INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Using Transposed Convolution for Articulatory-to-Acoustic Conversion from Real-Time MRI Data

Ryo Tanji, Hidefumi Ohmura, Kouichi Katsurada

Department of Information Sciences, Tokyo University of Science, Tokyo, Japan

6317072@ed.tus.ac.jp, ohmura@is.noda.tus.ac.jp, katsurada@rs.tus.ac.jp

Abstract

We herein propose a deep neural network-based model for articulatory-to-acoustic conversion from real-time MRI data. Although rtMRI, which can record entire articulatory organs with a high resolution, has an advantage in articulatory-to-acoustic conversion, it has a relatively low sampling rate. To address this, we incorporated the super-resolution technique in the temporal dimension with a transposed convolution. With the use of transposed convolution, the resolution can be increased by applying the inversion process of resolution reduction of a standard CNN. To evaluate the performance on the datasets with different temporal resolutions, we conducted experiments using two datasets: USC-TIMIT and Japanese rtMRI dataset. Results of the experiments performed using mel-cepstrum distortion and PESQ showed that transposed convolution is effective for generating accurate acoustic features. We also confirmed that increasing the magnification of the super-resolution leads to an improvement in the PESQ score.

Index Terms: rtMRI, articulatory-to-acoustic conversion, transposed convolution

1. Introduction

Articulatory movement, which is the movement of articulators such as the tongue, lips, and jaws, is a fundamental motion when producing speech sounds. These movements can be recorded using equipment such as electromagnetic articulography (EMA) [1], ultrasound tongue imaging (UTI) [2], permanent magnetic articulography (PMA) [3], surface electromyography (sEMG) [4], and video of lip movements [5]. In recent years, real-time magnetic resonance imaging (rtMRI), which uses powerful magnetic forces and radio waves to visualize organs in the body, is now being used for recording articulatory movements [6, 7, 8]. The advantage of rtMRI is that dynamic information of the entire midsagittal plane of the upper airway can be obtained with a relatively high image resolution, even when a speaker produces continuous speech. It can capture not only lingual, labial, and jaw motions, but also the articulation of the velum and the pharyngeal region, which is difficult to record with other articulatory acquisition techniques such as EMA.

Articulatory-to-acoustic conversion has been investigated for the generation of acoustic features from the recorded movements of articulators using statistical models that can represent the relationship between articulatory movements and changes in acoustic features [9]. With significant advances in machine learning, recent studies have demonstrated the use of deep neural networks (DNNs) for articulatory-to-acoustic conversion [10, 11, 12, 13, 14]. RtMRI is a potentially suitable technique for articulatory-to-acoustic conversion. In fact, articulatory-to-acoustic conversion from rtMRI using a DNN model has shown better results than that obtained from ultrasound tongue imaging using the same model [15]. However, rtMRI has a relatively low sampling rate due to the limitations of MRI imaging. In speech synthesis, it has been shown that a long frame period leads to a deterioration in the sound quality [16, 17]. Therefore, techniques to increase the sampling rate, such as temporal super-resolution, are required to improve the sound quality in articulatory-to-acoustic conversion from rtMRI.

In this paper, we propose an articulatory-to-acoustic conversion using transposed convolution (also called fractionally strided convolution or deconvolution) [18]. By implementing transposed convolution, the resolution can be increased by applying the inversion process of resolution reduction of a standard CNN. Although it has mainly been used to increase the resolution of the image [19, 20, 21], it has also been applied to speech enhancement [22] in the recent years. Transposed convolution can perform super-resolution and parameter learning simultaneously, resulting in higher accuracy than interpolation-based methods such as spline interpolation [23]. To evaluate the effectiveness of transposed convolution, we evaluated the performance on two datasets with different temporal resolutions. We also confirm how the magnification of super-resolution affects the performance of articulatory-to-acoustic conversion by comparing different sizes of stride in the transposed convolution.

2. Related work

Csapó tested three models for articulatory-to-acoustic conversion from rtMRI [15]: (1) fully connected DNN, (2) CNN, and (3) CNN-LSTM, and showed that the CNN-LSTM network is the most effective for the conversion. The CNN-LSTM network is a trainable model that extracts features from each image using a CNN-based deep visual feature extractor and estimates the latent variables over the time domain using an LSTM-based temporal dynamics extractor. This is the same neural network model as the one used by the author in his previous experiments on articulatory-to-acoustic conversion using ultrasound tongue imaging [12].

The structure of the CNN-LSTM is shown in Figure 1. It has three convolutional layers (filter size: 3 × 3, stride: 1 × 1, number of filters: 8, 16, 32) with max pooling (filter size: 2 × 2, stride: 2 × 2). The output of the convolution layers is given to a two-layer LSTM with 512 units for the extraction of the temporal features. Finally, the acoustic features were extracted using two linear layers with 512 units, followed by a linear layer with 40 units of output. ReLU activation was used for all the hidden layers.

They estimated the MGC-LSP features extracted using the MGLSA vocoder [24]. Their results showed that the mel-

http://dx.doi.org/10.21437/Interspeech.2021-906

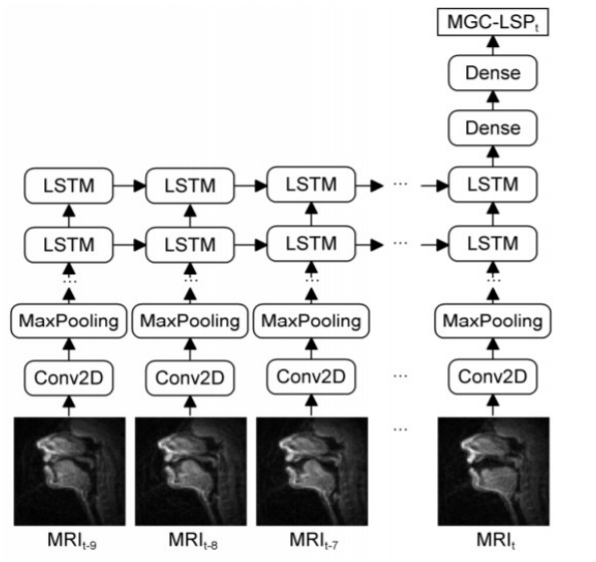

Figure 2: *Architecture of the CNN-LSTM [15]*

cepstrum distortion between the generated acoustic features and that of the original sound is better than that of the other two models. They also confirmed that the score is better than that using the ultrasound tongue images, which demonstrates that rtMRI is better than ultrasound tongue imaging in articulatory-to-acoustic conversion.

3. Articulatory-to-acoustic conversion using transposed convolution

A straightforward way to generate acoustic features from articulatory movement data is to convert the articulatory movement to an acoustic feature frame by frame. In this case, the acoustic and articulatory movements have the same temporal resolution. Therefore, if the temporal resolution of the data obtained using rtMRI is poor, the CNN-LSTM model generates acoustic features with a relatively long frame period. Because long frame periods lead to the deterioration of sound quality, some complementary techniques, such as super-resolution in the temporal dimension, should be applied. We employed transposed convolution to increase the temporal resolution.

Transposed convolution achieves accurate super-resolution by inverting the resolution reduction process of a standard CNN. An example of transposed convolution, where the length of the input data is three, the filter size is four, and the size of the stride is two, is shown in Figure 2. The output vector is generated as follows: First, the product of each element in the input and a filter vector is calculated. Next, the obtained vectors corresponding to elements in the input are added together while shifting the stride size. Finally, the last several elements, whose size is the filter size minus strides, are deleted from the vector. The size of the final output is the input size multiplied by the stride. Therefore, if the size of the stride is represented by *s*, every *element* in the output vector corresponds to an element in the input. As the deletion of the last several elements in the final step makes the output elements not to refer to the future input information, this process can be regarded as a causal convolution.

The structure of the proposed model is illustrated in Figure

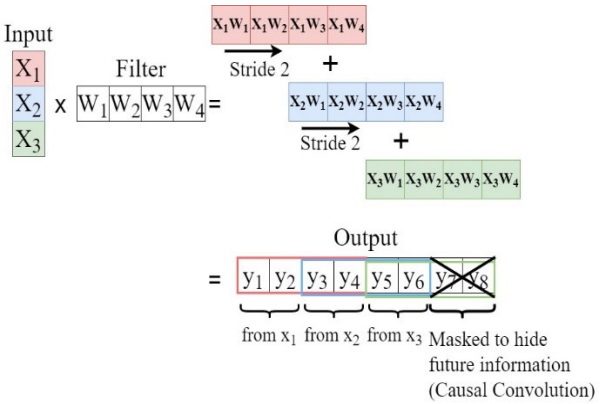

Figure 1: *Example of a transposed convolution*

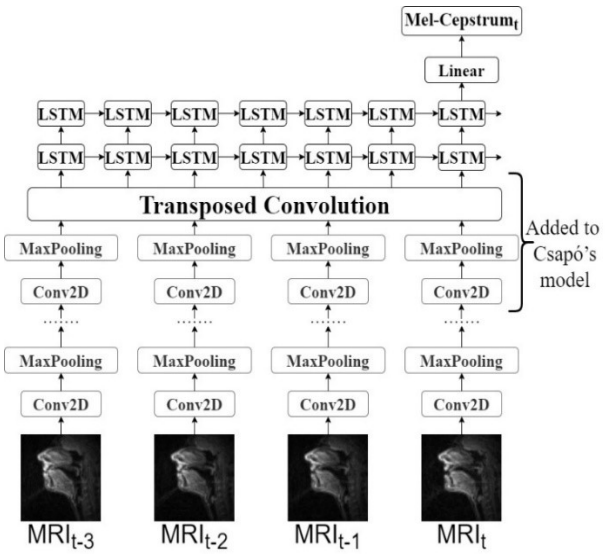

Figure 3: *Architecture of the CNN-TC-LSTM*

3. Transposed convolution was inserted between the CNN and the LSTM layers of Csapó's model (3). In addition, unlike Csapó's approach, we added a convolutional layer and a max-pooling layer in addition to Csapó's model (3). This addition has two advantages when training deep neural networks. One merit is the more complex relationship between the rtMRI videos and the acoustic features that can be modeled by incorporating additional layers. The other advantage is that the input dimension of the transposed convolution and the LSTM layer can be reduced by adding an additional max-pooling layer. Therefore, the addition of this layer leads to a reduction in the total number of parameters. Additionally, in this model, all the linear layers except the last one after the LSTM layers were removed to reduce the number of parameters.

4. Experimental setup

4.1. Dataset

To perform experiments with data of various temporal resolutions, we evaluated our model on two datasets: USC-TIMIT and Japanese rtMRI dataset.

4.1.1. USC-TIMIT

USC-TIMIT [6] is a speech production database that includes rtMRI data of five male and five female speakers of American English, and EMA data of five of these speakers. These two modalities were recorded in two independent sessions while the subjects spoke the same 460-sentences from the MOCHA-TIMIT corpus. We used only the rtMRI data from this corpus. The articulatory data were recorded with an image resolution of 68 × 68 pixels and a video frame rate of 23.18 fps. The audio was recorded simultaneously at a sampling frequency of 20 kHz inside the MRI scanner. Noise cancelation was applied using a custom adaptive signal processing algorithm, which exploits the periodic structure of the noise generated by the MRI scanner [25].

In the experiment, the sampling frequency was up sampled to 23,180 Hz to match the video frequency. We used two males ('m2' and 'm3')[1] and two female speakers ('f1' and 'f2').

4.1.2. Japanese rtMRI dataset

The Japanese rtMRI dataset [26] is currently under construction. This dataset contains 103 single Japanese morae, 676 combinations of two major morae, and 100 morae phoneme speech with rtMRI videos. The video resolution was 256 × 256 pixels, where one pixel corresponded to 1 mm. The temporal resolution of the articulation data is approximately 13.72 fps. The sound was recorded at 44.1 kHz.

In the experiment, we cropped a 150 × 150 pixel section of the articulators from the videos. In addition, the sound was down-sampled to 16 kHz. As no noise cancelation was performed on this dataset, we used spectral subtraction [27] to remove sound noise. The coefficient in spectral subtraction was set to 4.0, because the noise level was strong in the sound data. In this study, we used two male and two female speakers from the dataset.

4.2. Image pre-processing

RtMRI videos contain a large amount of noise and artifacts. In this study, we applied non-local means denoising [28] to the images. The non-local means filter is a noise reduction filter based on the convolution of support windows with adaptive weights. Non-local means denoising is suitable for processing rtMRI images because it retains the edges such as the boundary between the articulatory organs in the image more clearly as compared to other noise reduction methods.

4.3. Extraction of acoustic features

We used the World vocoder [29] and its modules Harvest, CheapTrick, and D4C to extract the fundamental frequency, mel-cepstrum, and aperiodicity from the speech, respectively. In the experiments, only the mel-cepstrum data were estimated using the DNN model. We used the other acoustic features (fundamental frequency and aperiodicity) when synthesizing the speech sound.

4.4. Network structure and comparison models

In the experiments, speaker-specific models were trained using 80% of the data. The remaining 10% of the data were used for validation and another 10% for testing. The input MRI videos were normalized to the zero mean and unit variance. No pre-training was conducted in any of the layers of the models. The weights of the CNN and transposed convolution were initialized using the initial value of He [30], and LSTM was initialized using a uniform distribution. We used the Adam optimizer, and the learning rate was set to 0.0001. We trained the networks for 100 epochs. Mel-cepstrum distortion [31], which was also used as an evaluation measure, was adopted as the cost function. The filter size of the transposed convolution and the number of filters were set to 6 and 128, respectively.

We compared our model (CNN-TC-LSTM) with Csapó's model (3) (CNN-LSTM) explained in Section 2. To evaluate the effect of transposed convolution, we constructed an additional model in which only the existence of the transposed convolution was different from CNN-LSTM (CNN-TC-LSTM (-1)). To confirm the effect of stride size in transposed convolution, we compared the stride sizes of 1, 2, 4, and 8.

Table1: *MCD on the USC-TIMIT [dB]*

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (43.2 ms)	6.45	5.71	5.67
2 (21.6 ms)		**5.68**	**5.64**
4 (10.8 ms)		5.70	5.68
8 (5.4 ms)		5.74	5.67

Table 2: *MCD on the Japanese rtMRI dataset [dB]*

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (72.5 ms)	7.06	6.16	6.18
2 (36.3 ms)		6.14	6.13
4 (18.1 ms)		**6.12**	6.12
8 (9.1 ms)		6.15	**6.08**

4.5. Evaluation metrics

We used mel-cepstrum distortion (MCD) [31] to measure the objective accuracy of the mel-cepstrum generated from the models. Additionally, we used the perceptual evaluation of speech quality (PESQ) [32] to evaluate the quality of the synthesized speech. PESQ was calculated from the synthesized speech generated using the estimated mel-cepstrum and stored acoustic features (fundamental frequency and aperiodicity). We adopted PESQ because the sound qualities were significantly different even though the MCDs were the same because the frame shift of the output mel-cepstrum differed depending on the size of the stride in the transposed convolution.

5. Results and discussion

Tables 1 and 2 present the MCD results between the generated mel-cepstrum and that of the original speech in dB on the USC-TIMIT [2] and Japanese rtMRI datasets. Based on the results, we can confirm that the use of CNN-TC-LSTM led to a reduction in the MCD by 0.81 dB and 0.98 dB on the USC-

[1] Speaker 'm1' was not used in this study due to incomplete synchronization between speech and rtMRI video.

[2] The MCD values in this experiment are higher than those in [15] because we used World vocoder instead of MGC-LSP used in [15] and we did not normalize the acoustic features.

TIMIT and Japanese rtMRI dataset, respectively. On the contrary, the results also indicate that the stride size does not have a significant impact on the performance of MCD. We can also confirm that CNN-TC-LSTM is slightly better than CNN-TC-LSTM (-1), which demonstrates that adding a convolution and a max-pooling layer and deleting two linear layers is only marginally effective for estimating mel-cepstrum from the rtMRI data.

When the stride size is set to one, the only difference between the CNN-LSTM and CNN-TC-LSTM (-1) models is the existence of the transposed convolution layer. In the transposed convolution layer, only filtering and summation operations are executed. However, the MCD scores are very different between these two models for both the USC-TIMIT and Japanese rtMRI datasets. As the filtering operation is included in the other layers of the network, we believe that the summation process affects this difference in the results. As mentioned in Section 3, the summation behaves similar to a causal convolution, which is not included in the CNN-LSTM model. Although the CNN-LSTM model handles past information of inputs in the LSTM layers, the transposed convolution layer explicitly considers the last several frames of inputs to generate the current output. This difference has a significant impact on the accuracy of the mel-cepstrum. To confirm this, we constructed a simple model based on CNN-LSTM in which the CNN layer was replaced by a 3DCNN layer, making it capable of taking several consecutive images as input. The results were 5.87 dB and 6.14 dB on the USC-TIMIT and Japanese rtMRI dataset, respectively. As can be seen, these scores are closer to those of CNN-TC-LSTM (-1) than CNN-LSTM. Therefore, we assume that taking several consecutive rtMRI images as input is important for the accurate estimation of mel-cepstrum.

Tables 3 and 4 present the results of the sound quality that was evaluated using PESQ on the USC-TIMIT and Japanese rtMRI dataset, respectively. As can be observed, the CNN-TC-LSTM (-1) and CNN-TC-LSTM models outperformed the CNN-LSTM model when the stride size was more than one. It can also be confirmed that the PESQ scores improved according to the size of the stride, whereas the effect of a changing stride size in the experiments of MCD could not be identified. This is because of the difference in the ways MCD and PESQ were calculated. Since MCD is computed on a frame-by-frame, if mel-cepstrum estimation on a frame is desirable, MCD will record a good value regardless of the quality of the output audio. On the contrary, PESQ was calculated using the entire speech with a small frame shift. Hence, the PESQ score improved as the stride size increased, because the speech quality of the intermediate frames between the image frames improved.

We also observe that the results of the CNN-LSTM model on the Japanese rtMRI dataset (0.83) were inferior to those on USC-TIMIT (1.74). We attributed this to the temporal resolution of the rtMRI videos in the datasets. While the frame shift of USC-TIMIT is 43.2 ms, that of the Japanese rtMRI dataset is 72.5 ms, which leads to a deterioration in the quality of the synthesized speeches. In contrast, the CNN-TC-LSTM model could improve the PESQ score on both the datasets by increasing the temporal resolution. Therefore, an improvement of more than 2.0, in both the datasets, was observed when using the CNN-TC-LSTM model.

In short, the CNN-TC-LSTM and CNN-TC-LSTM (-1) models outperformed the CNN-LSTM model, and CNN-TC-LSTM demonstrated slightly better results than CNN-TC-LSTM (-1) in the MCD experiment. However, the gains in terms of performance were almost the same between the two proposed models. The most significant difference was in terms of the number of parameters used in the models, as presented in Table 5. As mentioned in Section 3, the number of units in the transposed convolution and LSTM can be reduced by including an additional convolutional layer, followed by a max pooling layer. Because the number of parameters in the CNN-TC-LSTM model is fewer than that of the CNN-LSTM model, it can be concluded that the CNN-TC-LSTM will be more effective for estimating mel-cepstrum from rtMRI data, as compared with the other models.

Table 3: *PESQ on the USC-TIMIT*

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (43.2 ms)	1.74	1.92	1.95
2 (21.6 ms)		2.10	2.15
4 (10.8 ms)		2.20	2.20
8 (5.4 ms)		**2.29**	**2.29**

Table 4: *PESQ on the Japanese rtMRI dataset*

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (72.5 ms)	0.83	0.82	0.80
2 (36.3 ms)		1.47	1.42
4 (18.1 ms)		1.91	1.90
8 (9.1 ms)		**2.01**	**2.02**

Table 5: *Number of parameters in the models*

Stride size	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1	16.4 M	8.7 M	4.6 M
2		13.5 M	5.8 M
4		23.0 M	8.3 M
8		42.1 M	13.0 M

6. Conclusions

We proposed a new model for articulatory-to-acoustic conversion from rtMRI data. To address the low temporal resolution in the rtMRI data, we incorporated a super-resolution method in the temporal dimension using transposed convolution. We compared our proposed model with a model that does not contain the transposed convolution and confirmed that our model outperforms the model without transposed convolution in MCD and PESQ. We also confirmed that the PESQ scores were improved by increasing the stride of the transposed convolution. Because the proposed model is lightweight in terms of the number of parameters in the model, our model with the transposed convolution layer is expected to be highly effective for estimating mel-cepstrum from rtMRI data whose temporal resolution is not sufficient for speech synthesis. In the future, we will develop a speaker-independent and end-to-end articulatory-to-acoustic conversion system using rtMRI.

7. Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (C) 19K12024 2021 and Scientific Research (B) 20H01265 2021 by MEXT, Japan.

8. References

[1] 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.

[2] Y. Akgul, C. Kambhamettu and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.298–303, 1998.

[3] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," Medical Engineering and Physics, vol. 30, no. 4, pp. 419–425, 2008.

[4] H. J. Hermens, B. Freriks, R. Merletti, D. Stegeman, J. Blok, G. Rau, C. Disselhorst-Klug, and G. Hägg, "European recommendations for surface electromyography," Roessingh research and development 8.2, pp.13–54, 1999.

[5] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "LIP2AUDSPEC: Speech reconstruction from silent lip movements video," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., Calgary, Canada, 2018, pp. 2516–2520.

[6] 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, Sep 2014.

[7] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, 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.

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

[9] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview—application of articulatory movements using machine learning algorithms—," Acoustical Science and Technology 36.6, pp. 467–477, 2015.

[10] K. Katsurada and K. Richmond, "Speaker-Independent Mel-Cepstrum Estimation from Articulator Movements Using D-Vector Input," in Proc. INTERSPEECH, 2020, pp. 3176–3180.

[11] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," in Proc. INTERSPEECH, Hyderabad, India, 2018, pp. 2499–2503.

[12] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in Proc. INTERSPEECH, Stockholm, Sweden, 2017, pp. 3672–3676.

[13] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct Speech Reconstruction from Articulatory Sensor Data by Machine Learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2362–2374, Dec. 2017.

[14] Z. C. Liu, Z. H. Ling, and L. R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," Speech Communication, vol. 99, pp.161–172, 2018.

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

[16] T. Kitamura, S. Imai, C. Furuichi, and T. Kobayashi, "Speech analysis-synthesis system and quality of synthesized speech using mel-cepstrum," Transactions of the Institute of Electronics and Communication Engineers of Japan. A, vol. 68, pp. 957–964, Sep. 1985.

[17] G. Miyashita and M. Morise, "Influence of frame shift in speech parameters on sound quality by high–quality speech analysis/synthesis system," IEICE Technical Report. vol. 117, no. 393, SP2017–72, pp. 35–38, Jan. 2018.

[18] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," arXiv preprint arXiv:1603.07285, 2016.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 3431–3440.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015. 2, 3.

[21] C. Dong, C. C. Loy, and X. Tang, "Accelerating the Super-Resolution Convolutional Neural Network," in Proc. European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, 2016, pp 391–407.

[22] S. Pascual, A. Bonafonte, and J. Serr, "Segan: Speech enhancement generative adversarial network," in Proc. INTERSPEECH, 2017, pp. 3642–3646.

[23] I. J. Schoenberg, "Cardinal interpolation and spline functions," Journal of Approximation theory, pp. 167-206, 1969.

[24] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, pp. 10–18, 1983.

[25] 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, pp. 1791–1794, 2006.

[26] K. Maekawa, "A real-time MRI study of Japanese moraic nasal in utterance-final position," in Proc. International Congress of Phonetic Sciences (ICPhS), 2019, pp. 1987–1991.

[27] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, Vol.27, pp.113–120, 1979.

[28] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA 2005, pp. 60–65.

[29] 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, pp. 1877–1884, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034, 2015.

[31] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., Victoria, Canada, 1993, pp. 125–128.

[32] 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. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 2001, pp. 749–752.