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

Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals

Abdolreza Sabzi Shahrebabaki[1], *Sabato Marco Siniscalchi*[2], *Giampiero Salvi*[1,3], *Torbjørn Svendsen*[1]

[1]Department of Electronic Systems, NTNU
[2]Department of Computer Engineering, Kore University of Enna
[3] KTH Royal Institute of Technology, Dept. of Electrical Engineering and Computer Science

{abdolreza.sabzi, giampiero.salvi, torbjorn.svendsen}@ntnu.no,
marco.siniscalchi@unikore.it

Abstract

We propose a new acoustic-to-articulatory inversion (AAI) sequence-to-sequence neural architecture, where spectral subbands are independently processed in time by 1-dimensional (1-D) convolutional filters of different sizes. The learned features maps are then combined and processed by a recurrent block with bi-directional long short-term memory (BLSTM) gates for preserving the smoothly varying nature of the articulatory trajectories. Our experimental evidence shows that, on a speaker dependent AAI task, in spite of the reduced number of parameters, our model demonstrates better root mean squared error (RMSE) and Pearson's correlation coefficient (PCC) than a both a BLSTM model and an FC-BLSTM model where the first stages are fully connected layers. In particular, the average RMSE goes from 1.401 when feeding the filterbank features directly into the BLSTM, to 1.328 with the FC-BLSTM model, and to 1.216 with the proposed method. Similarly, the average PCC increases from 0.859 to 0.877, and 0.895, respectively. On a speaker independent AAI task, we show that our convolutional features outperform the original filterbank features, and can be combined with phonetic features bringing independent information to the solution of the problem. To the best of the authors' knowledge, we report the best results on the given task and data.

Index Terms: Acoustic-to-articulatory inversion, deep learning, sequence-to-sequence neural models, 1-D convolution.

1. Introduction

The acoustic to articulatory inversion (AAI) problem is about estimating the vocal tract shape in the form of articulator positions based on the uttered speech. The actual articulatory positions can be obtained from speakers through different techniques, such as MRI [1], X-ray microbeam [2], and electromagnetic articulography (EMA) [3]. In recent years, AAI has attracted increasing attention because of its suitability in different applications, namely speech synthesis [4, 5], second language learning [6, 7], and automatic speech recognition (ASR) [8]. In a companion paper submitted to this conference [9], we show that AAI is beneficial for the continuous phone recognition task. Unfortunately, this inversion problem is highly nonlinear and non-unique [10, 11], which means that different articulator configurations can produce the same sound. In addition *coarticulation* [12], i.e. the impact of adjacent phonemes on the articulators' movement, makes the AAI problem harder.

Different machine learning techniques and various input representations have been proposed to address the AAI task. For example, search of joint acoustic and articulatory space codebooks [13], Gaussian mixture models (GMMs) [14], hidden Markov models (HMMs) [7], mixture density networks (MDNs) [15], deep neural networks (DNNs) [16], and recurrent neural networks (RNNs) [17, 18]. It is reported to obtain better accuracy than the DNN-based solution proposed in [16] exploiting an RNN-based AAI approach [19]. This result was mainly due to the better capability at capturing temporal dynamics that the RNN has through its memory elements. Different acoustic representations, such as line spectral frequencies (LSF), Mel-frequency cepstral coefficients (MFCC) and filterbank energies (FBE) have also been employed as input of the AAI system [17, 18]. Linguistic features have also been proven useful when used as stand-alone input features [20], or together with acoustic features [21, 18]. Such linguistic features are for example: phonemic (PHN) and attribute (AF) features [18]. Those features can be estimated by using a phone recognizer [22] or, a forced phone aligner [18] whenever we have access to the transcription of the uttered speech, e.g. in language learning or speech synthesis applications.

Although LSTM-based RNNs are promising for tackling the AAI task, the AAI accuracy could be further improved by exploiting ad-hoc connectionist components that can help remove redundant information in the speech signal. In fact, there exist many sources of information in the speech acoustic signal, which are not all relevant for the target task. Deep learning methods can reduce the effects of that irrelevant information leveraging upon large amounts of training material and parameters; however, lack of ad-hoc corpora providing an appropriate amount of data is a peculiar curse of the AAI problem. Therefore, the use of connectionist blocks that can better exploit the intrinsic characteristic of the speech signal could be beneficial to improve AAI results. We know that the vocal tract movements encode the linguistic message, and the speech signal reflects these movements. Non-linguistic components in the speech signal have a rate of change that lies outside the typical rate of the change of the vocal tract. 1-D convolutional connectionist components can intrinsically be more robust to the speech variability by suppressing spectral components that change more slowly or quickly than the typical range of change of the speech signal. Furthermore, convolutional components offer the advantage to reduce the amount of connectionist parameters with respect to fully connected components, which implies that a smaller amount of data can be sufficient to learn the 1-D convolutional filters. Bi-directional recurrent components with LSTM gates can instead be used to capture temporal relationships and better estimate the articulatory parameters. In this work, we thus propose 1-D convolutional layers prior to the BLSTM-based recurrent blocks to project FBE features to a new space to deal with lack of data and temporal variability. Moreover, the scarcity of relevant speaker specific data makes build-

http://dx.doi.org/10.21437/Interspeech.2020-1140

ing speaker dependent (SD) systems challenging, and the performance typically drops significantly when moving to speaker independent (SI) systems, where data from the test speaker is not used in the training stage of the neural architecture. To overcome the drop in performance caused by data scarcity in the SI configuration, we proposed to combine the feature maps from 1-D convolutions and phonetic features.

The rest of paper is organized as follows. We describe the proposed AAI approach in Section 2. The experimental setup is given in Section 3, where the “Haskins Production Rate Comparison database ”(HPRC) [23], input features and output parameters, and network parameters are presented. The experimental results are discussed in Section 3. Section 5 concludes our work.

2. Proposed method

In this work, we propose a new AAI approach, where spectral sub-bands are independently processed in time by 1-D convolutional filters of different sizes. The learned features maps are then combined and processed by an RNN with BLSTM gates for preserving the smoothly varying nature of the articulatory trajectories. We use mel filterbank energies as features in the present work to have a higher resolution for low frequency bands.

1-D convolutional layers are mostly known as the feature extraction layers from sequences and widely used in many speech applications, e.g. ASR [24, 25], speech synthesis [26], and machine translation [27]. This is the first time, to the best of the authors’ knowledge, that 1-D convolutional layers on the features are employed in the AAI task. Here we employ convolutional layers along the time axis: we consider the output of the filterbank in each of the frequency bands as a one dimensional data stream and apply the filters on it. These filters’ outputs are then linearly combined and represent new feature maps. The computations are formulated as:

$$\boldsymbol{y}_{i,j}^{\text{cnn}} = b_j + \sum_{k=1}^{L_{i-1}} \mathbf{F}_i * \boldsymbol{y}_{i-1,k}^{\text{cnn}}, \qquad (1)$$

where, $*$ shows the convolution operation of weights $\mathbf{F}_i$ in convolutional layer i with the feature maps $\boldsymbol{y}_{i-1,k}^{\text{cnn}}$ from the previous layer $i-1$. A bias b_j is added to the result of the convolution, to calculate the new feature map $\boldsymbol{y}_{i,j}^{\text{cnn}}$ for the j^{th} channel feature map. Zero padding is used to guarantee that the input sequence (acoustic space) and output sequence (articulatory space) have the same length. The 1D-CNN layers are used and concatenated along the channel axis as depicted in Fig. 1. The filter length is different in each of the CNN layers which provides more information about adjacent frames with different resolutions along the time axis. The first convolutional layer plays an important role by high-passing or low-passing different frequency bands. In our architecture, this layer has the goal of sensing significant energy changes in the speech spectrum, which may indicate a phone transition. It is built of first order FIR filters in the form $b_0 + m_0 z^{-1}$, where b_0 is a bias and m_0 a multiplicative factor. These can be either low-pass filters when b_0 and m_0 have the same sign, or high-pass, otherwise. The next convolutional layers tries to capture more temporal information and filter out undesired temporal variabilities. After those convolutional layers, two BLSTM layers are employed to capture dynamical information and estimate smoothly varying articulator trajectories. Further analysis with regards to the extracted feature maps and their representation is presented in Section 4.

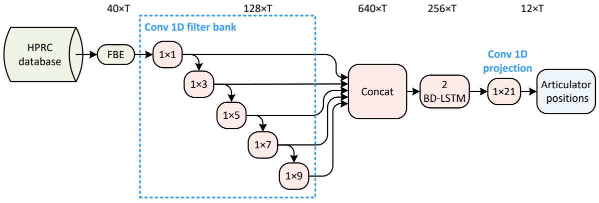

Figure 1: *Architecture of our proposed AAI method.*

3. Experimental Setup

3.1. Database

The EMA method is one of the most used techniques for recording of articulatory data which also allows for simultaneous recording of the speech. One of the available databases is the “Haskins Production Rate Comparison”(HPRC) [23], which covers material from eight native American English speakers, namely four female (F1-F4), and four male (M1-M4) speakers. There are 720 sentences available in this database with the normal and fast Speaking Rate (SR). For some of the normal speaking utterances, there are repetitions available.

Speech waveforms are sampled at rate of 44.1 KHz, and the synchronously recorded EMA data are sampled at 100 Hz. EMA data are measured from eight sensors capturing information about the tongue rear or dorsum (TR), tongue blade (TB), tongue tip (TT), upper and lower lip (UL and LL), mouth left (ML), jaw or lower incisors (JAW) and jaw left (JAWL). The articulatory movements are measured in the midsagittal plane in X, Y and Z direction, which denote movements of articulators from posterior to anterior, right to left and inferior to superior, respectively. In this work, we used the X and Z directions of TR, TB, TT, UL, LL and JAW for the speaker dependent AAI. In case of SI modeling, we employed nine tract variables (TV) [28] which are obtained by geometric transformations on EMA measurements. Those TV are Lip Aperture (LA), Lip Protrusion (LP), Jaw Angle (JA), Tongue Rear Constriction Degree (TRCD), Tongue Rear Constriction Location (TRCL). In a similar way for TB and TT we have TBCD, TBCL, TTCD and TTCL, respectively.

3.2. Input representation

In our experiments, acoustic features are extracted from a downsampled waveform at 16 KHz using an analysis window of length $25ms$ with frame shift of $10ms$, yielding a frame rate that matches the EMA recordings. Acoustic features are calculated from 40 filters, which are linearly spaced on the Mel-scale frequency axis. Energies in the overlapping frequency bands are called filterbank energy (FBE) features. Phonetic (PHN) features are extracted by the Penn phonetics lab forced aligner [29]. Each PHN feature is represented as one-hot 39 dimensional vector [18], and the attribute features (AF) are directly mapped from PHN features as in [18].

3.3. Neural parameters & settings

We compare three different neural architectures. In the first and most simple configuration, referred to as BLSTM, the unprocessed filterbank energies are directly fed at the input of the neural architecture, which is BLSTM-based RNN. Two fully connected layers are introduced between the FBEs and the BLSTM-based RNN in the second configuration, referred to as FC-BLSTM. The third configuration, 1D-CNN-BLSTM, is our proposal, and 1-D convolutional filters are employed between the FBEs and the BLSTM-based RNN. In all cases 2 BLSTM layers with 128 cells for each of the forward and backward layers are used. Sigmoid and tanh activation functions are used for the recurrent layers[18]. The output layer has 12 nodes, corresponding to the EMA dimension with linear activation function. In FC-BLSTM the first two layers are fully connected with 512 nodes with ReLU activation functions. In the 1D-CNN-BLSTM, 5 convolutional layers are used for feature extraction with the filter size of [1, 3, 5, 7, 9], respectively for each layer with ReLU non-linearity. The channel number for each of the convolutional layers are kept the same as $L_i = 128$. A batch size of 5 is used.

The experimental material is chosen from the subsets “N1” and “N2”, which have the normal speaking rate. The training data consist of 576 sentences, validation and test data each contains 72 sentences. The data splitting for the HPRC database is as in [30]. Experiments were performed in an utterance by utterance fashion, which requires that all of the utterances are zero padded to 4 sec in the feature domain for ease of training implementation. The same strategy was applied to mean normalized EMA utterances in order to obtain 4 sec duration. The Adam optimizer [31] is chosen for training the network. Keras [32] with TensorFlow backend [33] were used to train all of the neural networks. An early stopping patience of 10 iterations has been employed by checking the validation loss function to prevent over-fitting to the training data.

3.4. Performance measures

To measure the accuracy of the AAI approach, root mean squared error (RMSE) and Pearson’s correlation coefficient (PCC) are chosen. The first criterion reports the mean deviation between estimated and the ground-truth trajectories, and the latter measures the similarity of the two trajectories. The measures are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}\big(y(i) - \hat{y}(i)\big)^2}, \qquad (2)$$

$$\text{PCC} = \frac{\sum_{i=1}^{N}(y(i) - \bar{y})(\hat{y}(i) - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}\big(y(i) - \bar{y}\big)^2 \sum_{i=1}^{N}\big(\hat{y}(i) - \bar{\hat{y}}\big)^2}}, \qquad (3)$$

where $y(i)$ and $\hat{y}(i)$ are the ground-truth and estimated EMA values of the i^{th} frame, respectively; $\bar{y}$ and $\bar{\hat{y}}$ are mean values of $y(i)$ and $\hat{y}(i)$.

4. Experimental Results

In this section, we compare and contrast the three architectures described in Section 3.3. We also present additional analysis to gain a better understanding of the proposed approach.

4.1. Performance evaluation for acoustic features

For each of the three methods, we conduct 20 simulations in order to eliminate the effect of random initialization of the network parameters. Table 1 shows RMSE values, and we can observe that the proposed method outperforms both baseline approaches by almost 0.1 and 0.2 mm RMSE. A t-test shows that the reduction in RMSE with respect to both baselines is significant with p-values less than 0.05. The proposed method outperforms FC-BLSTM with lower number of parameters, as it can be observed by comparing the number of parameters reported in the table. Finally, PCC scores, related to the similarity between trajectories, are given in Table 2 and show a similar trend to that observed for RMSE.

Table 1: *RMSE for various baselines and proposed method for different speakers for AAI system.*

	Neural Architecture & No.Parameters		
Speaker	BLSTM	FC-BLSTM	1D-CNN-BLSTM
	571657	*1748233*	*1585033*
F1	1.363	1.226	**1.090**
F2	1.588	1.546	**1.380**
F3	1.296	1.231	**1.160**
F4	1.355	1.309	**1.200**
M1	1.211	1.133	**1.053**
M2	1.645	1.550	**1.435**
M3	1.523	1.479	**1.368**
M4	1.228	1.154	**1.048**
Avg.	1.401	1.328	**1.216**

Table 2: *PCC for various baselines and proposed method for different speakers for AAI system.*

	Neural Architecture & No.Parameters		
Speaker	BLSTM	FC-BLSTM	1D-CNN-BLSTM
	571657	*1748233*	*1585033*
F1	0.917	0.932	**0.945**
F2	0.852	0.858	**0.887**
F3	0.827	0.841	**0.861**
F4	0.916	0.921	**0.933**
M1	0.865	0.887	**0.902**
M2	0.861	0.880	**0.893**
M3	0.816	0.841	**0.860**
M4	0.825	0.856	**0.875**
Avg.	0.859	0.877	**0.895**

4.2. Feature extraction layers analysis

As we discussed in Section 2, 1D-CNN extract new features from FBEs. These feature maps are weighted sums of sub-band signals which have been processed by filters with different frequency responses. Fig. 2 shows an example of FBEs, and network activations through the 1D-CNN model. We can see some channel activations match phonemic segments in the first layer. Going to the next layers, the filter outputs become sparser and activations become more intense within the phoneme boundaries. For justifying our claim about channel output activations during the phonemic segments, we picked some channels output from the first layer by using correlation analysis with PHN and AF features as the reference patterns. This analysis provided a better insight for choosing the corresponding filter outputs with regards to PHN and AF features with higher correlation. As an example, we have chosen attribute fricative and phoneme /ʒ/ which are depicted in Fig. 3. The corresponding filters’ output which are chosen after doing correlation analysis are depicted in Fig. 3. We can see that these filters outputs have high energies when the chosen attribute and phoneme are active. Therefore, we can say these 1D-CNN layers are extracting the linguistic information from FBEs. This is inline with our expectation of sensing the significant energy changes at the phone transition. Furthermore, we can see for the second CNN layer compared to the first CNN layer, we have less activation outside the ground truth activation times of the chosen attribute and phoneme. By comparing the results for the SD experiment using i) the proposed architecture, and (ii) the BLSTM model that uses PHN features along with FBEs, from Fig. 4, it can be observed, the proposed architecture’s better performance could be explained by its inherent capability of 1D-CNN layers at ex-

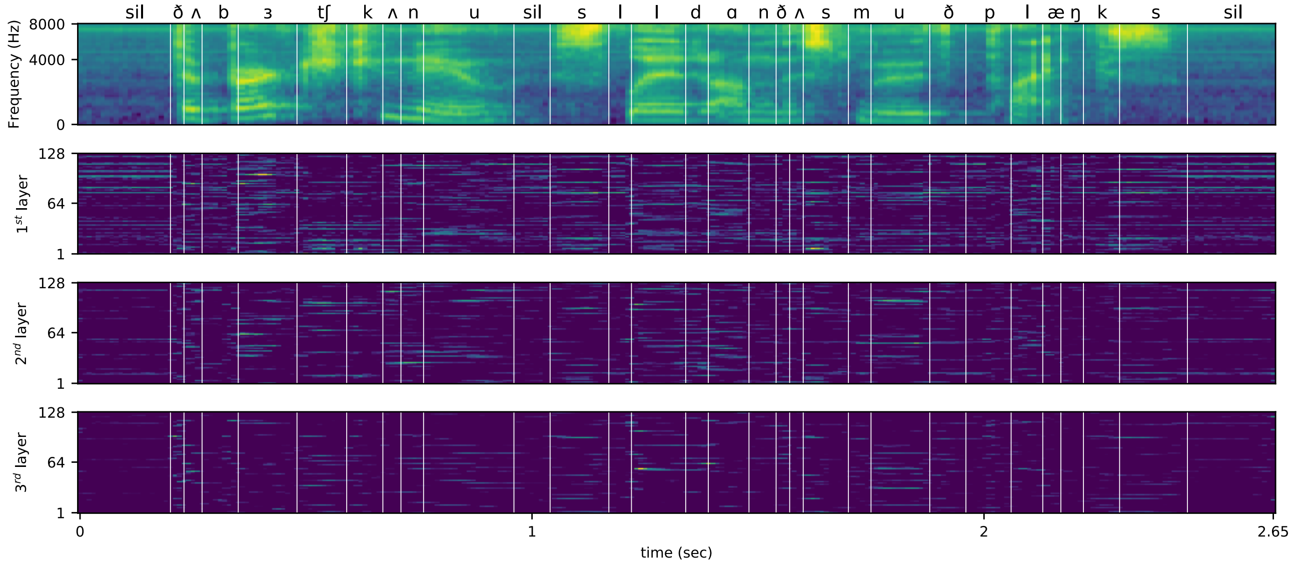

Figure 2: *FBE features for utterance "The birch canoe slid on the smooth planks." and the resulted convolutional feature maps for the* 1^{st}, 2^{nd} *and* 3^{rd} *layers.*

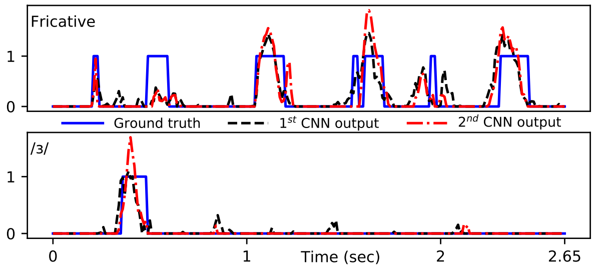

Figure 3: *AF feature for fricative and PHN features for phoneme /ɜ/ (———) and channel output from the* 1^{st} *1D-CNN layer (- - - -) and* 2^{nd} *1D-CNN layer (-·-·-).*

tracting speaker dependent information not available in one-hot encoded PHN features.

4.3. Speaker independent analysis

For evaluating the proposed method for SI training, we adopt a leave-one-out strategy, where each speaker is in turn considered as the testing speaker, and the rest of speakers are used in training. For articulatory data, we use TV trajectories as targets, and FBE and PHN features are fed at the input of the neural architectures. We used PCC as the performance measure, because of its intrinsic normalized nature that makes it less dependent on the differences between speakers' anatomy, and range of movements. We can observe from Fig. 4 that 1D-CNN improves the performance of both SD and SI configurations. Moreover, by comparing the performances of 1D-CNN with FBE, PHN and their combination, we can observe 1D-CNN has extracted more speaker dependent information while it is less speaker independent compared to FBEs. Using processed FBE features with 1D-CNN filters together with PHN features enhances the system performance in both SD and SI.

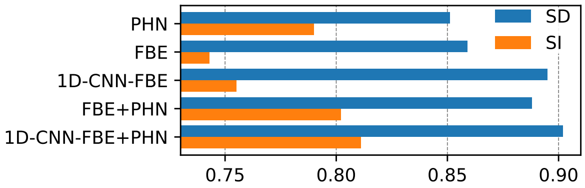

Figure 4: *PCC for the three different neural architecture tested in the present work in both speaker dependent, SD, (blue bar) and speaker independent, SI, (Orange bar) conditions.*

5. Conclusion

In this paper, we address the problem of articulatory inversion by employing 1D-CNNs as preprocessing layers to BLSTM layers. We show that this architecture improves the performance for the SD AAI task compared both to a BLSTM network alone, but also to BLSTM whose input is obtained with fully connected layers with a larger number of parameters. We also show that the representations obtained by the 1D-CNNs can be combined with phonetic features to improve performance both for SD and SI systems. The best result from only acoustic features for SD AAI of TV trajectories is PCC=0.895 and by considering phonetic features is PCC=0.901. As a comparison the SD results obtained by [30] on the same data set but with another architecture, is PCC=0.826. Our best results from only acoustic features for SI AAI of TV trajectories is PCC=0.755, by considering phonetic features with the proposed architecture, we reached averaged PCC equals to 0.810 for SI system. For the future works, we will focus on language learning and miss pronunciation detection by employing AAI systems while we have the transcription in this application.

6. Acknowledgements

This work has been supported by the Research Council of Norway through the project AULUS, and by NTNU through the project ArtiFutt. The second author is supported by the PRIN 2007 project nr. JNKCYZ_002.

7. References

[1] S. Narayanan, K. N. S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

[3] 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*, vol. 31, no. 1, pp. 26–35, 1987.

[4] K. Richmond and S. King, "Smooth talking: Articulatory join costs for unit selection," in *ICASSP*, 2016, pp. 5150–5154.

[5] R. Li and J. Yu, "An audio-visual 3d virtual articulation system for visual speech synthesis," in *2017 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*, Oct 2017, pp. 1–6.

[6] A. B. Youssef, T. Hueber, P. Badin, and G. Bailly, "Toward a multi-speaker visual articulatory feedback system," in *Interspeech*, 2011, pp. 589–592.

[7] T. Hueber, A. Ben Youssef, G. Bailly, P. Badin, and F. Elisei, "Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory HMM for pronunciation training," in *Interspeech*, 2012, pp. 783–786.

[8] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[9] A. S. Shahrebabaki, N. Olfati, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-sequence articulatory inversion through time convolution of subband frequency signals," in *submitted to Interspeech*, 2020.

[10] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[11] V. Mitra, "Articulatory information for robust speech recognition," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2010.

[12] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis using a variable-order switching shared gaussian process dynamical model," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1755–1768, Dec 2013.

[13] 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.

[14] 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.

[15] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[16] 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.

[17] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP*, 2019, pp. 5931–5935.

[18] A. S. Shahrebabaki, N. Olfati, A. S. Imran, S. M. Siniscalchi, and T. Svendsen, "A Phonetic-Level Analysis of Different Input Features for Articulatory Inversion," in *Interspeech*, 2019, pp. 3775–3779.

[19] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *ICASSP*, 2015, pp. 4450–4454.

[20] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated RNN," in *Interspeech*, 2018, pp. 3112–3116.

[21] P. Zhu, X. Lei, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *Interspeech*, 2015, pp. 2192–2196.

[22] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information," in *Interspeech 2016*, 2016, pp. 1497–1501.

[23] M. Tiede, C. Y. Espy-Wilson, D. G. V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[24] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct 2014.

[25] P. von Platen, C. Zhang, and P. Woodland, "Multi-Span Acoustic Modelling Using Raw Waveform Signals," in *Proc. Interspeech 2019*, 2019, pp. 1393–1397. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2454

[26] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech 2017*, 2017, pp. 4006–4010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[27] J. Lee, K. Cho, and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 365–378, 2017.

[28] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2014.

[29] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[30] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-Corpus Acoustic-to-Articulatory Speech Inversion," in *Interspeech 2019*, 2019, pp. 859–863.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[32] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[33] M. Abadi and et al., "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.