INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Synthesis from Articulatory Movements Recorded by Real-time MRI

Yuto Otani, Shun Sawada, Hidefumi Ohmura, Kouichi Katsurada

Department of Information Sciences, Tokyo University of Science, Tokyo, Japan
6319023@ed.tus.ac.jp, {sawada, katsurada}@rs.tus.ac.jp, ohmura@is.noda.tus.ac.jp

Abstract

Previous speech synthesis models from articulatory movements recorded using real-time MRI (rtMRI) only predicted vocal tract shape parameters and required additional pitch information to generate a speech waveform. This study proposes a two-stage deep learning model composed of CNN-BiLSTM that predicts a mel-spectrogram from a rtMRI video and a HiFi-GAN vocoder that synthesizes a speech waveform. We evaluated our model on two databases: the ATR 503 sentences rtMRI database and the USC-TIMIT database. The experimental results on the ATR 503 sentences rtMRI database show that the PESQ score and the RMSE of F_0 are 1.64 and 26.7 Hz. This demonstrates that all acoustic parameters, including fundamental frequency, can be estimated from the rtMRI videos. In the experiment on the USC-TIMIT database, we obtained a good PESQ score and RMSE for F_0. However, the synthesized speech is unclear, indicating that the quality of the datasets affects the intelligibility of the synthesized speech.

Index Terms: real-time MRI, articulatory movement, speech synthesis, speech waveform generation

1. Introduction

As neural networks are commonly used in text-to-speech (TTS), the quality of synthesized speech has become indistinguishable from human speech [1]. However, general TTS models capture the statistical relationship between text and speech and thus do not consider actual articulatory movements. This study proposes a speech synthesis model that uses articulatory movements as the inputs for generating speech. It would be a fundamental component of a text-to-MRI and MRI-to-speech pipeline, anticipated to be available in computer-assisted language learning (CALL) systems and offering substantial support for individuals with dysarthria.

Several methods have been proposed to capture articulatory movements. These include electromagnetic articulography (EMA), which measures the movement of the coils attached to the articulators such as the lips and tongue [2, 3]; ultrasound tongue imaging (UTI), which captures the movement of the tongue using ultrasound [4, 5]; and real-time magnetic resonance imaging (rtMRI), which records the mid-sagittal plane of the upper airway using fast MRI [6, 7, 8, 9]. UTI can easily record a high-frame-rate video at approximately 100 fps. However, it can only record tongue movements. The EMA can be recorded at a high sampling rate of approximately 500 Hz. However, it can only provide position information regarding several points where the sensors are attached. By contrast, rtMRI can record all articulatory organs, including the soft palate and larynx, which are challenging to record using other methods. Although rtMRI videos have a relatively low frame rate of approximately 30 fps, they contain considerable helpful information for speech synthesis owing to their high spatial resolution. Therefore, we used rtMRI videos as the inputs for speech synthesis.

Several models for synthesizing speech from rtMRI videos have been proposed, in which the MGC-LSP or mel-cepstrum are estimated from a series of MRI images. However, the estimated parameters only contain vocal tract shape information; hence, they require additional information about the fundamental frequency to synthesize a speech waveform. Although a model of speech synthesis from EMA has been proposed [10], few study have generated speech from rtMRI videos directly. This study proposes a two-stage model composed of CNN-BiLSTM that predicts a mel-spectrogram as an intermediate representation from a rtMRI video and a HiFi-GAN vocoder that synthesizes a speech waveform. We evaluate our model on the ATR 503 Japanese sentences rtMRI and the USC-TIMIT MRI databases. We show that the two-stage model can generate reasonable speech sounds with the adequate fundamental frequency.

2. Related work

2.1. Estimation of MGC-LSP from rtMRI videos

Csapó proposed three models: FC-DNN, CNN, and CNN-LSTM, using rtMRI videos [11] to estimate MGC-LSP [12], an acoustic feature representing vocal tract shape. They showed that the CNN-LSTM model had the highest estimation accuracy. The model consisted of a CNN part with three convolutional and max pooling layers, an LSTM part with two LSTM layers, and two densely connected layers. This model is reasonable because the CNN part extracts the image features from each frame of the rtMRI video, and the LSTM part captures the time-series features. Because the MGC-LSP estimated in this model is a vocal tract shape parameter, it is necessary to provide LP residual signals extracted from the original speech to synthesize a speech waveform.

2.2. Estimation of mel-cepstrum from rtMRI videos using temporal super-resolution with transposed convolution

Tanji et al. [13] proposed a model that applies a super-resolution process along the temporal dimension using transposed convolution to estimate the mel-cepstrum from a rtMRI video. Although rtMRI has the advantage of recording entire articulatory organs with high spatial resolution, it has a disadvantage because the temporal resolution is comparatively lower than other methods for recording articulatory movements. They increased the temporal resolution by using transposed convolution to address this issue. Their proposed CNN-TC-LSTM model, based on the CNN-LSTM model proposed by Csapó, inserts a transposed convolutional layer between the CNN and LSTM parts. A convolutional and a max pooling layers were added to the end of the CNN part, and the densely connected layers were eliminated from the CNN-LSTM model to optimize the model. The CNN-TC-LSTM model achieved a

10.21437/Interspeech.2023-286

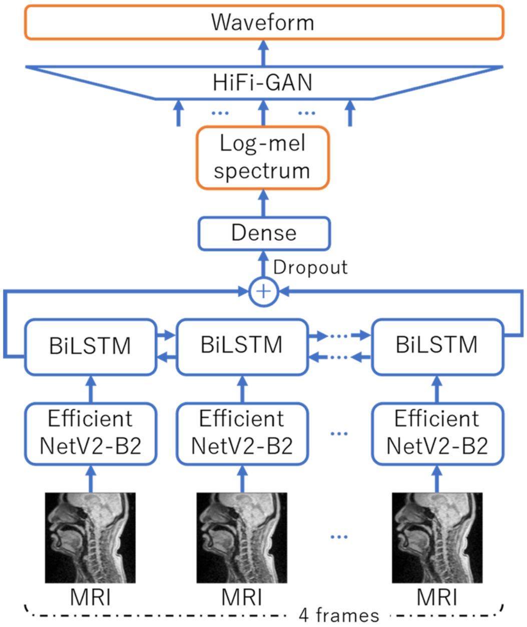

Figure 1: *Outline of proposed model*

better MCD [14] in the estimated mel-cepstrum and better perceptual evaluation of speech quality (PESQ) in the synthesized speeches than the CNN-LSTM model. However, the model also requires fundamental frequency information and an aperiodicity index of the original speech to synthesize a speech waveform.

3. Proposed model

We propose a model that estimates the mel-spectrogram as an intermediate representation and generates a speech waveform using a neural vocoder. Figure 1 illustrates the outline of the proposed model. The input rtMRI video is processed with a two-stage network. The first stage estimates the mel-spectrogram from the rtMRI video, and the second stage synthesizes a speech waveform from the estimated mel-spectrogram.

We constructed a new model that modifies the CNN-LSTM model for mel-spectrogram estimation, showing high estimation accuracy in related studies. Table 1 lists the configuration details. The network estimated the corresponding spectrum from four consecutive frames of the rtMRI video. The CNN was implemented using EfficientNetV2 [15], which achieved high accuracy and superior efficiency in image classification tasks. EfficientNetV2 includes various models with different network depths, widths, and input image sizes. We employed the B2 model because the rtMRI image size was suitable for this model. We used up to the global average pooling layer of the B2 model and removed the densely connected layer that was appended for classification in the original B2 model. We enabled more complex image feature extraction in the CNN part by replacing the simple convolution and max pooling layers in the CNN-LSTM model with EfficientNetV2. We did not insert the transposed convolution layer in the CNN part because the up-sampling process was conducted in the HiFi-GAN vocoder. In the LSTM part, we used a single bidirectional LSTM (BiLSTM) layer to extract bidirectional time-series features. The number of hidden units was set to 640, and the bidirectional outputs were merged by adding them. After a dropout [16] rate of 0.5 was applied to the outputs of the BiLSTM layer to suppress overfitting, they were sent to a densely connected layer. This first stage has 19M parameters.

Table 1: *Configuration of the first stage network*

Operator		Timesteps	Size	Layers
(Input)		4	256, 256, 1	-
ENv2 -B2	Conv3x3	4	128, 128, 32	1
	Fused-MBConv	4	32, 32, 56	5
	MBConv	4	8, 8, 208	16
	Conv1x1	4	8, 8, 1408	1
	Pooling	4	1, 1, 1408	-
BiLSTM		1 fwd/1 bwd	1, 1, 640	1
Summation		1	1, 1, 640	-
Dropout		1	1, 1, 640	-
Dense		1	1, 1, 64	1

We used HiFi-GAN [17], an efficient and high-fidelity neural vocoder for speech synthesis. HiFi-GAN is a GAN-based non-autoregressive neural vocoder designed to efficiently capture periodic patterns in speech waveforms, thus enabling high-quality and fast speech synthesis. We used the HiFi-GAN V1 model, which has the largest model size and the highest quality of generated speech. The HiFi-GAN generator consists of combinations of convolutional and transposed convolutional layers. It generates speech waveforms by repeating up-sampling along the time dimension with transposed convolution. In the proposed model, the mel-spectrum frame period corresponds to the frame interval of the rtMRI videos. Therefore, the detailed up-sampling factors in the HiFi-GAN generator vary according to the frame rates of the rtMRI videos and the generated waveforms in the datasets.

4. Experimental setup

4.1. Datasets

4.1.1. ATR 503 sentences rtMRI database

Currently under construction, the ATR 503 sentences rtMRI database[1] [18] contains rtMRI videos of ATR 503 phoneme-balanced sentences [19] read by a single male speaker. The database consists of 10 sets (A–J) of 50 phoneme-balanced spoken Japanese sentences (set J contains 53). Videos of the mid-sagittal plane of the head, including the entire vocal tract and audio of spoken sentences, were recorded using MRI equipment. The resolution of the video was 256 × 256 pixels, the frame rate was 27.17 fps, and the audio sampling rate was 44,100 Hz. In the experiment, sets A to I were used as the training data and set J was randomly divided into the validation and test data.

4.1.2. USC-TIMIT MRI database

The USC-TIMIT MRI database [20] comprises rtMRI videos of same sentences as the MOCHA-TIMIT corpus [21] read by five male and female American English speakers. The resolution of the video was 68 × 68 pixels, the frame rate was

[1] https://rtmridb.ninjal.ac.jp/

Table 2: *Hyperparameters of HiFi-GAN generators*

Model	Stride	Kernel size
Original HiFi-GAN V1	(8, 8, 2, 2)	(16, 16, 4, 4)
ATR503 database	(10, 7, 3, 2)	(20, 15, 7, 4)
USC-TIMIT database	(8, 8, 4, 2)	(16, 16, 8, 4)

23.18 fps, and the audio sampling rate was 20,000 Hz. Because the audio contains loud noise generated during MRI recording, noise reduction was applied using a custom adaptive filter [22]. In the experiment, we used the data from a male speaker (M3) randomly split into 8:1:1 subsets: training, validation, and test.

4.2. Preprocessing

4.2.1. ATR 503 sentences rtMRI database

For the rtMRI videos, luminance normalization was applied to smoothen the luminance changes between frames. For speech audio, noise reduction with spectral subtraction [23] removes loud noises generated by MRI equipment. The sampling rate of the audio is down-sampled to 11,413 Hz such that the frame rate of the rtMRI videos becomes approximately a factor of the audio sampling rate, which is convenient for introducing HiFi-GAN in the original design. The number of dimensions of the mel-filter bank was set to 64. The resulting mel-power spectrum was compressed to a decibel scale and standardized.

4.2.2. USC-TIMIT MRI database

Because the size of the MRI videos in the USC-TIMIT MRI database is small (68 × 68 pixels), up-sampling by a factor of three is applied using a bilinear method to match the EfficientNetV2 input image size. Although the audio clips are denoised using an adaptive filter, they still contain MRI equipment noises that have not been entirely removed and reverberations that the adaptive filter may have caused. Because these noises and reverberations make model training challenging, we used the NVIDIA MAXINE Audio Effects SDK [1], which provides deep-learning-based voice quality improvement tools to reduce them. For the same reason as in the ATR 503 sentences rtMRI database, the sampling rates of the audio clips were down-sampled to 11,866 Hz.

4.3. Training setup of CNN-BiLSTM model

The loss function employed for training the CNN-BiLSTM model is the mean squared error (MSE) between the log-mel-spectrum of the original speech waveform and the estimated one. AdaBelief [24] was used as the optimization algorithm. The learning rate was set to 0.001 at the start of training and decayed by a factor of 10 for every four epochs of stagnation of loss improvement in the validation data. The training was terminated when the loss improvement stagnated for eight epochs, and the weight with the lowest loss in the validation data was adopted. The training was converged in 7.5 hours using an NVIDIA GeForce RTX 3090. The source code is available in our repository.[2]

4.4. Training setup of HiFi-GAN

In the original HiFi-GAN V1 model, the up-sampling factor, namely the stride of the four transposed convolutional layers of the generator, was (8, 8, 2, 2), and the kernel size was (16, 16, 4, 4). As shown in Table 2, the up-sampling factor and kernel size are modified to (10, 7, 3, 2) and (20, 15, 7, 4) in the experiments on the ATR 503 sentences rtMRI database, respectively. In the experiments on the USC-TIMIT database, they were changed to (8, 8, 4, 2) and (16, 16, 8, 4). These factors were decided to match the frame rate of the rtMRI videos and the audios in each database. Our experiments reduced the number of dimensions of the mel-filter bank from 80 to 64. The other designs and training settings were the same as those of the official HiFi-GAN implementation[3].

Table 3: *PESQ, F_0 RMSE (Hz), and V/UV error (%) on the ATR 503 sentences rtMRI database*

Evaluated speech	PESQ	F_0 RMSE	V/UV
Predicted	1.64	26.7	3.6
Copy synthesized	2.97	21.3	3.5

The ATR 503 sentences rtMRI and the USC-TIMIT databases' total speech length is approximately 50 min and 35 min, respectively, insufficient to train HiFi-GAN from scratch. Therefore, we conducted pretraining using substantial speech corpora. In the ATR 503 sentences rtMRI database experiments, we used the Japanese versatile speech (JVS) corpus [25], which contains about 26 hours of reading by many speakers. In the USC-TIMIT experiment, we used the LJ Speech Dataset [26], an English speech corpus of approximately 24 hours readings. Fine-tuning was applied using training data from the rtMRI datasets.

4.5. Evaluation metrics

The PESQ narrowband [27] was employed to evaluate the synthesized speech's quality objectively. PESQ ranges from -0.5 to 4.5, with higher values corresponding to better speech quality. The F_0 root mean squared error (RMSE) and voiced/unvoiced (V/UV) error were used to evaluate the accuracy of the estimated F_0 objectively. The correct F_0 is extracted using Harvest [28], the F_0 extractor provided by the WORLD vocoder [29]. F_0 RMSE is calculated only from the segments where both the original speech and the synthesized speech are determined to be voiced, and the V/UV error is calculated only from the speech segments.

5. Results and discussion

Table 3 shows the experimental results for the ATR 503 sentences rtMRI database. The PESQ score of the speech synthesized using the estimated mel-spectrogram was 1.64. The F_0 RMSE and V/UV errors were 26.7 Hz and 3.6%, respectively. These values for the copy synthesized speech obtained by inputting the ground truth mel-spectrogram into HiFi-GAN are 21.3 Hz and 3.5%, respectively. The results indicate that the generated speeches have almost the same quality in fundamental frequency and voiced/unvoiced judgment as the original speeches. Figure 2 shows the mel-spectrograms of the original speech and the speech synthesized from a rtMRI video. Although there were minor differences in detail, the overall spectral structure was well reconstructed. Therefore, the proposed model achieved the goal of this study.

Figure 3 shows an example of an F_0 trajectory of original speech and speech synthesized from a rtMRI video. The outline

[1] https://github.com/NVIDIA/MAXINE-AFX-SDK

[2] https://github.com/y-otn/mri-to-speech

[3] https://github.com/jik876/hifi-gan

Table 4: *PESQ, F_0 RMSE (Hz), and V/UV error (%) on the USC-TIMIT database*

Evaluated speech	**PESQ**	**F_0 RMSE**	**V/UV**
Predicted	2.07	27.6	17.2
Copy synthesized	3.15	24.5	12.0

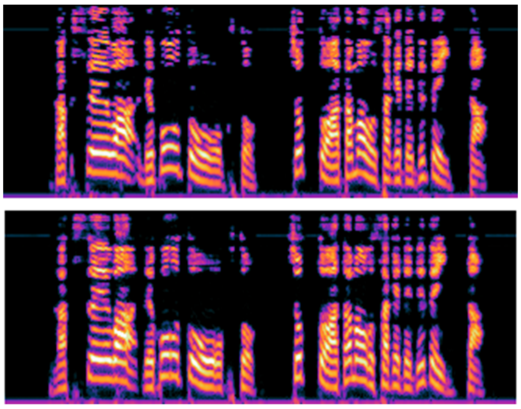

Figure 2: *Mel-spectrograms of the original speech (top) and the speech synthesized from rtMRI video (bottom)*

of F_0 of the predicted speech almost traces the original speech. However, we found a minor difference between them at approximately 5 seconds. This type of difference, thought to be caused by the strength of expiratory pressure that cannot be captured with rtMRI equipment, is sometimes found in the results. This difference slightly impacts the quality of the generated sound, such as pitch intonation. Some sample audio clips are available on our website[1].

Table 4 shows PESQ, F_0 RMSE, and the V/UV errors on the USC-TIMIT database. PESQ and F_0 RMSE of the speech synthesized from the estimated mel-spectrogram were 2.07 and 27.6 Hz, respectively. In contrast, the V/UV error was 17.2%, worse than the ATR 503 sentences rtMRI database experiment. Although the PESQ score is better than that on the ATR 503 sentences rtMRI database, the generated speeches have poorer speech quality. This is also observed when the generated speeches are evaluated using an ASR system. We calculated the generated speeches' word error rate (WER) using Microsoft Azure Speech to Text. The WER of the speech generated from the rtMRI videos on the USC-TIMIT database was 102.6%, showing that the speech quality was terrible, whereas that of the ATR 503 sentences was 0.7%, making it evident that the speech quality is excellent.

The quality of the recorded speeches and rtMRI videos seems to cause this significant difference between the generated speeches. The V/UV error of the reconstructed speech obtained by inputting the ground truth mel-spectrogram into HiFi-GAN is 12.0% in the USC-TIMIT database, while that on the ATR 503 sentences rtMRI database is 3.5%. This shows that the quality of speech in the USC-TIMIT database is worse than that in the ATR 503 sentences rtMRI database. Additionally, in silent speech recognition, the inputs for the recognizer are USC-TIMIT rtMRI videos, and the outputs are texts; the WER is reported to be 42.1% [30]. This result suggests that capturing speech features from the USC-TIMIT rtMRI videos is challenging. As shown in Figure 4, the resolution of USC-TIMIT MRI images is low. The quality is always unstable owing to intense noise, artifacts, and uneven luminance in some areas of the image [31]. In addition, there were several parts where the speech synthesized from the rtMRI video was not in sync with the video at 0.5 seconds, caused by some misalignments in the database. In contrast, the ATR 503 sentences rtMRI database has excellent image quality and does not have such problems, which yields accurate results in generating high-quality speech.

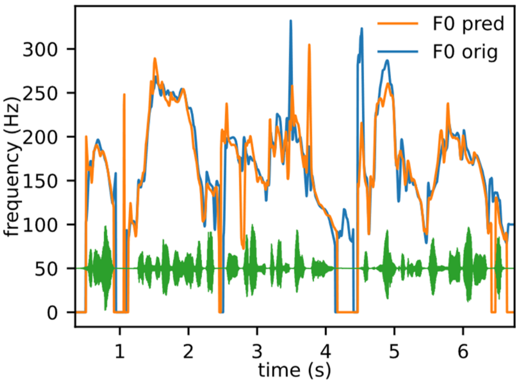

Figure 3: *F_0 trajectories of the original speech and the speech synthesized from rtMRI video*

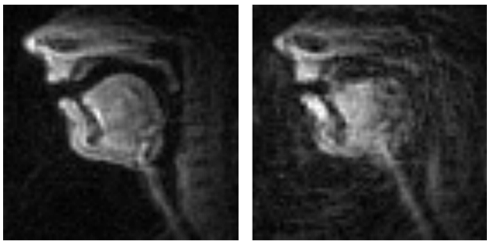

Figure 4: *MRI images of the USC-TIMIT*

6. Conclusions

This study proposes a two-stage model for estimating the mel-spectrogram from rtMRI videos, enabling speech waveform synthesis from rtMRI videos. The ATR 503 sentences rtMRI database experiment confirmed that the synthesized speech's F_0 RMSE and V/UV errors were low, and the speech content was accurately reproduced. These results show that all acoustic parameters, including fundamental frequency, can be estimated from only four frame consecutive rtMRI images, while a previous study demonstrated it was possible from 13 frames [32]. However, in the experiment on the USC-TIMIT database, the generated speech lacked intelligibility. Although an apparent reason has not been identified, the quality of the USC-TIMIT MRI database is expected to have several problems. We would like to evaluate our approach with other high-quality rtMRI datasets [33].

In the future, we plan to investigate why F_0 and voiced/unvoiced were well estimated from rtMRI videos in which vocal fold vibration was not recorded.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP19K12024, JP22K12100 (Kouichi Katsurada), and JP20H01265 (Kikuo Maekawa).

[1] https://y-otn.github.io/mri-to-speech-demo/

8. References

[1] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao, and T.-Y. Liu, "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," *arXiv preprint arXiv:2205.04421*, 2022.

[2] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader and B. Conrad, "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.

[3] H. Horn, G. Göz, M. Bacher, M. Müllauer, I. Kretschmer and D. Axmann-Krcmar, "Reliability of electromagnetic articulography recording during speaking sequences," *European Journal of Orthodontics*, pp. 647–655, 1997.

[4] Y. Akgul, C. Kambhamettu and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," in *Proc. CVPR*, 1998, pp. 298–303.

[5] Maureen Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[6] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1771–1776, 2004.

[7] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 123–132, 2008.

[8] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Toger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Computer Speech and Language*, vol. 52, pp. 1–22, 2018.

[9] A. Toutios, D. Byrd, L. Goldstein, and S. Narayanan, "Advances in vocal tract imaging and analysis," *The Routledge Handbook of Phonetics*, pp. 34–50, 2019.

[10] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep Speech Synthesis from Articulatory Representations," in *Proc. INTERSPEECH*, 2022, pp. 779-783.

[11] T. G. Csapó, "Speaker Dependent Articulatory-to-Acoustic Mapping Using Real-Time MRI of the Vocal Tract," in *Proc. INTERSPEECH*, 2020, pp. 2722–2726.

[12] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Melgeneralized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, pp. 1043–1046.

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

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

[15] M. Tan, and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. ICML*, 2021, pp. 10096–10106.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

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

[18] H. Takemoto, T. Goto, Y. Hagihara, S. Hamanaka, T. Kitamura, Y. Nota, and K. Maekawa, "Speech Organ Contour Extraction Using Real-Time MRI and Machine Learning Method," in *Proc. INTERSPEECH*, 2019, pp. 904–908.

[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[20] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[21] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 305–308.

[22] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[23] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, vol. 27, no. 2, pp. 113–120, 1979.

[24] J. Zhuang, T. Tang, Y. Ding, S. C. Tatikonda, N. Dvornek, X. Papademetris, and J. Duncan, "AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients," in *Proc. NeurIPS*, 2020, pp. 18795–18806.

[25] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[26] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[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, pp. 749–752.

[28] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH*, 2017, pp. 2321–2325.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A VocoderBased High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[30] L. Pandey, and A. S. Arif, "Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI," in *Proc. SIGGRAPH*, 2021, no. 27.

[31] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech mri," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[32] Y. Yu, A. H. Shandiz and L. Tóth, "Reconstructing Speech from Real-Time Articulatory MRI Using Neural Vocoders," in *Proc. EUSIPCO*, 2021, pp. 945-949.

[33] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P.-A. Vuissoz, "Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers," *Scientific Data*, vol. 8, no. 258, 2021.