INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Acoustic-to-Articulatory Inversion with Deep Autoregressive Articulatory-WaveNet

Narjes Bozorg , Michael T. Johnson

Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, U.S.A
narjes.bozorg@uky.edu, mike.johnson@uky.edu

Abstract

This paper presents a novel deep autoregressive method for Acoustic-to-Articulatory Inversion called Articulatory-WaveNet. In traditional methods such as Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), mapping the frame-level interdependency of observations has not been considered. We address this problem by introducing the Articulatory-WaveNet with dilated causal convolutional layers to predict the articulatory trajectories from acoustic feature sequences. This new model has an average Root Mean Square Error (RMSE) of 1.08mm and a correlation of 0.82 on the English speaker subset of the ElectroMagnetic Articulography-Mandarin Accented English (EMA-MAE) corpus. Articulatory-WaveNet represents an improvement of 59% for RMSE and 30% for correlation over the previous GMM-HMM based inversion model. To the best of our knowledge, this paper introduces the first application of a WaveNet synthesis approach to the problem of Acoustic-to-Articulatory Inversion, and results are comparable to or better than the best currently published systems.

Index Terms: acoustic-to-articulatory inversion, speaker-dependent, WaveNet, deep autoregressive model

1. Introduction

Acoustic-to-Articulatory Inversion (AAI) is the non-linear regression problem of estimating articulatory trajectories from the acoustic signal. AAI is an ill-posed problem since it is highly non-linear and different combinations of articulatory movements can be the source of similar generated speech signals. The accurate approximation of articulatory movements and positions from acoustic signals can be useful in several domains such as audio-visual synthesis [1], Computer-Aided Language Learning (CALL) and Computer-Aided Pronunciation Training (CAPT) [2, 3].

Previously, many approaches have been proposed for speaker-dependent AAI, including codebook [4, 5], Kalman filtering [6], Gaussian Mixture Model (GMM) [7] and Hidden Markov Model (HMM) [8]. We also applied the GMM-HMM method for our previous AAI framework [9]. In this method, two synchronized streams of acoustic-articulatory data are trained separately for each individual speaker, and then at the inversion phase, the recovered articulatory trajectories are estimated from acoustic samples by deriving the optimal HMM state from acoustic model and finding the corresponding articulatory HMM state from its parallel model.

Recently these traditional models for acoustic-articulatory mapping have been upgraded by using various deep architectures such as deep Artificial Neural Networks (ANN) [10],various Deep Neural Network (DNN) architectures [11, 12, 13, 14], deep Mixture Density Networks (MDNs)[15, 16], different versions of Recurrent Neural Networks (RNNs) [17, 18, 19, 20, 21] and various combinations of Convolutional Neural Networks (CNNs) [22, 23].

AAI performance tends to vary somewhat across corpuses, especially as directly measured by RMSE. Correlation is a more consistent metric for comparing across different datasets. The best AAI approaches currently have correlations in the range of 0.8-0.85 [14, 24, 22, 16, 18, 19, 20] on the MOCHA [25] and MNGU0 [26] datasets. In comparison, traditional methods such as GMM and GMM-HMM are substantially lower, between 0.55-0.65 [7, 8, 9].

In this paper, we propose a new deep autoregressive AAI model, Articulatory-WaveNet, which uses a waveform-based speech synthesizer approach to the task articulatory inversion. Inspired by the huge success of Google's WaveNet[27] for text-to-speech applications in generating natural humanlike speech signals, we hypothesized that applying a stacked dilated convolutional layer approach would help the AAI to reach the higher accuracy compared to the previous methods.

This new system, like many other modified WaveNet versions [28, 29, 30, 31], has been conditioned on acoustic features, using Mel-spectrograms instead of linguistic features [27]. Also, the speed of synthesis has been substantially increased by applying the Fast-WaveNet [32] approach which caches previous computations instead of recomputing them from scratch to predict the new sample. The novel approach, Articulatory-WavNet, for speaker-dependent AAI and its performance for EMA-MAE dataset are discussed in sections 4 and 5 of the paper.

2. DataSet

The Articulatory Wavenet approach is evaluated using the ElectroMagnetic Articulography corpus of Mandarin Accented English (EMA-MAE) [33]. EMA-MAE includes 40 total speakers, including both native (L1) and second language (L2) speakers balanced across gender. There are an L1 group of 10 males and 10 female native English speakers (upper Midwest accent Standard American English) and an L2 group of 20 speakers. In the study presented here, we only consider the L1 group of speakers. Articulatory data for EMA-MAE were collected on a Northern Digital Inc.Wave Speech Research System with five degrees of freedom sensors (three-dimensional position plus two-dimensional sensor plane orientation) at a 400 Hz sampling rate. Data were recorded in a sound-attenuating acoustic booth, with time-synced acoustic data recorded through a cardioid pattern directional condenser microphone.

To record the articulatory data, sensors in the midsagittal plane collect information about the jaw, lower lip, upper lip, tongue body, and tongue tip. Moreover, two lateral direction sensors were also included, one at the right corner of the mouth and one in the right central midpoint of the tongue body. For each individual speaker, about 45 minutes of acoustic and ar-

http://dx.doi.org/10.21437/Interspeech.2020-1875

ticulatory data have been collected, including word, sentence, and paragraph-level speech samples. The articulatory sensors are the reference sensor (REF), jaw sensor at the lower Middle Incisor (MI), Lower Lip (LL), Upper Lip (UL), Tongue Dorsum (TD), and Tongue Apex (TA), all placed in the mid-sagittal plane, and the two lateral sensors, the Lateral Lip (LL) sensor at the left corner of the mouth to help indicate lip rounding and the Lateral Tongue (LT) sensor at the left central midpoint of the tongue body.

3. Feature Description

In this experiment, 10 articulatory features have been selected to model the kinematic space for the AAI process. There are 6 features to model the tongue movements, 3 lip related features and a feature to track the jaw movements. Table 1 represents the Vocal Tract (VT) articulatory feature set that has been applied for evaluating Articulatory-WaveNet.

Table 1: *Articulatory Feature Set for AAI*

VT Feature	Description
VT1	Tongue Dorsum Horizontal Position
VT2	Tongue Dorsum Vertical Height to HardPalate
VT3	Lateral Tongue Horizontal Position
VT4	Lateral Tongue Vertical Height to HardPalate
VT5	Tongue Tip Horizontal Position
VT6	Tongue Tip Vertical Height to HardPalate
VT7	Horizontal Lip Protrusion
VT8	Vertical Lip Separation
VT9	Lateral Lip Corner (Lip Corner Sensor)
VT10	Vertical Middle Incisor (Jaw)

For the acoustic data, Mel spectrogram features have been acquired to represent the acoustic information space.

4. Model Architecture

In this paper, we introduce Articulatory-Wavenet, which consists of the stacked dilated convolutional layer to model the conditional probability distribution and provides the accurate estimation of articulatory trajectories from acoustic signals. This fully probabilistic autoregressive architecture predicts the articulatory trajectories from the given acoustic signal by utilizing the causal conditional predictive distribution of samples.

The core of this specific deep autoregressive architecture is the causal or masked convolutional layers. For causal convolutional operations, the occurrence of each sample x is conditioned on the previous samples $(\mathrm{x}_1, .., \mathrm{x}_{\mathrm{t}-1})$. By this assumption, the dependencies on future events or samples are eliminated and all $\mathrm{P}(\mathrm{x}_\mathrm{t}|\mathrm{x} < \mathrm{t})$ can be generated in one forward pass [27, 34]. Therefore, Articulatory-WaveNet models the time series articulatory trajectories with the shifted convolutional results for the required timesteps.

The property of dilated casual convolution not only captures the long-term dependencies between samples but also significantly grows the receptive field of the network. The receptive field width can be obtained by the following equation [27]:

$$\text{Receptive Field} = \text{Number of Layers} + \text{Filter Length} - 1 \quad (1)$$

To provide a wide receptive field, we have to either add the number of neural network layers or acquire a bigger filter for spanning the space. Dilated convolutional layers use the masking convolution technique to dilate the original filter with zeros and grow the receptive field while saving the resolution of inputs and outputs at the same level. A vanilla CNN can be considered as a dilated convolutional architecture with dilation set at 1 [27, 34].

Articulatory-WaveNet has been built up from the stacked convolutional layers. Each stack contains non-linear activation unites for modeling the nonlinear acoustic-articulatory time-series signal and it follows up by residual and parametrized skip connections to speed up the convergence and enable us to design a deeper architecture.

The goal of Articulatory-WaveNet is to model the sequence of articulatory trajectories that have been conditioned on the sequence of time-series acoustic features. The predicted articulatory trajectories are synthesized from the fully trained network. The conditional probability distribution of articulatory samples x_t is represented by the following equation:

$$\mathrm{p}(\mathrm{x}|\mathrm{h}_\mathrm{t}) = \prod_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{p}(\mathrm{x}_\mathrm{t} \mid \mathrm{x}_1, ..., \mathrm{x}_{\mathrm{t}-1}, \mathrm{h}_\mathrm{t}) \quad (2)$$

where h_t represents the conditioning Mel-Spectrogram features. Figure 1 illustrates the Articulatory-WaveNet architecture and gate activation for the input x and output gate z is computed by the following equation:

$$\mathrm{z} = \tanh(\mathrm{w}_{\mathrm{f},\mathrm{k}} * \mathrm{x} + \mathrm{V}_{\mathrm{f},\mathrm{k}} * \mathrm{h}(\mathrm{t})) \odot \sigma(\mathrm{w}_{\mathrm{g},\mathrm{k}} * \mathrm{x} + \mathrm{V}_{\mathrm{g},\mathrm{k}} * \mathrm{h}(\mathrm{t})) \quad (3)$$

where $*$ represents the convolutional operator, $\odot$ is an element-wise multiplication operator, $\sigma(.)$ denotes a logistic sigmoid function, k is the layer index, f and g are filter and gate indices, respectively, and w, V are the convolutional filter weight matrices for articulatory and acoustic features respectively.

5. Experiments and Results

5.1. Metrics and Data Preparation

To measure the accuracy of the proposed system two metrics, RMSE, and correlation, have been considered in this experiment. The RMSE is calculated as:

$$\mathrm{E}_{\mathrm{rms}} = \sqrt{\frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{f}(\mathrm{x}_\mathrm{i}) - \mathrm{y}_\mathrm{i})^2} \quad (4)$$

Where y are the known values, $\mathrm{f}(\mathrm{x})$ is the estimated output trajectory, and m is the number of test files. Results are also evaluated using a Correlation Coefficient (CC) metric between actual and estimated trajectories:

$$\mathrm{CC} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{f}(\mathrm{x}_\mathrm{i}) - \overline{\mathrm{f}(\mathrm{x})})(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})}{\sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{f}(\mathrm{x}_\mathrm{i}) - \overline{\mathrm{f}(\mathrm{x})})^2 \sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2} \quad (5)$$

where y are the known values, $\mathrm{f}(\mathrm{x})$ is the estimated output, m is the number of test files and $\overline{\mathrm{f}(\mathrm{x})}$, $\bar{\mathrm{y}}$ are the utterance-level means of the estimated and actual trajectories.

For the acoustic features, the Mel-Spectrograms are extracted through a Hanning-windowed Short-Time Fourier Transform with 38.7 ms frame size and 9.7 ms frame hop. Log dynamic range compression is implemented using an 80 channel Mel filter bank spanning the range of 125 Hz to 7.6 kHz.

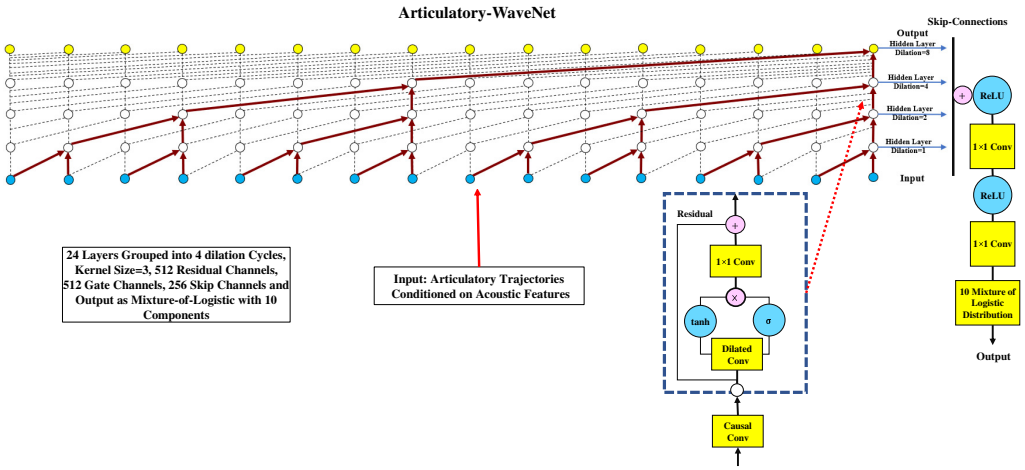

Figure 1: *Visualization of Articulatory-WaveNet, stacked causal convolutional layers with overview of the residual block and overall architecture.*

The articulatory features are the 10 static features described in section 3. To adapt the sensor articulatory trajectory information with the range of $[-1, 1]$ for activation function $\tanh$, we scaled the articulatory trajectories to this range using global dynamic range normalization. Therefore the caled Scaled Articulatory Feature (SAF) is computed by:

$$(\text{SAF})_i = 2\left(\frac{\text{Articulatory Feature} - Min_i}{Max_i - Min_i} - 1\right)$$

The dynamic range normalization is unique to each speaker and articulatory variable, with Max_i and Min_i representing the overall maximum and minimum of all articulatory trajectories for speaker i. This structure allows for easy conversion of predicted trajectories to the original feature space.

For this experiment, approximately 2500 utterances from EMA-MAE (both words and sentences) were selected across all 20 speakers (102-103 utterances per speaker) to train the Articulatory-WaveNet. For testing the performance of the network, another 350 utterances were selected separately from the training set.

5.2. Training and Synthesizing Articulatory Trajectories

While the training process can be executed in parallel, the synthesizing step is sequential for Articulatory-WaveNet. In this approach, we use the Fast-WaveNet [32] algorithm to speed up the sample generation process. Fast-WaveNet avoids redundant convolutions and caches the last computed states for the overlapping network states. Therefore, the computational efforts will be significantly reduced by utilizing the cached information from recurrent states instead of recomputing all states at each time step.

For modeling AAI, Articulatory-WaveNet deploys 24 layers with 4 dilation stacks. At each layer in each stack, the dilation increases with rate stepped geometrically by a factor of 2, which results in $1, 2, 4, ..., 512$ dilations for each stack. Causal dilated convolutions in Articulatory-WaveNet have a kernel size of 3 with 512 units in the gating layers and residual connection channels and 256 hidden units at the skip connection channel and $1 * 1$ convolution before the output layer. The output is modeled as a mixture of 10 logistic components for higher quality. To compute the logistic mixture distribution, the Articulatory-WaveNet stack output is passed through a ReLU activation followed by a linear projection to predict parameters $\theta = \{\text{Mean } \mu_i, \text{ Log Scale } S_i, \text{ Mixture Weight } \pi_i\}$ for each mixture component.

In this experiment, we considered loss as the negative log-likelihood of the ground truth sample, which is obtained by the following equation:

$$\mathrm{P}(\mathrm{x_t}|\theta, \mathrm{h_t}) = \sum_{\mathrm{i=1}}^{\mathrm{k=10}} \pi_{\mathrm{i}}[\sigma(\frac{\tilde{\mathrm{x}}_{\mathrm{ti}} + 0.5}{\mathrm{S_i}}) - \sigma(\frac{\tilde{\mathrm{x}}_{\mathrm{ti}} - 0.5}{\mathrm{S_i}})] \quad (6)$$

where $\tilde{\mathrm{x}}_{\mathrm{ti}} = \mathrm{x_t} - \mu_{\mathrm{i}}$ and $\mathrm{P}(\mathrm{x_t}|\theta, \mathrm{h_t})$ is the probability density function of the articulatory trajectory conditioned on mel-spectrogram $\mathrm{h_t}$. The Articulatory-WaveNet network was trained for 20,000 epochs using the ADAM optimizer. There are 8 mini-batches with each minibatch containing a maximum of 8000 timesteps (roughly 302ms).

5.3. AAI results

The performance of Articulatory-WaveNet for AAI has been evaluated across 20 Native English speakers for the 6 tongue articulatory features, 3 lip-related articulatory features, and 1 jaw feature.

Table 2 reports the result of RMSE and Correlation Coefficient (CC) scores across all articulatory features for two different architectures: Articulatory-WaveNet (ART-WN) and GMM-HMM. It has also demonstrates the percentage of Improvement (Imp%) for these methodes.

Our results from Articulatory-WaveNet suggests that using stacked causal convolutional layers have significantly improved the AAI performance compared to the previous GMM-HMM baseline method across all the L1 speakers from EMA-MAE corpus. The averaged correlation for all the articulatory features improved from 0.63 to 0.82 (30.16% increase), while the

Table 2: *Performance Comparison Of The Articulatory-WAVENET and GMM-HMM*

VT	Metric	Model		Imp%
		ART-WN	GMM-HMM	
VT1	RMSE	1.13mm	3.02mm	62.6%
	CC	0.84	0.61	37.7%
VT2	RMSE	1.21mm	3.00mm	59.7%
	CC	0.81	0.66	27.7%
VT3	RMSE	0.91mm	2.32mm	60.08%
	CC	0.82	0.62	32.3%
VT4	RMSE	0.98mm	2.30mm	57.4%
	CC	0.82	0.68	20.6%
VT5	RMSE	0.93mm	3.01mm	69.9%
	CC	0.81	0.61	32.8%
VT6	RMSE	1.62mm	3.24mm	50%
	CC	0.81	0.63	28.6%
VT7	RMSE	0.20mm	3.22mm	93.8%
	CC	0.83	0.61	36.1%
VT8	RMSE	1.50mm	3.02mm	50.3%
	CC	0.81	0.63	28.6%
VT9	RMSE	0.20mm	0.81mm	75.3%
	CC	0.81	0.60	35%
VT10	RMSE	2.11mm	2.09mm	-0.9%
	CC	0.79	0.66	19%
Mean	RMSE	1.08mm	2.61mm	58.6%
	CC	0.82	0.63	30.2%

RMSE decreased from 2.61mm to 1.08mm (58.62% decrease). As described below, these results are comparable to the best current techniques for AAI.

The most significant improvements for RMSE are for the horizontal Lip Protrusion, which reduces error from 3.22mm to 0.20mm (93.8%), lateral Lip Corner, reducing error from 0.81mm to 0.20mm (75.3%), vertical and horizontal Tongue Dorsum, reducing error from 3.00, 3.02mm to 1.21, 1.13mm (60% and 62.5% decrease) respectively, and vertical and horizontal Tongue Tip, reducing error from 3.24, 3.09mm to 1.62, 0.93mm (50% and 70% decrease) respectively.

The average RMSE for tracking the vocal tract height at the three tongue sensors, key variables for capturing physiological characteristics of tongue motion, is 1.27mm, down from 2.84mm for the baseline method. Speaker horizontal tongue sensor positions have an average RMSE of 0.99mm, down from 2.78mm. Vertical lip separation had an RMSE of 1.50mm, down from 3.02mm. Horizontal lip protrusion and Lateral lip distance both show slightly lower RMSEs 0.20mm, down from 3.22mm and 0.81mm respectively. The middle incisor (jaw) sensor shows slightly higher RMSE 2.11mm compared to baseline 2.09mm, which is interesting since it showed an improved correlation.

Correlation result shows consistent improvement across all features, with all 10 of the articulatory feature trajectories having correlations around 80%, ranging from 79% to 84%.

It is difficult to compare AAI results across different datasets due to both data differences and sensor placement and measurement variations. However, several specific articulatory features including lips, tongue, and incisor Articulatory-WaveNet indicate improvement compared to the best-reported approaches. The average RMSE from Latent Trajectory DNN[35] approach for the vertical tongue (tip, body, and dorsum) is around 1.80mm while for Articulatory-WaveNet the vertical tongue (tip, lateral and dorsum) RMSE reduces to 1.27mm. The best reported results for averaged correlation and RMSE with CNN+BLSTM approach in [22] for 12 articulatory features including lip, jaw, and tongue are reported around 0.84 and 1.4mm respectively. The 1.08mm average RMSE for Articulatory-WaveNet represents an over 30% lower RMSE with a similar average correlation.

Figure 2 demonstrates two examples of utterances for different speakers and different articulators to compare the estimated trajectories with the true measured EMA.

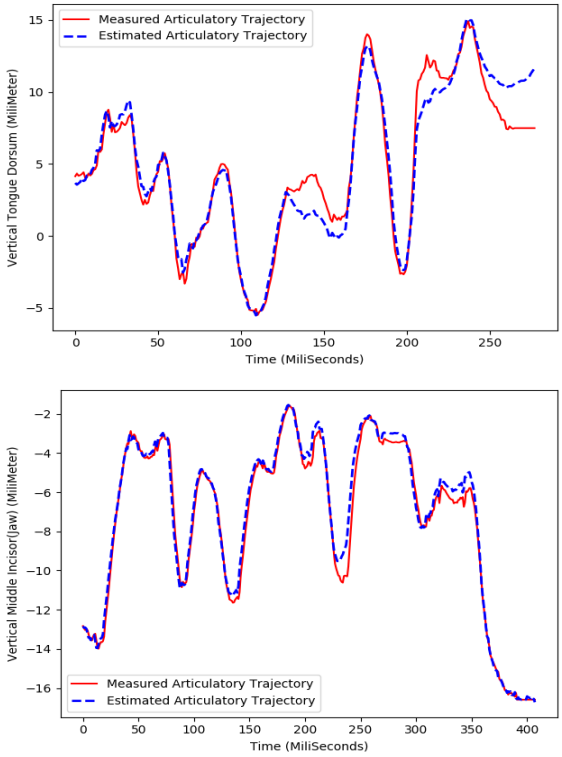

Figure 2: *Trajectories of selected articulatory features from a typical test sentence utterances. The plots show the trajectories that have been estimated by Articulatory-WaveNet alonside the target actual articulatory trajectories.*

6. Conclusions

The proposed Articulatory-WaveNet method represents a novel approach for acoustic-to-articulatory inversion. The results on the EMA-MAE corpus show significant improvement compared to the baseline GMM-HMM framework with an average correlation of 82% and RMSE of 1.08mm, demonstrating a similar correlation and substantially improved RMSE compared to the best current methods for inversion. Future work includes extending the approach to speaker-independent inversion and comparisons between subgroups of speakers across gender and dialect factors.

7. References

[1] G. Hofer and K. Richmond, "Comparison of hmm and tmdn methods for lip synchronisation," 01 2010, pp. 454–457.

[2] D. K. Jones, "Development of kinematic templates for auto-

matic pronunciation assessment using acoustic-to-articulatory inversion," 2017.

[3] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and blstm-based deep tone models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, 2019.

[4] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[5] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[6] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 237–240.

[7] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[8] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory hmm," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[9] N. Bozorg and M. T. Johnson, "Comparing performance of acoustic-to-articulatory inversion for mandarin accented english and american english speakers," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 1–5.

[10] G. Sivaraman, C. Y. Espy-Wilson, and M. Wieling, "Analysis of acoustic-to-articulatory speech inversion across different accents and languages." in *INTERSPEECH*, 2017, pp. 974–978.

[11] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion," *Proc. Interspeech 2019*, pp. 859–863, 2019.

[12] Z. Cai, X. Qin, D. Cai, M. Li, X. Liu, and H. Zhong, "The dku-jnu-ema electromagnetic articulography database on mandarin and chinese dialects with tandem feature based acoustic-to-articulatory inversion," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 235–239.

[13] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 1274–1277.

[14] A. Illa and P. K. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech & Language*, vol. 59, pp. 75–90, 2020.

[15] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[16] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *International Conference on Nonlinear Speech Processing*. Springer, 2007, pp. 263–272.

[17] A. Illa and P. K. Ghosh, "Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short term memory network," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL171–EL176, 2020.

[18] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4450–4454.

[19] X. Xie, X. Liu, T. Lee, and L. Wang, "Investigation of stacked deep neural networks and mixture density networks for acoustic-to-articulatory inversion," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Nov 2018, pp. 36–40.

[20] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated rnn," 2018.

[21] P. Maud, M. Juliette, and D. Ewan, "Independent and automatic evaluation of acoustic-to-articulatory inversion models," *arXiv preprint arXiv:1911.06573*, 2019.

[22] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5931–5935.

[23] R. Mannem, J. Mallela, A. Illa, and P. K. Ghosh, "Acoustic and articulatory feature based speech rate estimation using a convolutional dense neural network," *Proc. Interspeech 2019*, pp. 929–933, 2019.

[24] A. Illa, P. K. Ghosh *et al.*, "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5075–5079.

[25] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *In Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

[26] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[28] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for tts synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5906–5910.

[29] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[31] S. Maiti and M. I. Mandel, "Parametric resynthesis with neural vocoders," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 303–307.

[32] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast wavenet generation algorithm," *arXiv preprint arXiv:1611.09482*, 2016.

[33] A. Ji, J. J. Berry, and M. T. Johnson, "The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7719–7723.

[34] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[35] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1274–1277.