Speech Synthesis from IPA Sequences through EMA Data

Koki Maruyama, Shun Sawada, Hidefumi Ohmura, Kouichi Katsurada
Tokyo University of Science, Noda, Chiba 278-8510, Japan
E-mail: 6323542@ed.tus.ac.jp, {sawada, katsurada}@rs.tus.ac.jp, ohmura@is.noda.tus.ac.jp

***Abstract*—Previous research on articulatory synthesis has partially implemented and verified the speech synthesis systems that mimic the human vocalization process. This study proposes a speech synthesis system that more faithfully replicates the human vocalization process by constructing a model that generates speeches from the international phonetic alphabet (IPA) sequence through articulatory movement data. For training and evaluating this model, we used the ATR phoneme-balanced 503 sentence electromagnetic articulography (EMA) database that contains pair data of speech and EMA data of a male Japanese speaker. The experimental results showed that using articulatory movement data as intermediate information improved the quality of synthesized speeches regarding objective scores such as mel-cepstral distortion (MCD), phoneme error rate (PER), and perceptual evaluation of speech quality (PESQ). In particular, the effectiveness of using articulatory movement data was confirmed in the case of speech where articulation points were included in the EMA database.**

1. Introduction

Human beings generate speech by converting linguistic information into movements of the articulatory organs such as the lip and tongue, thereby controlling the airflow. This fact shows that considering articulatory movements on speech synthesis has the potential to improve the quality of synthesized speeches. Although many text-to-speech (TTS) systems express the movements of articulatory organs with Mel-spectrogram, which illustrates the resonant frequency generated from the organs, Mel-spectrogram is the final result of articulation that can be directly converted to waveform with the vocoder. Modeling the movements of articulatory organs between text and Mel-spectrogram can more accurately simulate the human speech production process; thus, potentially improving the quality of the generated speeches.

Based on this background, there have been some studies regarding the relationships between articulatory movements, linguistic information, and speech. There are several types of articulatory movement data, with three notable examples: first, real-time magnetic resonance imaging (rtMRI), which measures the movements of the articulatory organs using an MRI device [1-3]. Although recording rtMRI data is costly, it is characterized by its ability to capture the movements of the entire articulatory organs. Second, electromagnetic articulography (EMA), which measures the movements of the articulatory organs using coils attached to them [4-7]. Although EMA data do not include data on vocal fold movements, it is characterized by its high sampling rate and spatial resolution. Third, ultrasound tongue imaging (UTI), which uses ultrasound to measure the movements of the tongue [8,9]. UTI captures only the movements of the tongue but it is characterized by its ease of measurement. These studies generated articulatory movement data from linguistic features or speech from articulatory movement data. Thus, these studies have only partially implemented and evaluated speech synthesis systems that mimic the human vocalization process.

In this study, we propose a model that synthesizes speech from linguistic features through articulatory movement data, using EMA data as the articulatory movement. We compared three models; (1) generating Mel-spectrogram from the international phonetic alphabet (IPA) sequence through estimated EMA, (2) directly generating Mel-spectrogram from the IPA sequence, and (3) generating Mel-spectrogram from both the IPA sequence and estimated EMA. We confirm the effectiveness of using articulatory movement data with some objective scores such as Mel-cepstral distortion (MCD) [10], phoneme error rate (PER), and perceptual evaluation of speech quality (PESQ) [11] to evaluate the quality of generated speeches. We also analyzed which position of articulation data was effective for speech generation and which phones benefited from the information on articulatory movement data.

2. Related Work

2.1 EMA data generation from the IPA

We have proposed a method for generating EMA data from the IPA [12]. IPA is a symbol that contains a wealth of articulatory information which is defined in terms of articulation method and position of articulation. Each phone in any language is represented as a symbol of the IPA. EMA [13,14] is an instrument that measures the articulatory movement using a coil attached to the articulatory organs, and is characterized by its ability to capture the movements of

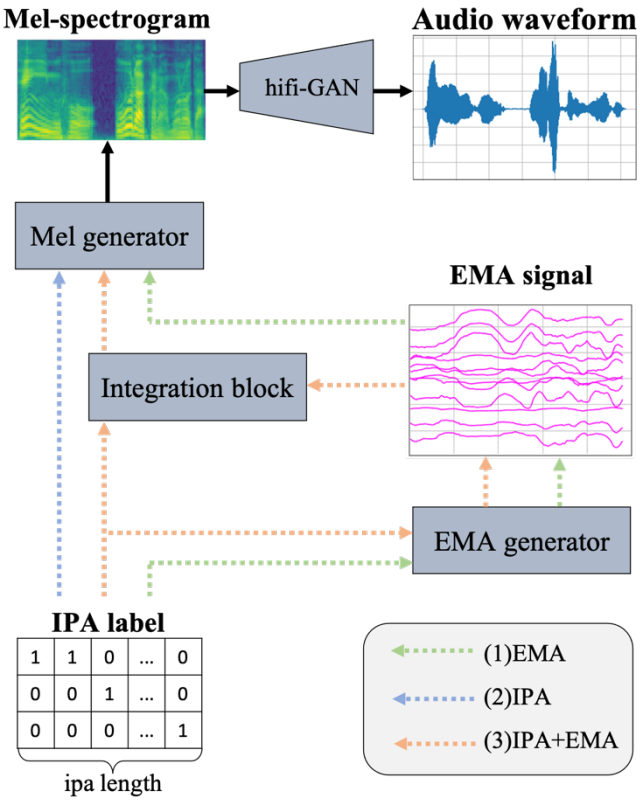

Fig. 1 Structure of speech synthesis model from the IPA sequence through EMA data.

articulatory organs at a high sampling rate and spatial resolution.

In [12], EMA data is generated from IPA using the 1DCNN-based SENet [15]. In this model, the receptive field, which is the input window width that affects the output, was optimized for generating EMA data from the IPA sequence. Additionally, dedicated models were constructed for each articulatory organ, and the final outputs were generated by combining the outputs of these multiple models as an ensemble model. This model enabled the generation of EMA data with high accuracy and speed.

2.2 Speech synthesis from EMA data

Kim et al. proposed a multi-speaker speech synthesis model from EMA data and speaker embedding through the Mel-spectrogram [16]. The proposed model consists of four modules; the first is an encoder consisting of the residual connections and CNN layers. The second is a generation model of F_0 and energy consisting of linear and CNN layers. The errors of F_0 and energy were used as the loss during training. The third is a model for generating the Mel-spectrogram, which consists of the Conformer [17]. The fourth is a vocoder that accepts Mel-spectrogram as input. This model uses the hifi-GAN [18] as a vocoder. The model showed high scores in both subjective and objective evaluations.

3. Proposed Models

The structure of the proposed model is shown in Fig. 1. To verify the effectiveness of using EMA data and the IPA information separately, we created three types of models: (1) EMA, which uses EMA data as intermediate information, (2) IPA, which generates speech directly from the IPA sequence, and (3) IPA+EMA, which uses both EMA and IPA sequence as the intermediate information and the input, respectively. For all models, hifi-GAN was used as the vocoder for generating speech from the Mel-spectrogram.

3.1 (1) EMA Model

The EMA Model first outputs EMA data from the IPA sequence and then generates a Mel-spectrogram from the output EMA data. IPA is considered as a suitable input for the EMA data generation model because it contains a wealth of articulatory information. In this model, the *EMA generator*, which generates EMA data from the IPA, consists of a 1DCNN-based SENet, and the *Mel generator*, which generates Mel-spectrogram from EMA data, is composed of the model proposed by Kim et al. [16], which learns losses including F_0.

3.2 (2) IPA Model

The IPA Model directly generates a Mel-spectrogram from the IPA sequence only. This model is the same as the *Mel generator* in the (1) EMA Model.

3.3 (3) IPA+EMA Model

The IPA+EMA Model first outputs EMA data from the IPA sequence and then generates a Mel-spectrogram from the output EMA data along with the IPA. The models that output EMA data from the IPA sequence and that generate Mel-spectrogram from the EMA data are the same as the *EMA generator* and *Mel generator* in the (1) EMA Model, respectively. The EMA data and IPA are combined in the *Integration block*, and this combined data is used as input to the *Mel generator*.

4. Experiment

In the experiment, we compared three models: (1) EMA, (2) IPA, and (3) IPA+EMA by using four evaluation metrics regarding the quality of the synthesized speeches.

4.1 Datasets

4.1.1 EMA database [19]

The EMA database contains IPA labeling, EMA, and speech data of the ATR phoneme-balanced 503 sentences [20] spoken by one male Japanese speaker. In the experiment, 450 sentences were used as training data, 27 sentences as validation data, and

Table 1 F_0RMSE, MCD, PER, and PESQ of synthesized speeches.

Model	F_0RMSE (Hz)	MCD (dB)	PER	PESQ
(1) EMA	39.34	13.36	0.26	1.09
(2) IPA	35.73	10.58	0.10	1.11
(3) IPA+EMA	**33.47**	**10.27**	**0.09**	**1.12**

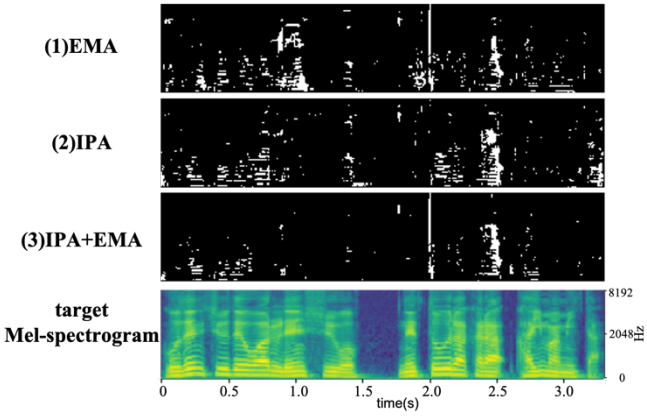

Fig. 2 Difference images of Mel-spectrograms between proposed models and ground truth.

26 sentences as test data, and were used for learning the *EMA generator* and *Mel generator* and fine-tuning the hifi-GAN.

Ninety-six IPA phones were included in the database and were preprocessed by converting them into one-hot vectors. EMA data consists of 12 dimensions of the back-forth and up-down movement data of six points on the mid-sagittal plane: the upper lip, lower lip, lower jaw, tongue tip, tongue body, and tongue dorsum. We preprocessed these data by normalizing them to range from -1 to 1. The speech data were transformed such that the Mel-spectrogram had 60 dimensions.

4.1.2 JVS corpus [21]

The JVS corpus contains speech data of ATR phoneme-balanced 503 sentences spoken by multiple Japanese speakers. In the experiment, only the data from male speakers were used for pre-training the hifi-GAN, with 5732 sentences as training data and 636 sentences as verification data. This is because the EMA database contains only the data from one male speaker. The speech data were transformed such that the Mel-spectrogram had 60 dimensions.

4.2 Evaluation metrics

As evaluation metrics, we used F_0 Root Mean Squared (RMSE) to measure the accuracy of F_0 estimation. Haverst [22] was used to extract F_0. Additionally, to evaluate the quality of the synthesized speech, we used MCD [10], PER and PESQ [11]. PER was calculated with the speech recognition

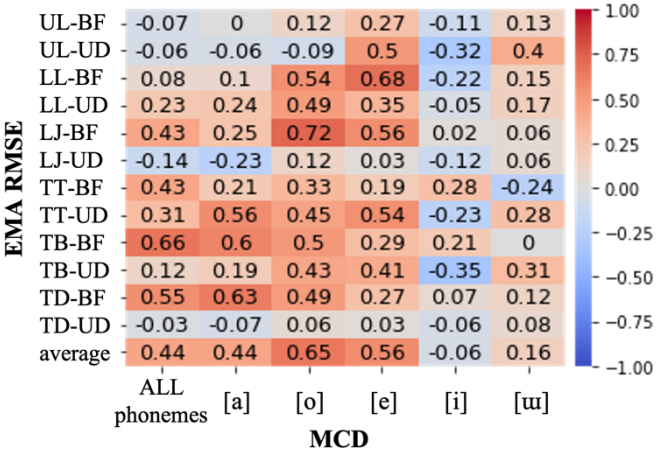

Fig. 3 Heatmap of Pearson correlation coefficients between the MCD and the RMSE of EMA data in the (3) IPA+EMA Model. "UL" means upper lip, "LL" means lower lip, "LJ" means lower jaw, "TT" means tongue tip, "TB" means tongue body, "TD" means tongue dorsum, "BF" means back-forth, and "UD" means up-down.

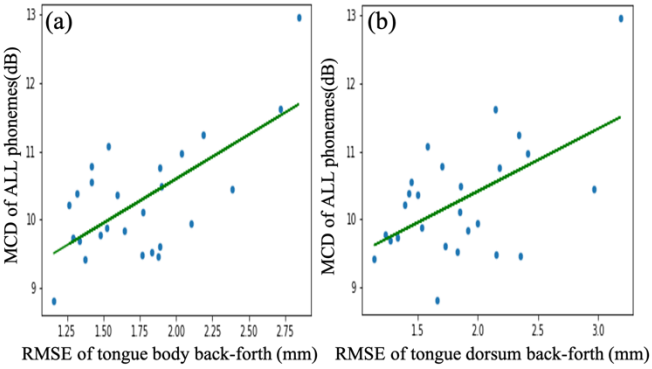

Fig. 4 Scatter plot (blue) and regression line (green).
(a) : "MCD of all phonemes" and "RMSE of tongue body back-forth position" in the EMA.
(b) : "MCD of all phonemes" and "RMSE of tongue dorsum back-forth position" in the EMA.

model Julius [23]. Lower values of F_0RMSE, MCD, and PER indicate better evaluation, whereas higher values of PESQ indicate better speech quality.

5. Results and Discussions

5.1 Overall results

Table 1 summarizes the overall experimental results. For all evaluation metrics, the (3) IPA+EMA Model showed the best results. The addition of EMA data as intermediate information enabled more accurate modeling of vocal tract characteristics, resulting in improved speech quality. However, F_0RMSE was 33.47 (Hz), indicating that synthesized speech cannot be considered natural. This is because EMA data do not contain information about the vocal folds, which are closely related to F_0.

Fig. 2 shows the difference images of Mel-spectrograms between the target speech and the synthesized speeches. White

Table 2 MCD for each phoneme category.

Model	MCD(dB)			
	vowel	consonant	semivowel	silence
(1) EMA	12.03	13.86	12.30	18.05
(2) IPA	9.85	10.68	**9.99**	11.63
(3) IPA+EMA	**9.69**	**10.37**	11.06	**11.07**

Table 3 MCD for each vowel.

Model	MCD(dB)				
	[a]	[o]	[e]	[i]	[ɯ]
(1) EMA	10.92	11.46	10.18	14.04	12.66
(2) IPA	9.97	9.79	8.81	**10.34**	**10.04**
(3) IPA+EMA	**9.55**	**9.50**	**8.52**	10.41	10.49

areas indicate where the Mel-spectrogram of the target speech was not accurately reproduced. Comparing the three models using the difference images, we can observe that the Mel-spectrogram of the speech synthesized from the (3) IPA+EMA Model exhibited the best reproducibility. However, none of the models accurately reproduced the target Mel-spectrogram around the 2.5 seconds. The sound at this point is [a], and the issue may be attributed to amount of articulatory movement; further analysis is required.

Fig. 3 shows the Pearson correlation coefficient between the RMSE of EMA data and the MCD in the (3) IPA+EMA Model. Focusing on the MCD of ALL phonemes, the correlation coefficient with the RMSE of the tongue body back-forth position (TB-BF) was the highest at 0.66, and the correlation coefficient with the RMSE of the tongue dorsum back-forth position (TD-BF) was the second highest at 0.55. Fig. 4 shows the scatter plot and regression line for the two pairs that recorded high correlation coefficients in Fig. 3. Fig. 4 (a) and (b) show the graphs which were depicted from the data of TB-BF and TD-BF, which showed the two best performances. The scatter plot and correlation coefficient indicate that the higher the accuracy of the estimation tongue body or dorsum back-forth position in the EMA, the better the MCD values for all phonemes. This is because MCD is an index of vocal tract characteristics, and the back-forward movement of the tongue has a significant effect on the shape of the vocal tract.

5.2 MCD for each phoneme category

Table 2 compares three models by the MCD of vowels, consonants, semivowels, and silence. The MCD of vowels, consonants, and silence was the best with the (3) IPA+EMA Model. However, that of semivowels was the best with the (2) IPA Model. The test data included only the semivowels [j] (voiced palatal approximant) and [w] (voiced labial–velar approximant), both of which are articulated at the palate as the point of articulation. However, because the EMA data do not contain information about the palate, the EMA data used as intermediate information was considered to be noise, leading to this result.

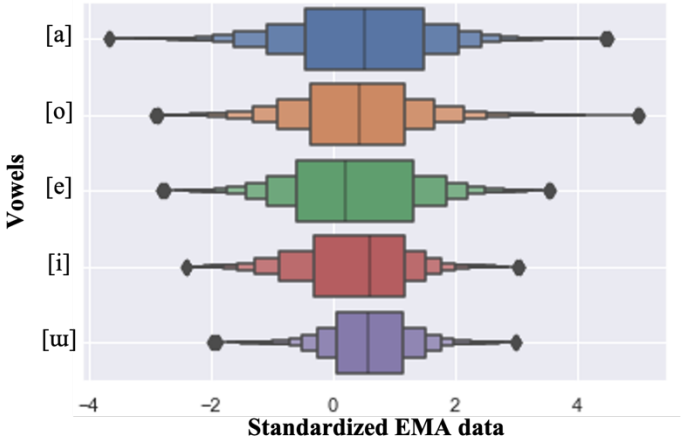

Fig. 5 Boxenplot of standardized EMA data for each vowel in test data.

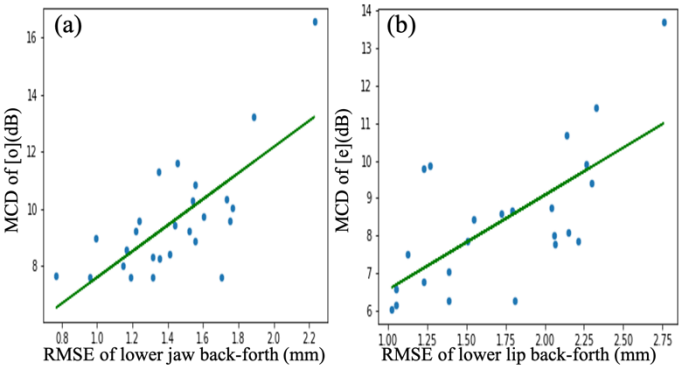

Fig. 6 Scatter plot (blue) and regression line (green).
(a): "MCD of [o]" and "RMSE of the lower jaw back-forth position" in the EMA.
(b) : "MCD of [e]" and "RMSE of the lower lip back-forth position" in the EMA

5.3 MCD for each vowel

Table 3 compares the three models using the MCD of single vowels. The MCD of [a], [o], and [e] was the best with the (3) IPA+EMA Model. These three vowel sounds were either open or close-mid vowels. Fig. 5 shows boxenplots that display the standardized EMA data categorized by vowels in the test data. The EMA data was standardized for each measurement point and converted from 12 dimensions to one dimension. From these plots, it can be observed that the data distribution is scattered in the order [a], [o], [e], [i], [ɯ]. That is, the articulatory movements are more intense in [a], [o], and [e] compared to [i] and [ɯ]. Therefore, three open or close-mid vowel sounds had a certain amount of articulatory movement, and the accuracy of MCD was improved by using the EMA data as intermediate information. However, the MCDs of [i] and [ɯ]

Table 4 MCD of consonants categorized by articulation points

Model	MCD(dB)						
	Labiodental	Palatal	Bilabial	Alveolar	Velar	Retroflex	Glottal
(1) EMA	17.59	13.78	14.11	12.60	14.85	15.37	17.58
(2) IPA	12.65	10.20	10.68	10.38	11.68	**11.10**	**10.83**
(3) IPA+EMA	**11.24**	**9.26**	**10.08**	**10.08**	**11.65**	11.31	11.17

were the best with the (2) IPA Model. These two vowel sounds were close vowels, and Fig. 5 indicates that close vowels have less articulatory movement. Thus, the MCDs in these two vowel sounds were not improved by using the EMA data as intermediate information.

Focusing on the MCD for each vowel in Fig. 3, the correlation coefficient between the MCD and RMSE of the lower jaw back-forth position of [o] ([o]-LJ-BF) was the highest at 0.72, and the correlation coefficient between the MCD and RMSE of the lower lip back-forth position of [e] ([e]-LL-BF) was the second highest at 0.68. Fig. 6 (a) and (b) show the scatter plots and the regression lines which were depicted from data [o]-LJ-BF and [e]-LL-BF, respectively. The results indicate that the higher the accuracy of the estimation lower jaw or lower lip back-forth position in the EMA, the better the MCD values of [o] or [e]. The vowel [o] is a close-mid back rounded vowel, pronounced with lip rounding. Because the lower jaw moves to create lip rounding, a high correlation was obtained. Fig. 7 shows boxenplots that display the standardized EMA data of [e] categorized by measurement points of EMA in the test data. Focusing on the mean values, LL-BF, UL-BF, and LJ-BF in this order, were the farthest from the mean value of zero. In other words, compared to other sounds except [e], the articulatory organs of these EMA points moved in unique positions. Therefore, it is considered that a high correlation coefficient between the MCD and [e]-LL-BF was observed. Additionally, [e]-UL-BF and [e]-LJ-BF also showed relatively high correlation coefficients when compared to other vowels.

However, the correlation coefficients between the MCD of close vowels [i][ɯ] and the RMSE of EMA data were lower compared to those of open or close-mid vowels. This indicates that no correlation between the estimation accuracy of the EMA data and MCD of close vowels existed. This may also be owing to the less articulatory movement of close vowels.

5.4 MCD for consonant categories

Table 4 compares three models using the MCD of each consonant articulation point. The MCD of labiodental, palatal, bilabial, alveolar, and velar sounds was the best with the (3) IPA+EMA Model. These sounds involved the lips or tongue as the place of articulation, which could be supplemented with EMA data, resulting in improved MCD.

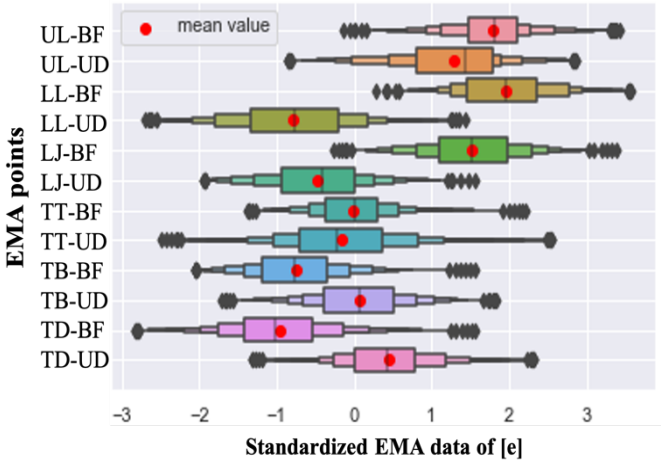

Fig. 7 Boxenplot of standardized [e]-EMA data for each EMA point of measurement in test data.

However, the MCD of the retroflex and glottal sounds was the best with the (2) IPA Model. Retroflex sounds are pronounced when the curled tongue approaches the hard palate. The EMA data do not contain information about the hard palate and the tongue movements are complex. Thus, the accuracy of MCD in the sounds was not improved by using the EMA data as intermediate information. Additionally, EMA data do not contain information about the glottis which is the articulation point of the glottal sound [h]; hence, the (2) IPA Model had the best results in the MCD of glottal sounds.

6. Conclusions and Future Work

In this study, we evaluated three models for generating speeches from IPA sequences through EMA data, which contains a wealth of information about articulatory movement. Experiments confirmed that using EMA data as intermediate information improves the quality of the synthesized speech. In particular, the effectiveness of using EMA data as intermediate information was confirmed in the case of speech where articulation points were included in the EMA data.

However, even when using EMA data, the estimation accuracy of F_0 was low. Therefore, in the future, we will investigate methods to accurately estimate F_0 using EMA data as input.

7. ACKNOWLEDGMENT

This work was supported by the JSPS KAKENHI Grant Numbers JP22K12100 (Kouichi Katsurada), and JP24K00071 (Kikuo Maekawa).

REFERENCES

[1] R. Tanji, H. Ohmura, and K. Katsurada, "Using Transposed Convolution for Articulatory-to-Acoustic Conversion from Real-Time MRI Data," *INTERSPEECH*, pp. 3176-3180, 2021.

[2] S. Udupa, and PK. Ghosh, "Real-Time MRI Video synthesis from time aligned phonemes with sequence-to-sequence networks," *ICASSP*, pp. 1-5, 2023.

[3] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," *INTERSPEECH*, pp. 127-131, 2023.

[4] Z. Wei, Z. Wu, and L. Xie, "Predicting articulatory movement from text using deep architecture with stacked bottleneck features," *ICASSP*, pp. 1-6, 2016.

[5] ZC. Liu, ZH. Ling, and LR. Dai, "Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks," *INTERSPEECH*, pp. 1502-1506, 2016.

[6] K. Katsurada, and K. Richmond, "Speaker-independent mel-cepstrum estimation from articulator movements using D-vector input," *INTERSPEECH*, ISCA, pp. 3176-3180, 2020.

[7] YW. Chen, KH. Hung, SY. Chuang, and J. Sherman, et al., "EMA2S: An End-to-End Multimodal Articulatory-to-Speech System," *ISCAS*, pp. 1-5, 2021.

[8] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," *the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-11, 2019.

[9] G. Gosztolya, T. Grósz, T. Tóth, and L. Markó, et al., "Applying DNN adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces," *Acta Polytechnica Hungarica*, 17(7), 109-124, 2020.

[10] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *ICASSP*, 1993, pp. 125–128.

[11] AW. Rix, JG. Beerends, MP. Hollier, and AP. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *ICASSP*, pp. 749–752, 2001.

[12] K. Maruyama, S. Sawada, H. Ohmura, and K. Katsurada, "Generation of Articulatory Movement Data from IPA using 1DCNN," *ASJ*, 3-Q-38, September 2023.

[13] PW. Schönle, K. Gräbe, P. Wenig, and J. Höhne, et al., "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, pp. 26–35, 1987.

[14] H. Horn, G. Göz, M. Bacher, and M. Müllauer, et al., "Reliability of electromagnetic articulography recording during speaking sequences," *Eur. J. Orthod,* pp. 647–655, 1997.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CVPR*, pp. 7132-7141, 2018.

[16] M. Kim, Z. Piao, J. Lee, and HG. Kang, "Style Modeling for Multi-Speaker Articulation-to-Speech," *ICASSP*, pp. 1-5, 2023.

[17] A. Gulati, J. Qin, CC. Chiu, and N. Parmar, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv*:2005.08100, 2020.

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, pp. 17022-17033, 2020.

[19] K. Wakamiya, F. Taguchi, R. Watanabe, and K. Katsurada, et al., "Collection of large-scale Japanese articulatory-acoustic parallel data," *IEICE Technical Report*, pp. 7-12, June 2019.

[20] A. Kurematsu, K. Takeda, Y. Sagisaka, and S. Katagiri, et al., "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," *Speech Commun*, vol. 9, no. 4, pp. 357–363, 1990.

[21] S. Takamichi, K. Mitsui, Y. Saito, and T. Koriyama, et al., "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint*, 1908.06248, Aug 2019.

[22] M. Morise, "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals," *INTERSPEECH*, pp. 2321-2325, 2017.

[23] A. Lee, and T. Kawahara, "Recent development of open-source speech recognition engine Julius," *APSIPA*, pp. 131-137, 2009.