REPRESENTATION LEARNING USING CONVOLUTION NEURAL NETWORK FOR ACOUSTIC-TO-ARTICULATORY INVERSION

Aravind Illa, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

ABSTRACT

Recent techniques employ end-to-end systems to learn relevant features for several speech related applications, including speech recognition, and speaker verification. In this work, we focus on the task of acoustic-to-articulatory inversion (AAI) for which we propose an end-to-end system that comprises a convolution neural network (CNN) and a bidirectional long short-term memory network (BLSTM). The aim of this work is to understand the nature of the features learnt by the end-to-end model and the importance of pre-emphasis in representation learning for AAI. Further, we propose a subject adaptation scheme to overcome the limitations of the availability of parallel acoustic-articulatory data to train an end-to-end AAI system. The AAI performance is evaluated with $\sim$3.19 hours of acoustic-articulatory data collected from 8 subjects. Experiments reveal that, the frequency response of filters learnt by the CNN in the proposed system resembles those of the mel-scale, and hence, the performance of the proposed system (RMSE=1.47mm) is on par with that using mel-frequency cepstral coefficients (1.42mm) as features. Using pre-emphasis reduces RMSE by 0.13mm, and also the proposed adaptation scheme performs better than a subject-specific AAI model by an RMSE of 0.21mm despite of limited acoustic-articulatory data from a subject.

Index Terms— CNN, BLSTM, acoustic-to-articulatory inversion, electromagnetic articulograph

1. INTRODUCTION

Speech is produced as a result of different temporally overlapping gestures of speech articulators (namely lips, tongue tip, tongue body, tongue dorsum, velum, and larynx) each of which regulates constriction in different parts of the vocal tract [1]. Knowing the position information of articulators along with the speech acoustics has shown to improve the performance in various applications like speech recognition [2, 3], speech synthesis [4], accent conversion [5] etc. Electromagnetic articulograph (EMA) is one of the promising devices to capture acoustic-articulatory data comprising synchronous recordings of articulatory movements and speech but the technology is still limited to lab setup. Also, it is challenging to collect acoustic-articulatory data for a long time with sensors attached to the articulators. So, there are limitations with this direct articulatory movement recording particularly in terms of the amount of acoustic-articulatory data from a subject. Therefore, a mapping function is typically learnt to estimate the articulatory movements from speech acoustic features with the available acoustic-articulatory data. This is known as acoustic-to-articulatory inversion (AAI). Various methods have been proposed in the literature for AAI, namely codebook based [6], Gaussian Mixture Model [7], Hidden Markov Model (HMM) [8], Deep Neural Networks [9, 10]. Bidirectional Long Short Term Memory (BLSTM) network architecture among recurrent neural networks (RNN) has been shown to give state-of-the-art performance for AAI [11, 12], which also preserves the smoothly varying nature of articulatory trajectories [13].

The choice of acoustic features is crucial for AAI, and it is typically chosen in such a way that it preserves the maximal information between acoustic and articulatory features. Using maximal mutual information criterion Ghosh et al. [13], have shown that mel frequency cepstral coefficients (MFCCs) are better than the linear prediction coefficients (LPCs), cepstral representation of LPC (LPCC), and variants of LPC (line spectral frequency (LSF), reflection coefficient (RC), log area ratio (LAR)). On the other hand, recently CNNs have been shown to be successful in learning acoustic features from the raw signal, which have been used to achieve state-of-the-art performances in automatic speech recognition (ASR) [14, 15] and speaker verification tasks [16, 17]. Inspired by these advancements, we, in this work, explore learning feature representations using CNN for AAI and compare their performance with that of traditional knowledge driven features.

In this work we propose an end-to-end network for AAI, where we incorporate a CNN in the first layer to learn the features from the raw waveform followed by a BLSTM network which learns mapping function to estimate the articulatory trajectories from features learnt by CNN. End-to-end networks often require large amount of data. However, the amount of training data available for AAI is typically much less compared to those for ASR and speaker verification tasks. Hence, it is challenging to train an end-to-end network for AAI. For this purpose, we deploy different training schemes to overcome the availability of the limited acoustic-articulatory data from a subject. Experimental results show that the features learnt from CNN show competitive performance with the knowledge driven features (MFCC). The rest of the paper is organized as follows, we begin with the explanation of the data collection process followed by the proposed approach for AAI. Section 4 presents the experimental setup followed by the results and discussion.

2. DATA COLLECTION

In this work, 460 MOCHA TIMIT sentences were chosen as speech stimuli to collect acoustic-articulatory data. EMA AG501 [18] was used to record speech and articulatory movements synchronously. AG501 has a capacity of 24 channels, which can capture 3-dimensional position information of 24 sensors. To capture articulatory movements, six sensors were attached to speech articulators namely, upper lip (UL), lower lip (LL), jaw (Jaw), tongue tip (TT), tongue body (TB) and tongue dorsum (TD) following the guidelines provided in [19]. For head movement correction two sensors were placed behind the ears. We considered the articulatory movements in the midsagittal plane as shown in Fig. 1, where X and Y denote horizontal and vertical directional movements of articulators, respectively. This results in 12 articulatory features denoted by, UL_x,

UL_y, LL_x, LL_y, Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, TD_y.

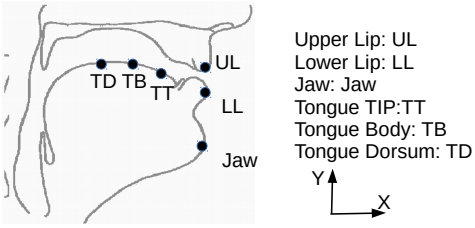

Fig. 1. Schematic diagram indicating the placement of EMA sensors [12].

Total eight subjects participated in this study, out of which 4 were male (M1, M2, M3, M4) and 4 were female (F1, F2, F3, F4). All the subjects were from an age group of 21-28 years and fluent speakers of English with no record of speech disorders in the past. For each subject, recording of all 460 sentences was done in a single session. Before starting the recordings, enough time was provided to the subject to get used to the sensors attached to the articulators. During the recording, each sentence was displayed on a computer monitor screen and a wireless slide navigator was provided to the subject to navigate through the sentences. Speech was recorded at 48kHz synchronously with articulatory movements (250Hz) using t.bone EM9600 shotgun, unidirectional electret condenser microphone [20]. We performed manual annotations for the recorded acoustic-articulatory data to remove start and end silence segments in each sentence. This results in parallel acoustic-articulatory data with a total duration of 3.19 hours and an average duration of 23.97 ($\pm$ 2.43) minutes per subject.

3. PROPOSED APPROACH

In this section, we present a brief overview of AAI followed by the proposed approach.

AAI is the task of mapping the acoustic features to the articulatory movements. This inverse mapping is known to be non-linear and non-unique in nature [3]. Also, the current articulatory position is determined not only by the corresponding phone but also by the preceding and succeeding phones. To incorporate this dependency, a context in time is provided to the acoustic features by concatenating the neighbour frames [10]. DNNs are often deployed to learn complex non-linear mapping, but the predicted articulatory movements by DNNs turn out to be jagged in nature; this, in turn, requires smoothing as a post-processing step. Recently, BLSTM networks have shown to provide a state-of-the-art performance for AAI [11, 12]. BLSTM networks also implicitly take care of providing context information to acoustic features and preserving smoothly varying nature of the predicted articulatory trajectories. Typically the acoustic features for AAI are chosen to be MFCC and shown to be the best in terms of the mutual information between acoustic and articulatory features [13]. On the other hand, CNNs are known to learn the local representations well. In this work, rather than using the knowledge driven features like MFCC, we propose a neural network architecture which extracts features directly from raw speech signal. This is done by incorporating a CNN layer as the first layer followed by the conventional BLSTM network for AAI [11, 12].

In the proposed approach, a speech signal is converted into short speech segments (also called speech frames) with a window length W_l and a window shift W_s. W_s is chosen to match the sampling rate of articulatory data, such that there is a one-to-one correspondence between the acoustic and articulatory features. To extract the features from the speech frames, we consider a single CNN layer as first layer. Let N_{cf} be the number of CNN filters with i-th filter F_i having length N_l denoted by $\mathbf{F}=\{\mathbf{F}_i\}_{i=1}^{N_{cf}}$, where $\mathbf{F}_i \in \mathbb{R}^{1\times N_l}$ with a bias vector $\mathbf{b}\in \mathbb{R}^{N_{cf}}$. At a frame index n, let $\mathbf{x}_n$ be the speech frame. We compute $\mathbf{Y}_n \in \mathbb{R}^{(W_l-N_l+1)\times N_{cf}}$, the output of the convolution filter by

$$\mathbf{Y}_n = \sigma(\log(|\mathbf{F} * \mathbf{x}_n + \mathbf{b}|)) \quad (1)$$

where, $*$ denotes the convolution operation and σ is a non-linear activation function. Before applying non-linear activation we compute the absolute value of CNN filter output followed by a logarithm [14]. Next on the output $\mathbf{Y}_n$, we perform max-pooling of size $(W_l - N_l + 1)$ which results in an $1 \times N_{cf}$ dimensional output $\mathbf{y}_n$, which could help in discarding short term phase information [15].

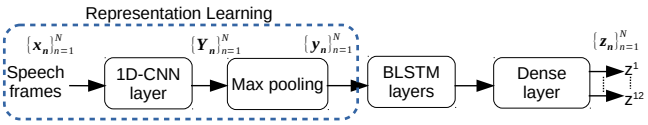

Fig. 2. Illustration of end-to-end AAI setup based on the proposed approach.

The proposed architecture is shown in Fig. 2. From all the speech frames in a given utterance, we extract features $\{\mathbf{y}_n\}_{n=1}^{N}$ using a CNN layer and max-pooling operation, where N denotes number of speech frames in an utterance. These features are fed to BLSTM hidden layers which are followed by the time-distributed linear regression (Dense) layer at the end to predict the articulatory features $\{\mathbf{z}_n\}_{n=1}^{N}$. The training is performed jointly on the CNN and BLSTM networks.

4. EXPERIMENTAL SETUP

Initially, we pre-process the recorded acoustic-articulatory data. To remove high frequency noise from the recorded articulatory data which occur during acquisition, low-pass filtering is done with a cut-off frequency of 25Hz [13]. Then articulatory data is down-sampled to 100Hz. Since the average position of EMA sensors could change from utterance to utterance, we remove mean from each dimension of articulatory feature vector at an utterance level. We down-sample the raw speech waveform to 8kHz, and perform pre-emphasis with $\alpha = 0.97$ (importance of pre-emphasis is explained in detail in Section 5.1) identical to the pre-emphasis used for MFCC computation [21]. Then using a Hamming window of length $W_l = 200$ (25msec) and with window shift $W_s = 80$ (10msec), the raw waveform is windowed to convert a given utterance into multiple speech frames.

For learning features from CNN, we use $N_l = 160$ (20msec). We have experimented with different values of $N_{cf} = 40$, 100, and 256. For the BLSTM, we choose 3 hidden layers with 150 units in each. We also performed experiments with the end-to-end architectures proposed in [14, 15] but there is no significant improvement in performance of AAI. The recorded 460 utterances of acoustic-articulatory data are divided into three sets for: train 80% (364), validation 10% (46) and test 10% (46) for each subject. An utterance by utterance training is performed by grouping all speech frames of an utterance to predict the complete articulatory trajectories for the utterance. To accelerate the speed of training we perform zero padding at the end of utterances of the acoustic-articulatory data in train and validation sets, to get a fixed length sequence of 4sec. This enables us to use a fixed batch size while training. A batch size of 5 is finally chosen, as it turns out to be the best among the choices considered namely, 5, 10, 25. We use mean squared error as an objective func-

tion to minimize and Adam as an optimizer [22]. All experiments are performed using Keras [23] with Tesorflow backend [24].

We deploy different training approaches to overcome limitation on the amount of available acoustic-articulatory data. The first approach is a standard way to train the AAI, where models are trained separately for all subjects in a subject specific manner. In the second approach, we first train a single AAI model by pooling data from all the subjects. From the single trained model, we fine-tune the weights further with the data from each subject separately. Early-stopping is imposed based on the validation data loss to avoid over-fitting.

To asses the performance of AAI models, we choose two error metrics for each articulator separately, namely root mean squared error (RMSE) and correlation coefficient (CC) [13, 12]. For i^{th} articulatory trajectory $RMSE^i$ and CC^i are given by

$$RMSE^i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(d_n^i - z_n^i)^2}, \tag{2}$$

$$CC^i = \frac{\sum_{n=1}^{N}(d_n^i - \bar{d}^i)(z_n^i - \bar{z}^i)}{\sqrt{\sum_{n=1}^{N}(d_n^i - \bar{d}^i)^2 \sum_{n=1}^{N}(z_n^i - \bar{z}^i)^2}}. \tag{3}$$

where, d_n^i and z_n^i are the original and predicted i^{th} articulatory data for n-th frame index, $\bar{d}^i$ and $\bar{z}^i$ are the corresponding mean of $\{d_n^i\}_{n=1}^N$ and $\{z_n^i\}_{n=1}^N$ across N number of frames. Note that, to indicate average CC among the articulators and/or subjects, we use the following notations: CC_{Aavg} is average CC across all the articulatory dimensions for a subject, CC_{Savg} is average CC of i^{th} articulatory dimension across all subjects and CC_{avg} is average across all articulatory dimensions and subjects. Similar notation is followed for RMSE as well.

5. RESULTS AND DISCUSSION

In this section, we present the results of the experiments which show the importance of pre-emphasis filter, followed by the performance evaluation of AAI in different training approaches. Finally we compare the proposed method with the baseline MFCC based AAI.

5.1. Analysis on Pre-emphasis

Here we perform experiments to train AAI models independently for all the subjects in subject specific manner, separately with and without pre-emphasis operation on the raw waveform. Table 1 reports the performance of AAI models with and without pre-emphasis operation for different choices of N_{cf}. We observe that, pre-emphasis of $\alpha = 0.97$ yields an improvement in $RMSE_{avg}$ of 0.13, 0.16, 0.20mm and in CC_{avg} of 0.03, 0.03, 0.04 over no pre-emphasis for $N_{cf} = 40$, 100, 256 respectively.

Table 1. Performance of AAI with and without pre-emphasis.

	N_{cf}	$RMSE_{avg}$	CC_{avg}
Without Pre-emphasis	40	1.81	0.78
	100	1.82	0.78
	256	1.86	0.77
Pre-emphasis = 0.97	40	1.68	0.81
	100	1.66	0.81
	256	1.66	0.81

To observe the difference in frequency characteristics of the learned filters with and without considering pre-emphasis, we plot the center frequencies of filters learned by the CNN for N_{cf} =256 as shown in Fig. 3. In Fig. 3, x-axis represents frequency and y-axis indicates the filter index sorted in increasing order of center frequencies. When no pre-emphasis is performed, we observe that more number of filters are centered in the frequency range 0-1000Hz. To quantify it, we calculate the number of filters with center frequencies below 1000Hz on average across all the subjects. It turns out that, without using pre-emphasis for N_{cf} = 40, 100, 256 the number of filters below 1000Hz are 24, 63.3, 171.6 as opposed to 16.6, 46.2, 129.2 respectively with using pre-emphasis. Thus we observe that pre-emphasis helps to boost the high frequency components, thereby higher formant regions and plays an important role in improving the performance of AAI. Therefore, all further experiments are carried out using pre-emphasis with α =0.97.

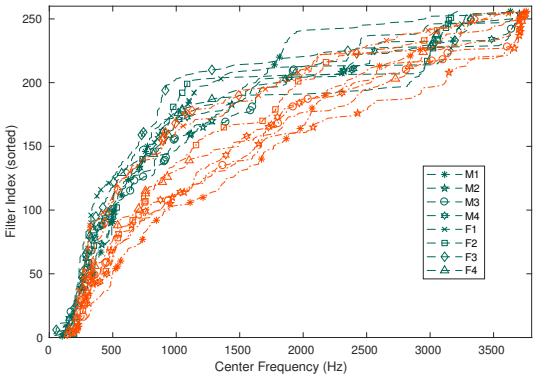

Fig. 3. Illustration of center frequencies learned using CNN with (-·-·-) and without (- - -) pre-emphasis operation for each subject.

5.2. Joint training and adaptation

Table 2. Performance of AAI in terms of $RMSE_{avg}$ (mm) with different training approaches.

Training	N_{cf} =40	N_{cf} =100	N_{cf} =256
Independent	1.68	1.66	1.66
Joint	1.56	1.63	1.60
Adaptation	1.47	1.50	1.49

At first, we perform training AAI models for each subject separately with ∼24 minutes of acoustic-articulatory data. The AAI performance using $RMSE_{avg}$ using such a setup is shown in the first row of Table 2. The acoustic-articulatory data from individual subject could be less to train a network from raw waveform. So, we pool acoustic-articulatory data from all subjects which results in ∼3.19 hours of data for jointly training an AAI model with all subjects. Second row in Table 2, summarizes the $RMSE_{avg}$ obtained by such joint training. We observe that with joint training an improvement in $RMSE_{avg}$ of 0.12, 0.03, 0.06mm is achieved for N_{cf} = 40, 100, 256 respectively. The frequency response of the filters learnt after joint training is shown in Fig. 4, where x-axis represents frequency, y-axis represents sorted filter index and color intensity variations indicate the magnitude response. Interestingly, the frequency response is band-pass in nature and center frequencies are found to be similar to those of mel-scale which is linear in lower frequency region (approximately < 1000Hz) and is non-linear from 1000Hz to 4000Hz frequency region. This could be due to the fact that the speech gestural information is maximally preserved when speech signal is processed by auditory filters such as mel-scale or bark-scale [25]. The frequency response learned for AAI is noted to be similar to the frequency response of filters learned from raw waveform in ASR literature [14, 15].

Table 3. Individual articulatory wise comparison of MFCC vs CNN features in terms of CC_{Savg}

Features	UL_x	UL_y	LL_x	LL_y	Jaw_x	Jaw_y	TT_x	TT_y	TB_x	TB_y	TD_x	TD_y	Average
MFCC	0.723 (0.15)	0.727 (0.16)	0.805 (0.12)	0.873 (0.08)	0.854 (0.08)	0.836 (0.12)	0.906 (0.06)	0.925 (0.04)	0.915 (0.05)	0.913 (0.04)	0.912 (0.05)	0.909 (0.05)	0.858 (0.12)
CNN	0.716 (0.15)	0.730 (0.16)	0.796 (0.12)	0.871 (0.07)	0.850 (0.09)	0.841 (0.11)	0.880 (0.09)	0.915 (0.05)	0.894 (0.07)	0.901 (0.05)	0.893 (0.07)	0.897 (0.05)	0.849 (0.12)

It is known that articulation style is subject specific in nature, which varies the timing and range of movements of articulators across different subjects thereby corresponding acoustic characteristics [26, 27]. Hence the AAI model learned by joint training could be a mapping function which is averaged across all the subjects. So, we further perform adaption on weights learnt during joint training by fine-tuning them to each subject's acoustic-articulatory data individually. In Table 2, last row represents the performance of AAI using adaptation. It is observed that, using joint training followed by adaptation, we achieve an improvement in $RMSE_{avg}$ of 0.21, 0.16, 0.17mm compared to the independent training for $N_{cf} = 40$, 100, 256 respectively. This could be due to the fact that the rich acoustic-articulatory mapping learnt from multiple subjects obtained by joint training would benefit in terms of better initialization of weights while adaptation [12]. We also note that an increase in N_{cf} from 40 to 256 does not seem to increase the benefit in the performance of AAI.

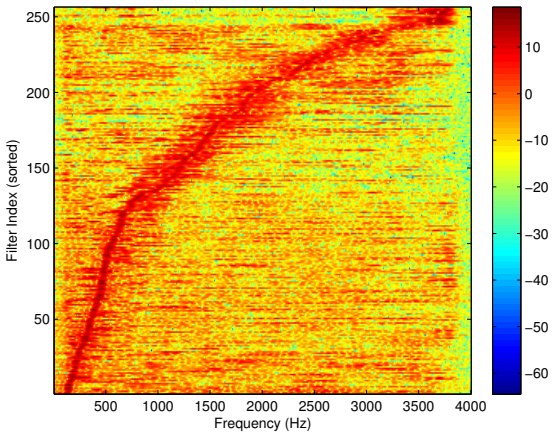

Fig. 4. Magnitude response of learned filters after joint training

5.3. Comparison with MFCC

As a baseline scheme for comparison, we compute MFCC from the acoustics. We feed MFCC features directly to BLSTM without having CNN layer and max pooling. To have a fair comparison we perform joint training and adaption for MFCC similar to the raw waveform approach with CNN. Fig. 5 shows the individual subject wise comparison of performance with knowledge based MFCC features and learnt features using CNN ($N_{cf} = 40$). The bar height for each subject in Fig. 5 indicates the mean CC_{Aavg} (top row) and $RMSE_{Aavg}$ (bottom row) across all the test utterances and the error bars indicate the corresponding standard deviation (SD). Also, we observe that the $RMSE_{avg}$ is found to be 1.47mm for CNN features which is on par with that using MFCC (1.42mm).

We also report the individual articulatory performance in terms of CC_{Savg} (SD), in Table. 3. We observe that for two of the articulators UL_y and Jaw_y, CNN performs slightly better (0.01) than MFCC. For the rest of the articulators, MFCC performs, on average, slightly better (0.01) than CNN. We perform 'ttest' on CC^i (for each subject separately), in order to verify whether the performance of AAI with CNN features and MFCC is significantly different. We observe that there is no significant ($p < 0.01$) difference in AAI performance using CNN and MFCC features, except for the articulators TT_x, TB_x and TD_x corresponding to subjects F1, F3 and F4. Fig. 6 illustrates a plot of tongue tip articulatory trajectory predicted using MFCC and CNN features with reference to original test utterance trajectory of subject F3. We observe that the trend in predicted trajectories is similar between CNN features and MFCC.

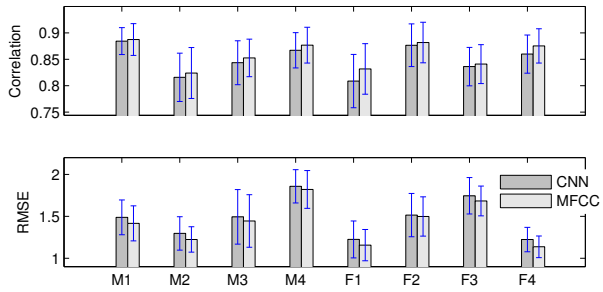

Fig. 5. Subject wise comparison of MFCC vs CNN features in terms of CC_{Aavg} (top) and $RMSE_{Aavg}$ (bottom in mm).

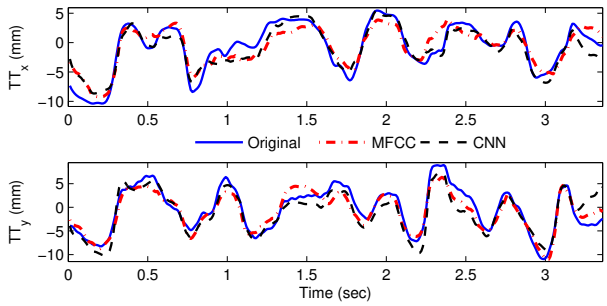

Fig. 6. Illustration of TT trajectories with MFCC and CNN features predicted using AAI with respect to the original trajectory.

6. CONCLUSION

We proposed an end-to-end network for AAI by cascading a CNN layer to the state-of-the-art BLSTM network. Experiments performed with 8 subjects revealed that the proposed CNN based approach (using a simple single 1D-CNN layer with 40 filters) performs on par with MFCC. Joint training and adaption of AAI model, which was deployed to overcome the limitation of available amount of acoustic-articulatory data, was shown to improve the performance of AAI. We showed an analysis on the benefit of pre-emphasis using a fixed coefficient $\alpha = 0.97$ for end-to-end network. In future, we will learn α by integrating it to the end-to-end AAI network. Also, it is interesting to perform an end-to-end AAI for each articulator separately to observe the learned frequency representations in an articulatory specific manner. These are the parts of our future work.

Acknowledgements: Authors thank all the subjects for their participation in the data collection. Authors thanks the Pratiksha Trust for their support.

7. REFERENCES

[1] Louis Goldstein and Carol A Fowler, "Articulatory phonology: A phonology for public language use," *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 159–207, 2003.

[2] Joe Frankel, Korin Richmond, Simon King, and Paul Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China (CD-ROM) 2000.

[3] Katrin Kirchhoff, *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefeld, 1999.

[4] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[5] Sandesh Aryal and Ricardo Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 433–446, 2015.

[6] Bishnu S Atal, Jih Jie Chang, Max V Mathews, and John 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.

[7] Tomoki Toda, Alan W Black, and Keiichi 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] Le Zhang and Steve Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[9] Korin Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping.," in *Proceedings of the ICSLP, Pittsburgh*, 2006, pp. 577–580.

[10] Zhiyong Wu, Kai Zhao, Xixin Wu, Xinyu Lan, and Helen Meng, "Acoustic to articulatory mapping with deep neural network," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9889–9907, 2015.

[11] Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, and Lianhong Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.

[12] Aravind Illa and Prasanta Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," *Proc. Interspeech*, pp. 3122–3126, 2018.

[13] Prasanta Kumar Ghosh and Shrikanth Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[14] Pegah Ghahremani, Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Acoustic modelling from the signal domain using CNNs.," in *INTERSPEECH*, 2016, pp. 3434–3438.

[15] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] H. Muckenhirn, M. Magimai.-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4884–4888.

[17] Heinrich Dinkel, Nanxin Chen, Yanmin Qian, and Kai Yu, "End-to-end spoofing detection with raw waveform CLDNNs," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2017, pp. 4860–4864.

[18] "3d electromagnetic articulograph, available online: http://www.articulograph.de/, last accessed:21/10/2018," .

[19] Ashok Kumar Pattem, Aravind Illa, Amber Afshan, and Prasanta Kumar Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[20] "Em 9600 shotgun microphone, available online: http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/, last accessed:21/10/2018," .

[21] S.J. Young and Sj Young, "The htk hidden markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory Ltd*, vol. 2, pp. 2–44, 1994.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[23] Franois Chollet, "keras," https://github.com/fchollet/keras, 2015.

[24] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[25] Prasanta Kumar Ghosh, Louis M Goldstein, and Shrikanth S Narayanan, "Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 4014–4022, 2011.

[26] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[27] Aravind Illa and Prasanta Kumar Ghosh, "Inferring speaker identity from articulatory motion during speech," *Workshop on Machine Learning in Speech and Language Processing*, 2018.