EMA2S: An End-to-End Multimodal Articulatory-to-Speech System

Yu-Wen Chen[1], Kuo-Hsuan Hung[1], Shang-Yi Chuang[1],
Jonathan Sherman[1], Wen-Chin Huang[23], Xugang Lu[4], Yu Tsao[1]
[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Graduate School of Informatics, Nagoya University, Japan
[4]National Institute of Information and Communications Technology, Japan

***Abstract*—Synthesized speech from articulatory movements can have real-world use for patients with vocal cord disorders, situations requiring silent speech, or in high-noise environments. In this work, we present EMA2S, an end-to-end multimodal articulatory-to-speech system that directly converts articulatory movements to speech signals. We use a neural-network-based vocoder combined with multimodal joint-training, incorporating spectrogram, mel-spectrogram, and deep features. The experimental results confirm that the multimodal approach of EMA2S outperforms the baseline system in terms of both objective evaluation and subjective evaluation metrics. Moreover, results demonstrate that joint mel-spectrogram and deep feature loss training can effectively improve system performance.**

***Index Terms*—articulatory movement, end-to-end, multimodal learning, neural network, speech synthesis**

I. Introduction

Silent speech interfaces enable people to communicate without the presence of an acoustic signal. Such techniques can provide patients who suffer from vocal cord disorders a more natural alternative way to communicate [1]. Also, these techniques can be helpful in situations requiring acoustic silence, or in high-noise environments, since acoustic signals are not required as input and thus background noises have a greatly reduced effect. Various silent speech technologies have been investigated, including magnetic resonance imaging [2], [3], electromyograms [4], permanent magnetic articulograph [5], and electromagnetic midsagittal articulography (EMA) [6]–[9].

In this study, we use EMA to collect the articulatory movements data. EMA records the articulatory movements by using an electromagnetic field to induce currents in sensors, which are attached to articulators such as lips and tongue. Previous studies have proposed several methods to convert EMA signals towards acoustic features. [10] uses a codebook to store articulatory and acoustic parameters pairs, and then estimates the spectrum of the articulatory data by selecting neighbor samples in the codebook. Also, statistical models such as Gaussian mixture models (GMM) [11], hidden Markov models (HMM) [12], fully connected neural network [13], and bidirectional long short-term memory (BLSTM) [14], [15] have been used to map the articulatory movements to acoustic signals. These studies have indicated that neural-network-based methods achieve better performance than GMM and HMM methods. However, they only use neural networks to map the articulatory movements to spectral features, and reconstruct the waveform with traditional parametric vocoders such as STRAIGHT [16] and WORLD [17]. Since the neural-network-based vocoders [18], [19] have shown much superior performance over traditional parametric vocoders, it is logical to investigate the performance of using neural networks for both articulatory-to-spectrum mapping and waveform reconstruction. We propose an end-to-end multimodal articulatory-to-speech system, EMA2S, that improves the existing speech synthesis systems by applying two techniques: (1) a neural-network-based vocoder and (2) a multimodal jointly training method with a combined loss. Concerning the first, in addition to demonstrating much superior performance over the traditional parametric vocoders, the neural-network-based vocoder allows further development in an end-to-end trainable system that directly converts articulatory movements into waveforms. It is not bounded or required to fit the constraints of parametric or independently trained vocoders. For the second, we jointly train with a combined loss of different acoustic features: (a) the spectrogram loss, (b) the mel-spectrogram loss, and (c) the deep feature loss of spectral embeddings and articulatory movement embeddings. The deep feature loss [20] measures the dissimilarity between articulatory movement embeddings and spectral embeddings. To calculate the deep feature loss, both articulatory movements and spectrograms are used as input data during training, but only articulatory movements are necessary during inference. The introduction of the deep feature loss allows synthesis models to learn a better representation of one modality (articulatory movements) from multiple modalities (spectrograms and articulatory movements).

Experimental results show that our proposed system outperforms a previous system in terms of mel-cepstral distortion (MCD) [21], perceptual evaluation of speech quality (PESQ) [22], short-time objective intelligibility (STOI) [23], character correct rate (CCR) of a pre-trained automatic speech recognition (ASR) system [24], and a listening test. For the reason that users will be more willing to use the device without using invasive sensors, we investigate the system performance with only four less invasive EMA sensors. The results reveal that our proposed system still performs better than the previous system.

The rest of the paper is organized as follows. Section II introduces the related works. The proposed EMA2S system is presented in Section III. Experimental details and results are given in Section IV to demonstrate the performance of the proposed approach. Section V concludes our work.

II. Related Work

In this section, we review Parallel WaveGAN (PWG) [19], multimodal learning [25], and deep feature loss [20] used in our proposed model.

A. Parallel WaveGAN

PWG [19] is a non-autoregressive, fast, and effective parallel waveform generation method based on a generative adversarial network [26]. PWG has shown superior performance to parametric vocoders, and can train and inference faster than autoregressive generative models such as WaveNet [18].

PWG uses a joint training method of the multi-resolution short-time Fourier transform (STFT) loss and the waveform-domain adversarial loss. To calculate the adversarial loss, PWG is composed of two separate neural networks: a generator and a discriminator. The input of the generator is auxiliary acoustic features, which are mel-spectrograms and random noises drawn from a Gaussian distribution, and the output of the generator is the raw waveform in parallel. The generator learns a distribution of realistic waveforms by trying to deceive the discriminator to classify the generated samples as real. On the contrary, the discriminator learns by correctly recognizing the generated sample as fake and the ground truth sample as real.

B. Multimodal Learning

Multimodal learning [25] aims to learn relating information from multiple modalities and fill the missing modality given the observed ones. Numerous research has investigated the effectiveness of incorporating different features into speech-related systems, including text [27]–[30], videos [31]–[33], bone-conducted microphone signals [34], electropalatography [35], and articulatory movements [36]–[38].

C. Deep Feature Loss

Deep feature loss [20] is defined as the dissimilarity of the embeddings in neural networks. Previous research has shown that deep features can capture the perceptual features of the input, and deep feature loss can effectively improve the model performance without adding the complexity of the processing network itself [20], [39], [40].

III. Proposed Method

In this work, we propose an end-to-end multimodal articulatory-to-speech system (EMA2S) that uses PWG [19] as a vocoder and incorporates multimodal learning [25] and deep feature loss [20]. Unlike previous studies, the deep feature loss in this work exploits the idea of multimodal learning, and calculates the dissimilarity of two modalities' embeddings (EMA embeddings and spectral embeddings) instead of one (EMA embeddings). The combination of multimodal learning and deep feature loss is designed for low resource data such as EMA signals since a network that extracts deep features of low resource data (e.g., EMA signals) is more difficult to obtain or train than a network that extracts deep features of high resource data (e.g., audio signals). Furthermore, given the objective to transform EMA signals to speech, the deep feature loss calculated by EMA embeddings and spectral embeddings aligns the training of the system.

A. Architecture

Fig. 1 depicts the proposed system, which includes an EMA encoder E_{ema}, a spectral encoder E_{spec}, and a shared decoder D. The spectral encoder is for guiding the training process of the EMA encoder and the shared decoder, and the spectral encoder will not be used in inference - only EMA features are required in testing. The output of the decoder is a spectrogram. This spectrogram will be transformed into a mel-spectrogram and then reconstructed back to waveform by the PWG model.

The spectral encoder contains two BLSTM layers, and the hidden units of the first and second layers are 196 and 256, respectively. The BLSTM layers are followed by a linear layer with 256 units and the rectified linear unit (ReLU). The EMA encoder consists of two BLSTM layers with 128 and 256 hidden units, followed by a linear layer with 256 units and the ReLU. The shared decoder consists of three BLSTM layers with 256 hidden units, a linear layer with 513 units, and the ReLU.

B. Training Stages and Loss Function

The training process contains two stages. The first stage is to train the spectral encoder E_{spec} and the shared decoder D. The second stage is to train the EMA encoder E_{ema} and shared decoder D.

In the first stage, E_{spec} and D are optimized by minimizing $L^{(1)}$ which is the reconstructed loss of spectral features, including spectrogram loss $L^{(1)}_{spec}$ and mel-spectrogram loss $L^{(1)}_{mel}$. $L^{(1)}$ is defined as follows:

$$\begin{aligned} L^{(1)}_{spec} &= |D(E_{spec}(s)) - s| \\ L^{(1)}_{mel} &= |M(D(E_{spec}(s))) - M(s)| \\ L^{(1)} &= L^{(1)}_{spec} + L^{(1)}_{mel} \end{aligned} \tag{1}$$

where s is the input spectrogram and M is the mapping from spectrogram to mel-spectrogram.

In the second stage, E_{ema} and D are optimized by minimizing $L^{(2)}$ which combines the reconstructed spectrogram loss $L^{(2)}_{spec}$, the reconstructed mel-spectrogram loss $L^{(2)}_{mel}$, and the deep feature loss $L^{(2)}_{df}$. In this stage, the spectrogram and mel-spectrogram are reconstructed from EMA embeddings rather than spectral embeddings. The deep feature loss $L^{(2)}_{df}$ measures the dissimilarity between EMA embeddings and spectral embeddings. We want the EMA embeddings close to spectral embeddings because we assume that we can more easily reconstruct the spectrograms and mel-spectrograms by the spectral embeddings. $L^{(2)}$ is defined as follows:

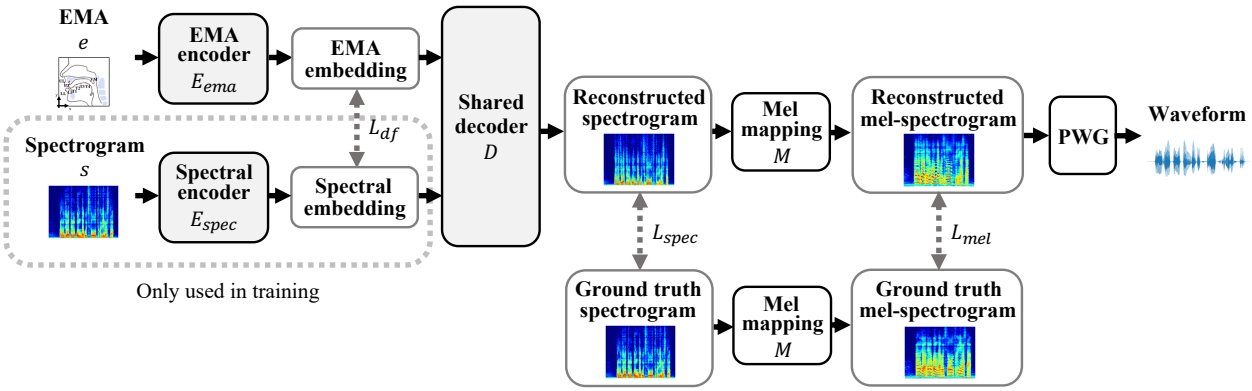

Fig. 1. Diagram of EMA2S system.

$$
\begin{aligned}
L_{spec}^{(2)} &= |D(E_{ema}(e)) - s| \\
L_{mel}^{(2)} &= |M(D(E_{ema}(e))) - M(s)| \\
L_{df}^{(2)} &= |E_{ema}(e) - E_{spec}(s)| \\
L^{(2)} &= L_{spec}^{(2)} + L_{mel}^{(2)} + L_{df}^{(2)}
\end{aligned} \tag{2}
$$

where e is the input EMA signal.

IV. Experiments

A. Experimental setup

In this study, we use the EMA data collected by NTT, Tokyo, Japan [41]. The sensor coils of EMA sensors 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. 2. EMA records the Cartesian coordinates of each sensor point at a sampling rate of 250 Hz, and the audio signals are recorded at the same time with a sampling rate of 16 kHz. The dataset contains articulatory movements and speech signals from three speakers, each providing 354 utterances. The training set includes 304 utterances from each speaker, and the testing set includes the remaining 50 utterances.

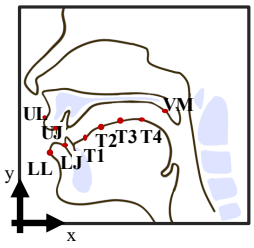

Fig. 2. The placement of EMA sensors.

We divide the EMA signals by the maximum value for normalization, and we concatenate ± frames for a five-frame context window. For speech signals, we convert waveforms to spectrograms by STFT and only use the magnitude components to train the proposed model. The STFT settings are the same as that of the pre-trained PWG model, which use a window size of 1024 and a hop length of 256. The PWG used in this work was pre-trained under the Japanese corpus JNAS [42], which has the same sampling rate as that of our dataset.

We evaluate our results with objective evaluation metrics including MCD [21], PESQ [22], STOI [23], and CCR of a pre-trained ASR [24] system. We measure speech quality by using MCD and PESQ, and we evaluate speech intelligibility with STOI and CCR of a pre-trained ASR system. The CCR is calculated using Levenshtein distance [43].

For subjective evaluation, we conduct an A/B test for subjective listening tests. The A/B test compares the baseline system and the proposed system to determine which one brings better signal quality. The testing data contain five questions for each of the three speakers, resulting in a total of 15 questions.

B. Baseline System

The baseline system is based on the work in [15], which is composed of three fully-connected layers, a layer normalization [44] layer, a sigmoid layer with 128 units, two layers of BLSTM with 256 units, and a fully-connected output layer. The input of the model is the EMA signal, while the target is the concatenated feature of mel-cepstrum, aperiodic parameters, F0 and voice activity detection (VAD). WORLD [17] is used to extract the feature parameters of spectral envelope, aperiodic parameters and F0. The spectral envelope is further processed into mel-cepstrum, and F0 is further processed into VAD. Each feature parameter was normalized to zero mean and unit standard deviation. During inference, WORLD generates a speech signal using the synthesized speech parameters. Note that we skip the dynamic features (delta features) and maximum likelihood parameter generation algorithm used in [15] because we have already considered the forward and backward time series of the input EMA data.

C. Experimental results

1) Perceptual Analysis: Fig. 3 visualizes the ground truth spectrograms as well as the reconstructed spectrograms. The results show that reconstructed spectrograms are visually close to the ground truths, which reveal that EMA signals can be successfully transformed into speech signals. Also, as

	Loss	E_{spec}	MCD	PESQ	STOI	CCR
Baseline	-	-	7.815	1.279	0.696	0.818
SI	$L^{(2)}_{spec}$	✗	8.264	1.259	0.679	0.796
SII	$L^{(2)}_{spec}, L^{(2)}_{mel}$	✗	7.334	1.320	0.702	0.841
SIII	$L^{(1)}_{spec}, L^{(2)}_{spec}, L^{(2)}_{df}$	✓	8.445	1.303	0.697	0.831
EMA2S	$L^{(1)}, L^{(2)}$	✓	**7.176**	**1.350**	**0.716**	**0.868**

TABLE I
TRAINING LOSS AND THE NUMERICAL ANALYSIS OF THE ARTICULATORY-TO-SPEECH SYSTEMS.

indicated in the red boxes, the proposed EMA2S generated speech with more details in the high-frequency bands and less unnatural formant structures than the baseline system.

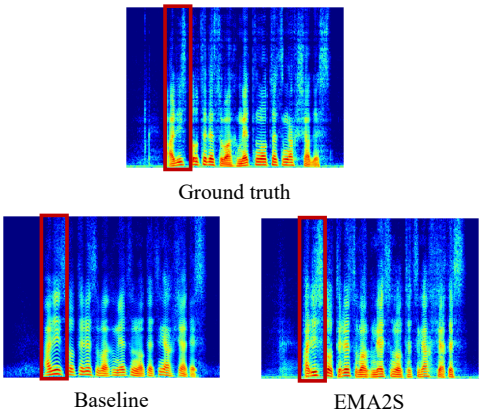

Fig. 3. Visualization of spectrograms.

2) Numerical Analysis: To evaluate the combined loss, we compared EMA2S performance with three configurations of the articulatory-to-speech system (SI, SII, and SIII). SI has no spectral encoder E_{spec}, and is trained with only spectrogram loss $L^{(2)}_{spec}$. SII is SI trained with both spectrogram loss $L^{(2)}_{spec}$ and mel-spectrogram loss $L^{(2)}_{mel}$. SIII is SI with multimodal jointly training. It has a spectral encoder, and is trained with spectrogram loss $L^{(1)}_{spec}$, $L^{(2)}_{spec}$, and deep feature loss $L^{(2)}_{df}$.

Table I organizes the corresponding training losses and shows the performance of the different articulatory-to-speech systems, including the baseline, SI, SII, SIII, and the proposed EMA2S. The check mark in the column E_{spec} indicates whether the system contains a spectral encoder that used for multimodal learning. The results reveal that EMA2S outperforms the baseline system in terms of MCD, PESQ, STOI, and CCR. Moreover, the performance of the articulatory-to-speech system can be improved by training the system with a combined loss of spectrogram and mel-spectrogram and using the multimodal jointly training method.

3) Listening Test: We recruited 10 participants for an A/B test. Each participant must answer 15 questions, and each question contains two speech waveforms of the same utterance. One of the waveforms is generated by the baseline system, and the other is generated by our EMA2S system. The participants are asked to choose which waveform they prefer. Experimental results in Fig. 4 reveals that an average of 83% participants voted for the proposed EMA2S system, and the remaining 17% of participants voted for the baseline system.

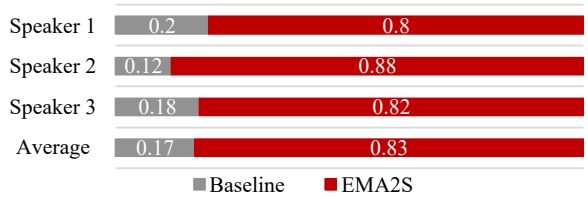

Fig. 4. The results (in percentage) of the A/B listening test.

4) Further Analysis: Because EMA requires a laboratory environment for recording, we test EMA2S with only four less invasive sensors (UL, LL, LJ, and T1) to improve the applicability of the system, reasoning that without using invasive sensors, users will be more willing to use the devices. Fig. 5 shows that EMA2S with only four sensors (denoted as fewer) can still achieve better performance than the baseline system.

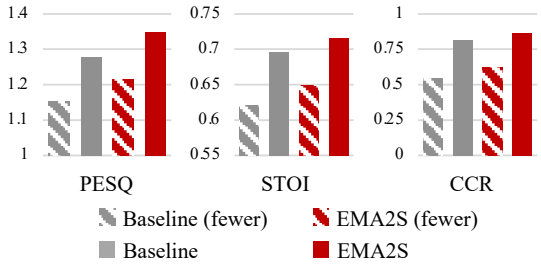

Fig. 5. The average scores of different articulatory-to-speech systems.

V. CONCLUSION

We propose EMA2S, an end-to-end multimodal articulatory-to-speech system that uses (1) a neural-network-based vocoder and (2) a multimodal jointly training method with a combined loss of spectrogram, mel-spectrogram, and the deep feature. Experimental results reveal that our proposed EMA2S system outperforms the baseline system in terms of objective evaluation metrics and a subjective listening test. In the future, we plan to increase the naturalness of the synthesized speech by incorporating a natural language model in the articulatory-to-speech system, and improve the performance of the system with limited sensors as input.

ACKNOWLEDGMENT

We thank NTT Communication Science Laboratories for permitting us to use the articulatory data.

REFERENCES

[1] Y. Lin, L. Wang, J. Dang, S. Li, and C. Ding, "End-to-end articulatory modeling for dysarthric articulatory attribute detection," in *Proc. ICASSP 2020*.

[2] P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[3] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," in *Proc. UAHCI 2007*.

[4] M. Janke and L. 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.

[5] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, "Permanent magnetic articulograph (PMA) vs electromagnetic articulograph (EMA) in articulation-to-speech synthesis for silent speech interface," in *Proc. SLPAT 2019*.

[6] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. NIPS 2010*.

[7] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845–856, 2012.

[8] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[9] S. Li and L. Wang, "Cross linguistic comparison of Mandarin and English EMA articulatory data," in *Proc. INTERSPEECH 2012*.

[10] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. ICSLP 1998*.

[11] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *Proc. SSW5 2004*.

[12] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance hmm," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[13] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[14] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks." in *Proc. INTERSPEECH 2016*.

[15] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory." in *Proc. INTERSPEECH 2018*.

[16] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[18] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: a generative model for raw audio," in *Proc. SSW9*, 2016.

[19] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP 2020*.

[20] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. INTERSPEECH 2019*.

[21] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. PACRIM 1993*.

[22] 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*.

[23] 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.

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

[25] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML 2011*.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS 2014*.

[27] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. INTERSPEECH 2015*.

[28] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *arXiv preprint arXiv:1912.06813*, 2019.

[29] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *Proc. ICASSP 2020*.

[30] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer learning from speech synthesis to voice conversion with non-parallel training data," *arXiv preprint arXiv:2009.14399*, 2020.

[31] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of lombard effect," *Speech Communication*, vol. 115, pp. 38–50, 2019.

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

[33] 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.

[34] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-w. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[35] P.-H. Chen, R. T.-H. Tsai, and Y. Tsao, "Multimodal electropalatography-audio speech enhancement," *submitted to ICASSP 2021*.

[36] I. Steiner, S. Le Maguer, and A. Hewer, "Synthesis of tongue motion and acoustics from text using a multimodal articulatory database," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2351–2361, 2017.

[37] R. Li and J. Yu, "Multimodal 3D visible articulation system for syllable based mandarin chinese training," in *Proc. VCIP 2017*.

[38] Y.-W. Chen, K.-H. Hung, S.-Y. Chuang, J. Sherman, X. Lu, and Y. Tsao, "A study of incorporating articulatory movement information in speech enhancement," *arXiv preprint arXiv:2011.01691*, 2020.

[39] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. NIPS 2015*.

[40] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Proc. WACV 2017*.

[41] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 453–463, 2001.

[42] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

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

[44] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *stat*, vol. 1050, p. 21, 2016.