fig. 2. Visualization of the EMMA data.

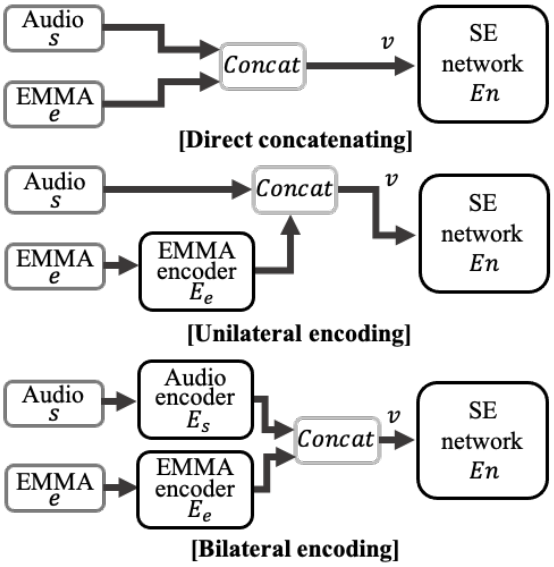

Fig. 3. The three fusion strategies.

signal x. We define s as: $s = x + n$, where x denotes the clean speech signal and n represents the noise signal.

AAMSE is a multimodal problem. The physical meaning of an audio signal is the sound intensity, while that of a articulatory movement is the trajectory of organs in the vocal tract. We reason that these different mechanisms, both containing information about speech, should improve performance over the single audio modality when combined. We test three fusion strategies: (1) direct concatenating, (2) unilateral encoding, and (3) bilateral encoding to integrate audio and articulatory movement data. Fig. 3 illustrates the structure of the three fusion strategies. The audio speech signal and the EMMA signal are denoted as s and e, and v is the input of the SE model. We aim to find an audio encoder E_s, an EMMA encoder E_e, and a SE network En such that the enhanced signal $\hat{x} = En(v)$ is close to the clean signal x.

- Direct concatenating:

$$v = Concat(s, e) \tag{1}$$

- Unilateral encoding:

$$v = Concat(s, E_e(e)) \tag{2}$$

- Bilateral encoding:

$$v = Concat(E_s(s), E_e(e)) \tag{3}$$

4. EXPERIMENTS

4.1. Experimental setup

The EMMA dataset contains articulatory and speech signals from 3 speakers, each providing 354 utterances. The two types of signals are recorded at the same time with a sampling rate of 250Hz for EMMA and 16kHz for speech. We align the two signals by upsampling the EMMA signals to 16 kHz. The training set includes 304 utterances from each speaker, and the testing set includes the remaining 50 utterances. We use 100 different noises [26] to prepare the noisy training data at 8 SNRs (±1dB, ±4dB, ±7dB, and ±10dB). Each clean utterance is contaminated by 5 randomly selected noises at the 8 SNRs. In testing data, each clean utterance is corrupted by 7 noises (car noise, engine noise, pink noise, white noise, background talkers, and two kinds of street noises) at 6 SNRs (-8dB, -5 dB, -2dB, 0dB, 2dB and 5dB).

Experimental results are evaluated using PESQ [27] and STOI [28] for speech quality and intelligibility, respectively. We also test our results on a pre-trained ASR system [29] and calculate the character correct rate (CCR) by Levenshtein distance [30].

4.2. Implementation details

The structures of the waveform-mapping-based and the spectral-mapping-based SE systems are shown in Table 1. All waveform-mapping-based FCN [14] models are trained with L2 loss and the Adam optimizer [31] with a learning rate of 0.001. For the spectral-mapping-based models, we use STFT with window size 512, hop length 128 and log1p magnitude spectrograms [32] as our audio input feature. All spectral-mapping-based TDNN [25] and BLSTM models are trained with L1 loss and Adam optimizer [31] with a learning rate of 0.0001. For each SE model, we keep the same SE network structure under the audio-only condition and the audio-articulatory-movement condition with the fusion strategy of direct concatenating.

4.3. Experimental results

The spectrograms of the enhanced audio signals in Fig. 4 show that all the models reduce distortions. However, the results of the AAMSE models are better than the audio-only SE baselines, which can be observed in the silent region.

The PESQ and STOI of the original noisy speech and that of the audio-only baselines are shown in Table 2. All the waveform-mapping-based and spectral-mapping-based audio-only SE systems yield higher scores than the original noisy speech. Table 3 and Table 4 present the average scores (white part) and the improvement (gray part, compared to audio-only models) of PESQ and STOI. Except for the FCN with unilateral encoding, all AAMSE models achieve higher scores than the audio-only SE models. A potential reason for the poor performance of FCN with unilateral encoding is the information loss on channel reduction. SE is an audio-dominant task, so we designed the EMMA channel number lesser or equal to the audio channel number. The unilateral EMMA encoder encodes EMMA signals from 18 channels to only 1 channel, which is the same size as audio signals. Note that the bilateral EMMA encoder encodes EMMA signals to 18 channels without a channel reduction.

Fig. 5 shows the SE improvement of the best audio-only SE model (BLSTM) and the best AAMSE model (BLSTM with unilateral encoding) compared to the scores of the original noisy signals in different SNRs. The performance of both models improve in terms of PESQ and STOI, and the AAMSE

	Audio encoder	EMMA encoder	SE network
FCN			
Audio only	-	-	Conv1d(f:128, k:55)×7 Conv1d(f:1, k:55)
Direct concatenating	-	-	Conv1d(f:128, k:55)×7 Conv1d(f:1, k:55)

Unilateral encoding	-	Conv1d(f:128, k:256)
		Conv1d(f:128, k:128)
		Conv1d(f:1, k:55)
			Conv1d(f:128, k:55×4
			Conv1d(f:1, k:55)

Bilateral encoding	Conv1d(f:128, k:55)
	Conv1d(f:128, k:55)
	Conv1d(f:18, k:55)
		Conv1d(f:128, k:128)
		Conv1d(f:128, k:128)
		Conv1d(f:18, k:64)
			Conv1d(f:128, k:55)×4
			Conv1d(f:1, k:55)

TDNN			

Audio only	-	-	TDNN(257)×3
			Dense(771)
			Dense(257)
			TDNN(257)×4

Direct concatenating	-	-	TDNN(257)×3
			Dense(771)
			Dense(257)
			TDNN(257)×4

Unilateral encoding	-	TDNN(18)×2	TDNN(257)×2
			Dense(771)
			Dense(257)
			TDNN(257)×4

Bilateral encoding	TDNN(257)	TDNN(18)×2	TDNN(257)×2
			Dense(771)
			Dense(257)
			TDNN(257)×3

BLSTM			
Audio only	-	-	BLSTM(500)×3 Dense(257)
Direct concatenating	-	-	BLSTM(500)×3 Dense(257)

Unilateral encoding	-	BLSTM(36)×3
		Dense(36)×2
			BLSTM(514)×2
			BLSTM(257)
			Dense(257)

Bilateral encoding	BLSTM(257)
	Linear(257)
		BLSTM(18)×4
		Dense(18)
			BLSTM(514)×2
			BLSTM(257)
			Dense(257)

Table 1. The structures of the waveform-mapping-based and the spectral-mapping-based SE systems. In waveform-mapping-based FCN [14], f is the number of the output filters, and k is the kernel size. In spectral-mapping-based TDNN [25] and BLSTM, the numbers in brackets represent the output size.

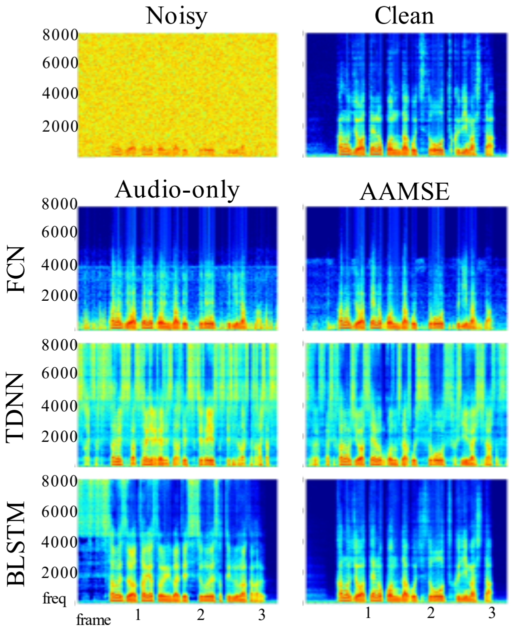

Fig. 4. Spectrograms of audio signals.

model outperforms the audio-only SE model. The CCR of the audio-only SE model decreases like the result reported in [33], while that of the AAMSE model increases, indicating that the articulatory movement features can provide more information about intelligibility. We test the performance of our best AAMSE model (BLSTM with unilateral encoding) with 4 less invasive sensors (UL, LL, LJ, and T1). Experimental results in Fig. 6 show that the AAMSE (fewer) model achieves better performance than the audio-only SE model, indicating that a lesser combination of articulatory movement features may be sufficient for SE tasks.

5. CONCLUSION

In this paper, we propose AAMSE on both waveform-mapping-based and spectral-mapping-based SE systems and investigate three fusion strategies. Experimental results show that articulatory movements can effectively improve the SE performance, and they are helpful especially in low SNRs. The contributions of this paper are twofold. First, we confirmed the effectiveness of incorporating articulatory movements into SE systems. Second, we verified that even with only 4 sensors, the extra articulatory features can still provide useful information to SE tasks. The findings from the present study is positive and serves as a useful guide to design devices to collect articulatory movement data.

	Noisy	Audio-only		
		FCN	TDNN	BLSTM
PESQ	1.530	2.311	2.064	**2.329**
STOI	0.686	**0.814**	0.738	0.801

Table 2. PESQ and STOI of different audio-only SE models.

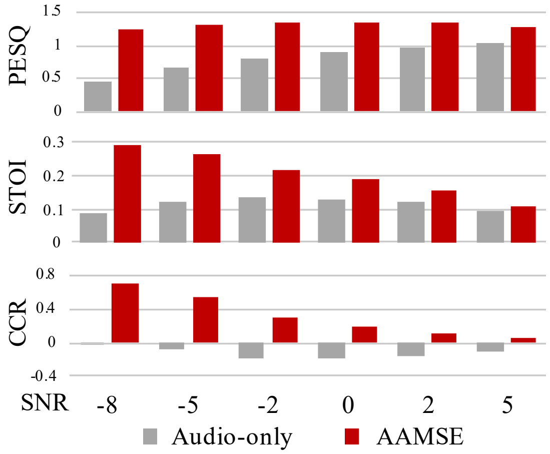

Fig. 5. The SE improvement of the best audio-only SE model (BLSTM) and the best AAMSE model (BLSTM with unilateral encoding) in different SNRs.

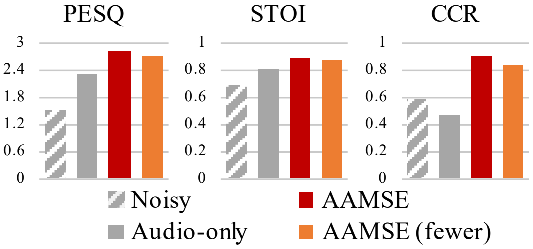

Fig. 6. The average scores of different SE systems.

6. ACKNOWLEDGEMENT

The authors would like to thank NTT Communication Science Laboratories for permitting them to use the articulatory data.

	Audio only	Direct concatenating		Unilateral encoding		Bilateral encoding	
FCN	2.311	2.653	+0.342	2.251	-0.060	2.575	+0.264
TDNN	2.064	2.402	+0.338	2.434	+0.370	2.390	+0.326
BLSTM	2.329	2.793	+0.464	**2.839**	**+0.510**	2.470	+0.141

Table 3. PESQ of different SE models (noisy=1.530).

	Audio only	Direct concatenating		Unilateral encoding		Bilateral encoding	
FCN	0.814	0.881	+0.067	0.796	-0.018	0.862	+0.048
TDNN	0.738	0.816	+0.078	0.827	+0.089	0.820	+0.082
BLSTM	0.801	0.885	+0.084	**0.891**	**+0.090**	0.825	+0.024

Table 4. STOI of different SE models (noisy=0.686).

7. REFERENCES

[1] A. El-Solh, A. Cuhadar, and R. A. Goubran, “Evaluation of speech enhancement techniques for speaker identification in noisy environments,” in *Proc. ISM 2007*.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[3] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[4] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” in *Proc. Interspeech 2017*.

[5] H. Levit, “Noise reduction in hearing aids: An overview,” *J. Rehabil. Res. Develop.*, vol. 38, no. 1, pp. 111–121, 2001.

[6] E. W. Healy, M. Delfarah, E. M. Johnson, and D. Wang, “A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation,” *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1378–1388, 2019.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder.” in *Proc. Interspeech 2013*.

[8] B. Xia and C. Bao, “Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification,” *Speech Communication*, vol. 60, pp. 13–29, 2014.

[9] P. G. Shivakumar and P. G. Georgiou, “Perception optimized deep denoising autoencoders for speech enhancement.” in *Proc. Interspeech 2016*.

[10] D. Liu, P. Smaragdis, and M. Kim, “Experiments on deep learning for speech denoising,” in *Proc. Interspeech 2014*.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[12] M. Kolbæk, Z.-H. Tan, and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[13] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, “Convolutional maxout neural networks for speech separation,” in *Proc. ISSPIT 2015*.

[14] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” in *Proc. APSIPA 2017*.

[15] A. Pandey and D. Wang, “A new framework for cnn-based speech enhancement in the time domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[16] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr,” in *Proc. LVA/ICA 2015*.

[17] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, “Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks,” in *Proc. Interspeech 2015*.

[18] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, “Multiple-target deep learning for lstm-rnn based speech enhancement,” in *Proc. HSCMA 2017*.

[19] J.-C. Hou, S.-S. Wang, Y.-H. Lai, J.-C. Lin, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-visual speech enhancement using deep neural networks,” in *Proc. APSIPA 2016*.

[20] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-visual speech enhancement using multimodal deep convolutional neural networks,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[21] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, “Robust articulatory speech synthesis using deep neural networks for bci applications,” in *Proc. Interspeech 2014*.

[22] F. Taguchi and T. Kaburagi, “Articulatory-to-speech conversion using bi-directional long short-term memory.” in *Proc. Interspeech 2018*.

[23] J. Wei, Y. Ji, J. Zhang, Q. Fang, W. Lu, K. Honda, and X. Lu, “Study of articulators’ contribution and compensation during speech by articulatory speech recognition,” *Multimedia Tools and Applications*, vol. 77, no. 14, pp. 18 849–18 864, 2018.

[24] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an hmm-based speech production model,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[25] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. Interspeech 2015*.

[26] G. Hu, “100 nonspeech environmental sounds,” 2004 (accessed October 18, 2020), http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP 2001*.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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.

[29] A. Zhang, “Speech recognition (version 3.8),” 2017 (accessed October 18, 2020), https://github.com/Uberi/speech_recognition#readme.

[30] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR 2015*.

[32] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, “Lite audio-visual speech enhancement,” in *Proc. Interspeech 2020*.

[33] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring speech enhancement with generative adversarial networks for robust speech recognition,” in *Proc. ICASSP 2018*.