Reconstructing Speech from Real-Time Articulatory MRI Using Neural Vocoders

Yide Yu
Institute of Informatics
University of Szeged
Szeged, Hungary
mr_yideyu@163.com

Amin Honarmandi Shandiz
Institute of Informatics
University of Szeged
Szeged, Hungary
shandiz@inf.u-szeged.hu

László Tóth
Institute of Informatics
University of Szeged
Szeged, Hungary
tothl@inf.u-szeged.hu

Abstract—**Several approaches exist for the recording of articulatory movements, such as eletromagnetic and permanent magnetic articulagraphy, ultrasound tongue imaging and surface electromyography. Although magnetic resonance imaging (MRI) is more costly than the above approaches, the recent developments in this area now allow the recording of real-time MRI videos of the articulators with an acceptable resolution. Here, we experiment with the reconstruction of the speech signal from a real-time MRI recording using deep neural networks. Instead of estimating speech directly, our networks are trained to output a spectral vector, from which we reconstruct the speech signal using the WaveGlow neural vocoder. We compare the performance of three deep neural architectures for the estimation task, combining convolutional (CNN) and recurrence-based (LSTM) neural layers. Besides the mean absolute error (MAE) of our networks, we also evaluate our models by comparing the speech signals obtained using several objective speech quality metrics like the mean cepstral distortion (MCD), Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ) and Signal-to-Distortion Ratio (SDR). The results indicate that our approach can successfully reconstruct the gross spectral shape, but more improvements are needed to reproduce the fine spectral details.**

Index Terms—**Real-Time MRI, articulatory-to-acoustic mapping, deep learning**

I. Introduction

Human speech production is a complex process that requires precisely coordinated movements from the respiratory organs, the larynx and the articulators. Theoretically, the configuration of the articulatory organs determines the speech signal that is being produced. Articulatory-to-acoustic mapping seeks to determine whether it is possible to estimate the speech signal if we know the physical positions of these organs. During the last decade, several types of devices have been proposed to record the movement of the articulators. The most important of these are ultrasound tongue imaging (UTI) [1]–[6], electromagnetic articulography (EMA) [7]–[10], permanent magnetic articulography (PMA) [11], [12], and surface electromyography (sEMG) [13]–[17]. While articulatory-to-acoustic mapping can also be considered as a theoretical problem, it has an important application, which is the creation of silent speech interfaces [18]. Silent speech interfaces (SSI) seek to generate speech from the articulatory movement without actually having access to the speech signal, with the aim of aiding the communication of speaking impaired people who can move their articulators, but cannot actually produce speech. SSIs may also be applied in situations where the speech signal cannot be heard, as in extremely noisy industrial environments and certain military applications.

Currently, the devices listed above typically record noisy and low resolution signals. Magnetic resonance imaging (MRI) may offer an alternative solution, as it can produce high resolution images. Moreover, while the other methods capture only the lingual, labial and jaw motions, MRI also covers the pharyngeal and nasal regions, which are not reachable by the other methods [19]. Of course, MRI is costly and it requires a huge machine, so currently this technology cannot provide a basis for wearable silent speech interfaces, but treating the articulatory-to-acoustic mapping problem with MRI as input is a very interesting theoretical research topic in itself. The current developments of MRI technology allow us to create real-time recordings of the speech production process at rates of about 20-50 frames per second. While this is lower than the sampling rate of other tools such as ultrasound tongue imaging, the spatial resolution of MRI is typically much better [20]. Therefore, compared to the other tools, it provides complementary information, and thus examining the applicability of MRI recordings for articulatory-to-acoustic conversion may lead to some intriguing conclusions.

Several previous studies have applied articulatory MRI for speech-related tasks. The majority of these tried to perform speech recognition using the MRI as input [21]–[23]. Interestingly, some authors tried to estimate the MRI from the speech signal, which is the inverse of the problem we study here [24]. However, relatively few attempts have been made to synthesize speech signals based on the MRI [25], [26]. Here, we extend the recent study by Tamás G. Csapó [26]. Similar to his approach, we use the MRI recordings of the USC-TIMIT data set as input [27]. However, while he used a simple conventional vocoder for the speech synthesis process, here we apply more recent neural vocoders for this task. As the name suggests, these are based on deep learning models, and they have been reported to produce higher quality speech [28]. Hence, the approach we present here is purely neural. In the next section we shall introduce the concept of our articulatory-to-acoustic mapping framework, with a special focus on the application of neural vocoders.

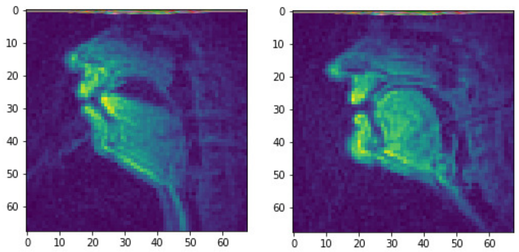

Fig. 1. *Sample MRI images (64x64 pixels) from two speakers.*

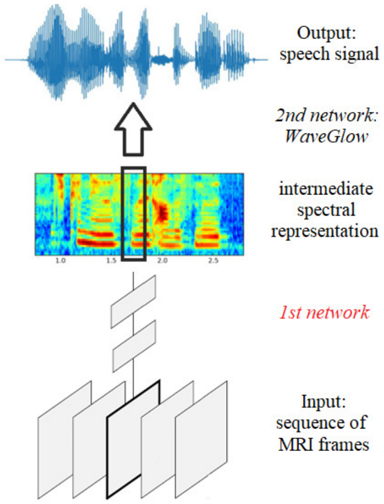

Fig. 2. *Schematic diagram of the MRI-to-speech conversion process applied here. Out goal is to find the optimal structure and parameters for the first network of the processing chain.*

II. MRI-based Articulatory-to-Acoustic Conversion

A. Real-Time Articulatory MRI

As the input MRI data we used the freely available USC-TIMIT data set [27]. This data set contains synchronized speech and real-time MRI recordings from American English speakers. The subjects lying in an MRI device were asked to read 460 sentences from the MOCHA-TIMIT database. The MRI data was captured using an 1.5 Tesla Sigma Excite HD MRI scanner. The resolution of the recorded images is 64x64 pixels, and their orientation is adjusted to the mid-saggital plane. The time resolution of the recordings was 23 frames per second. The speech signal was recorded simultaneously with the MRI video inside the MRI scanner at a sampling rate of 20 kHz. Due to the noisy operation of the scanner, the recorded speech signals were very noisy, so they had to be post-processed using noise cancellation algorithms.

Fig 1 shows two example images from a male and a female speaker from the database. As can be seen, the images have significant inter-speaker differences, so in the experiments we created speaker-dependent models, that is, separate models for each speaker. The database contains samples from five male and five female subjects, and we worked with the data of two males and two females. Since Csapó reported earlier that the data of speaker 'M1' has significant misalignment problems [26], we worked with speakers 'M2', 'M3', 'F2' and 'F3'.

B. Speech Generation using Neural Vocoders

As shown in Fig 2, our goal is to convert a sequence of MRI images into a speech signal. This task can be formalized as a sequence-to-sequence mapping problem. For such tasks, various sophisticated deep neural network (DNN) structures have been proposed recently that do not even require aligned data. However, our MRI and speech sequences were synchronized, allowing us to use much simpler architectures that estimate the mapping in a pairwise manner, that is, give one output vector for each element of the input sequence.

Thus, the input of our network is an MRI image, or a short sequence of consecutive images. However, the optimal choice for the output or target vector is not trivial. Creating DNNs that output speech signals directly is feasible [29], [30]. However, such networks generate tens of thousands of speech samples per second, and, consequently, these models are enormous and their training requires a huge amount of speech data. This is why we applied an alternative, indirect approach here, which is depicted in Fig 2. This approach was motivated by neural speech synthesis, where DNNs are applied in two steps [30]. First, there is a network that estimates a spectral representation from the text to be synthesized. Then, there is a second network that generates the speech signal from the spectral representation. Adapting this approach to our task requires modifications only in the first network, as our input is an MRI video and not a text. However, the second step is the same, so we can borrow large, pre-trained networks for the second task from text-to-speech synthesis. Doing so, we had to create and train only the first network, which had to estimate a dense spectral representation instead of the speech waveform itself. As for the second, the speech generation task, several neural vocoders are available [28], and we chose to use the WaveGlow model [30], as it worked well in our earlier study [31]. WaveGlow requires a sequence of 80-dimensional mel-scaled spectral vectors as the input, so our task was to create a network that can estimate such a spectral vector for each frame of the MRI video. As the default frame rate of WaveGlow (86 fps) is almost 4 times higher than the frame rate of our MRI video (23 fps), we estimated the missing spectral vectors by applying interpolation (using the *resize* operation of *scikit-image*) before the actual speech synthesis step.

III. DNN Architectures for MRI-to-Spectrogram Conversion

To prepare the MRI images for DNN training, the pixel intensities of each image were min-max normalized to the $[-1, 1]$ range, and the speech recordings were resampled at 22050 Hz, as this is the sampling rate required by WaveGlow. The speech signals were then converted to a mel-spectral representation, and the resulting 80-dimensional mel-spectral vectors were standardized before using them as the DNN

TABLE I
THE LAYERS OF THE 3 NETWORK ARCHITECTURES, ALONG WITH THEIR MAIN PARAMETERS. TO SAVE SPACE, SOME DETAILS SUCH AS DROPOUT LAYERS AND REGULARIZATION PARAMETERS ARE NOT SHOWN.

2D-CNN+BiLSTM	3D-CNN	3D-CNN+BiLSTM
TimeDistributed(Conv2D(30,(13,13),strides=(2,2))	Conv3D(30, (5, 13, 13), strides=(sts, 2, 2)	Conv3D(30, (5, 13, 13), strides=(sts, 2, 2)
TimeDistributed(Conv2D(60,(13,13),strides=(2,2))	Conv3D(60, (1, 13, 13), strides=(1, 2, 2)	Conv3D(60, (1, 13, 13), strides=(1, 2, 2)
TimeDistributed(MaxPooling2D((2,2)))	MaxPooling3D((1, 2, 2))	MaxPooling3D((1, 2, 2))
TimeDistributed(Conv2D(90,(13,13),strides=(1, 1))	Conv3D(90, (1, 13, 13), strides=(1, 1, 1)	Conv3D(90, (1, 13, 13), strides=(1, 1, 1)
TimeDistributed(Conv2D(85,(13,13),strides=(2,2))	Conv3D(120, (5, 3, 3), strides=(1, 2, 2)	Conv3D(120, (5, 3, 3), strides=(1, 2, 2)
TimeDistributed(MaxPooling2D((2,2))	MaxPooling3D((1, 2, 2))	MaxPooling3D((1, 2, 2))
TimeDistributed(Flatten())	Flatten()	Reshape((5, 480))
Bidirectional(LSTM(320)	Dense(1000)	Bidirectional(LSTM(370)
Dense(80, activation='linear')	Dense(80, activation='linear')	Dense(80, activation='linear')

training targets. From the 92 sentences of each subject, 4 were used for validation and 2 for testing.

Formally, our networks has to map each MRI image to a spectral vector. However, using several consecutive input frames instead of a single frame can significantly improve the results [6], [32]. Hence, the input for all our network configurations was a 3D array, treating time as the the third axis besides the two spacial axes of the images. Table I summarizes the main parameters of our experimental network configurations, and we also give a brief description of each below. Our networks were implemented in Keras. We applied the Swish activation function [33] in the hidden layers. The input window of the network contained 13 consecutive video frames. The number of trainable parameters was approximately the same for all 3 configurations. The networks were trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 100. As the loss function we applied the mean absolute error (MAE), as it was reported to give slightly better results than the mean-squared error for speech-related tasks [34].

2D-CNN+BiLSTM: When working with a sequence of images, a popular technique is to process each image using a convolutional neural network (CNN), and then combine the results along the time axis using recurrent neural structures such as the long short-term memory (LSTM) layer [35]. Thus, our first network combined 2D-CNN layers that process each image with an LSTM layer on top to fuse the information along the time axis. As shown in Table I, the 2D convolutional and max-pooling layers are applied to each image of the input sequence using the TimeDistributed() function of Keras. Then their outputs are combined along the time axis using an LSTM layer, and the actual regression task with respect to the spectral target vectors is performed by the topmost linear layer.

3D-CNN: Several authors argued recently that good video classification is also achievable using purely convolutional structures by extending the convolution to the temporal axis [6], [32], [36]. This is why a 3D convolutional model served as our second model (see Table I). Its lowest layer processes the input in 5-frame blocks with a hop size controlled by the sts parameter. This parameter allows us to analyze input blocks that are placed at bigger time intervals. In an earlier study we processed ultrasound videos which had a much larger frame rate, and the optimal value for sts was found to be 5 [32]. Here, we got the best performance with $sts = 3$, which is reasonable as the frame rate was much lower. The subsequent Conv3D layers essentially process each input block separately (their filter size along the time axis being set to 1), and the uppermost Conv3D layer performs the fusion along time. This structure was motivated by the findings of Tran et al. [36]. Finally, the regression is performed by a Dense layer.

3D-CNN+BiLSTM: A drawback of the 2D-CNN+BiLSTM model is that it cannot skip input frames, while the drawback with the 3D-CNN model is that recurrent layers may be more effective in fusing information obtained at several points along the time axis. Hence, we also experimented with a third model that combines the advantages of the two previous architectures. As Table I shows, this model retained the basic architecture of the 3D-CNN network, but we replaced the uppermost Dense layer with an LSTM layer.

IV. RESULTS AND DISCUSSION

Table II shows the MAE values obtained by the three network configurations on the validation and test sets of each speaker. The results indicate that there are large interpersonal differences compared to the average shown in the bottom row. However, the three networks architectures produced very similar results. Based on the average score on the validation

TABLE II
MEAN ABSOLUTE ERROR VALUES FOR THE VALIDATION AND TEST SETS.

	Mean absolute error (validation / test)		
speaker	2D-CNN+BiLSTM	3D-CNN	3D-CNN+BiLSTM
'F2'	0.26 / 0.26	0.28 / 0.28	0.26 / 0.26
'F3'	0.48 / 0.40	0.48 / 0.40	0.45 / 0.40
'M2'	0.37 / 0.33	0.36 / 0.33	0.36 / 0.32
'M3'	0.30 / 0.32	0.31 / 0.32	0.29 / 0.32
avg.	0.35 / 0.33	0.36 / 0.33	0.34 / 0.33

TABLE III
MCD SCORES FOR THE TEST SET.

	Mel-Cepstral Distortion (dB)		
speaker	2D-CNN+BiLSTM	3D-CNN	3D-CNN+BiLSTM
'F2'	5.87	6.02	5.84
'F3'	5.75	5.74	5.84
'M2'	5.54	5.47	5.40
'M3'	5.06	5.09	4.95
average	5.56	5.58	5.51

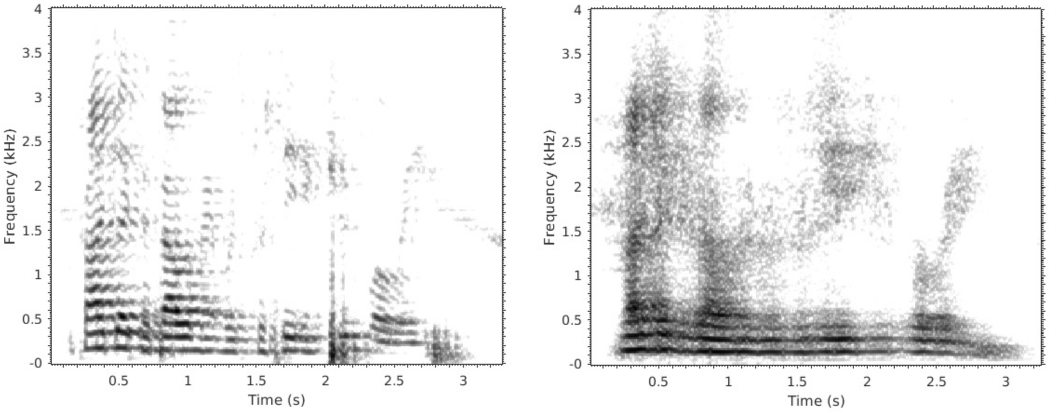

Fig. 3. *Spectrograms of a sample sentence from speaker M3. Left: original, right: reconstructed from MRI.*

set, the 3D-CNN+BiLSTM model seems to be slightly better than the 2D-CNN+BiLSTM network, and the latter is slightly better than the 3D-CNN, but the differences are negligible, and the average loss values are equivalent on the test set.

To compare not only the DNN loss values but also the synthesized speech signals, we calculated the mel-cepstral distortion (MCD) of the test files. MCD is a frequently used objective measure of speech quality in speech synthesis [37]. The MCD values reported in Table III are not in complete accord with the MAE loss values obtained for the various speakers. This reflects the fact that the simple DNN loss functions such as the MAE or MSE do not necessarily coincide with human perception [38]. However, the average MCD values have the same tendency as the MAE values in the sense that the 3D-CNN+BiLSTM model seems to be slightly better than the other two networks, but the difference is minimal.

To compare our results with those of the literature, Csapó performed a similar experiment using a conventional MGLSA vocoder, and he reported MCD scores around 4.5 for speakers ’F2’ and ’M2’ [26]. However, he attempted to estimate only the spectral envelope, and he used the residual component of the original speech signal during the synthesis. In contrast, our approach attempts to reconstruct the full spectrum from the MRI only, including the fine spectral details such as the pitch information. To illustrate the performance of our solution in this respect, we give an example of the narrow-band spectrogram (created from the output of the 3D-CNN+BiLSTM net) for speaker ’M3’ in Fig. 3. Comparing the estimated signal with the original, we see that our network is quite successful in reconstructing the rough spectral shape, but it fails with the fine spectral details. The horizontal stripes that reflect the fundamental frequency and its harmonics are preserved only in the lower spectral region. This is probably due the the mel-scale used by WaveGlow, as it represents the higher frequencies with a lower resolution. There is considerable smearing present along the time axis as well. For example. the plosive burst at 2 seconds in the original spectrogram is missing in the reconstructed signal. A further analysis is required to see whether the relatively low frame rate of the MRI is responsible for this.

We evaluated additional objective metrics to assess the quality and intelligibility of the synthesized signals (see Table IV). The Short-Time Objective Intelligibility (STOI) metric [39] returns values between -1 and 1, so the average score of 0.43 we attained is fair, but not that good. The Perceptual Evaluation of Speech Quality (PESQ) metric [40] returns values on the mean opinion scale between 1 and 5. On this scale, our average score of 1.75 lies between ”bad” and ”poor”. Lastly, the Signal-to-Distortion Ratio (SDR) metric [41] behaves much like the signal-to-noise ratio, so the negative values we got indicate that there is a large rate of distortion in our signals. However, we emphasize that the original speech recordings were also of low quality, as they were recorded in an MRI device, and they were post-processed by noise cancellation methods.

TABLE IV
OBJECTIVE SPEECH QUALITY SCORES FOR THE TEST SET.

speaker	STOI	PESQ	SDR
’F2’	0.51	1.67	-22.7
’F3’	0.27	1.86	-25.7
’M2’	0.52	1.65	-18.9
’M3’	0.41	1.81	-22.4
average	0.43	1.75	-22.4

V. CONCLUSIONS

Here, we investigated the feasibility of performing articulatory-to-acoustic mapping using neural vocoders such as WaveGlow. Compared to earlier attempts, the approach we proposed estimates the whole spectral content of the signal and not just the spectral envelope. Our experiments showed that our method is able to reconstruct the gross spectral shape and also some fine details such as the lowest pitch harmonics, but the spectrograms we got were quite blurred, and hence the resulting speech signals are at the lower end of the scale, according to several objective speech quality metrics. In the future we plan to incorporate perceptually motivated loss functions in the training process, as this would allow the direct optimization of the output speech quality [38].

Acknowledgments

This study was supported by grant NKFIH-1279-2/2020 of the Ministry for Innovation and Technology, Hungary, and by the Ministry of Innovation and the National Research, Development and Innovation Office by grant FK 124584 and the framework of the Artificial Intelligence National Laboratory Programme. The GPU card used for the computations was donated by the NVIDIA Corporation. The authors are grateful for the comments and source code of Tamás Gábor Csapó.

References

[1] T. Hueber, E. Benaroya, B. Denby, and G. Chollet, "Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface," in *Proc. Interspeech*, 2011, pp. 593–596.

[2] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, and B. Denby, "An articulatory-based singing voice synthesis using tongue and lips imaging," in *Proc. Interspeech*, 2016, pp. 1467–1471.

[3] 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*, 2017, pp. 3672–3676.

[4] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using b-mode ultrasound images," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL531–EL537, 2017.

[5] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proceedings of ICASSP*, 2017, pp. 2971–2975.

[6] P. Saha, Y. Liu, B. Gick, and S. Fels, "Ultra2speech – a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images," in *Proc. MICCAI*, 2020, pp. 473–482.

[7] J. Wang, A. Samal, J.R. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. ICASSP*, 2012, pp. 4985–4988.

[8] W. Jun, S. Ashok, and G. Jordan, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," in *Proc. SLPAT*, 2014, pp. 38–45.

[9] M. Kim, B.g Cao, T. Mau, and J. Wang, "Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2323–2336, 2017.

[10] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," in *Interspeech*, 2018, pp. 2499–2503.

[11] 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.

[12] J. A. Gonzalez, R. K. Moore, J. M. Gilbert, L.A. Cheah, S. Ell, and J. Bai, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer, Speech and Language*, vol. 39, pp. 67–87, 2016.

[13] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0," in *Proc. ICASSP*, 2011, pp. 573–576.

[14] Y. Deng, J. T. Heaton, and G. S. Meltzner, "Towards a practical silent speech recognition system," in *Proc. Interspeech*, 2014, pp. 1164–1168.

[15] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using Deep Neural Networks," in *Proc. IJCNN*, 2015, pp. 1–7.

[16] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2386–2398, 2017.

[17] M. Wand, T. Schultz, and J. Schmidhuber, "Domain-adversarial training for session independent EMG-based speech recognition," in *Proc. Interspeech*, 2018, pp. 3167–3171.

[18] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[19] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Godstein, K.S. nazak, and S. Narayanan, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," *Computer Speech and Language*, vol. 52, pp. 1–22, 2018.

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

[21] K.I. Douros, A. Katsamanis, and P. Maragos, "Multi-view audio-articulatory features for phonetic recognition on RTMRI-TIMIT database," in *Proc. ICASSP*, 2018, pp. 5514–5518.

[22] A. Katsamanis, E. Bresch, V. Ramanarayanan, and S. Narayanan, "Validating rt-MRI based articulatory representations via articulatory recognition," in *Proc. Interspeech*, 2011, pp. 2841–2844.

[23] P. Saha, P. Srungarapu, and S. Fels, "Towards automatic speech identification from vocal tract shape dynamics in real-time MRI," in *Proc. Interspeech*, 2018, pp. 1249–1253.

[24] H. Li, J. Tao, M. Yang, and B. Liu, "Estimate articulatory mri series from acoustic signal using deep architecture," in *Proc. ICASSP*, 2015, pp. 4854–4858.

[25] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Proc. Interspeech*, 2016, pp. 1492–1496.

[26] T. G. Csapó, "Speaker-dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," in *Proc. Interspeech*, 2020, pp. 2722–2726.

[27] 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, Katsamanis A., and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *Journal of the Acoustical Society of America*, vol. 136, pp. 1307–1311, 2014.

[28] P. Govalkar, J. Fisher, F. Zalkov, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. ISCA Speech Synthesis Workshop*, 2019.

[29] K. Kumar, R. K. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, vol. 32, pp. 14910–14921.

[30] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.

[31] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis," in *Proc. Interspeech 2020*, 2020, pp. 2727–2731.

[32] L. Tóth and A. H. Shandiz, "3D convolutional neural networks for ultrasound-based silent speech interfaces," in *Proc. ICAISC*. Springer, 2020, pp. 159–169.

[33] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: a Self-Gated Activation Function," *ArXiv e-prints 1710.05941*, 2017.

[34] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. ICASSP*, 2018, pp. 5414–5418.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018.

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

[38] M. Kolbæk, Z.-H. Tan, S.H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 825–838, 2020.

[39] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[40] J. Martín-Doñas, A. Gomez, Gonzalez L.J., and A. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680 – 1684, 2018.

[41] J. Le Roux, S. Wisdom, H. Erdogan, and J.R. Hershey, "SDR - half-baked or well done?," in *Proc. ICASSP*, 2019.