Inferring speaker identity from articulatory motion during speech

Aravind Illa, Prasanta Kumar Ghosh

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

aravindi@iisc.ac.in, prasantg@iisc.ac.in

Abstract

Speaker characteristics in speech are unique mainly because of vocal tract morphology and learned habits of articulation. In this work, we try to quantify the amount of speaker identity information encoded in articulation alone in a text independent manner. We approach this problem as a speaker identification task. For experiments, we recorded movements of the articulators namely, tongue, jaw and lips, with simultaneous speech using AG501 from 31 subjects. Experiments are performed using long short term memory (LSTM), convolution neural networks (CNN), and CNN-LSTM as classifiers. Among these, the CNN-LSTM classifier using articulatory features yields the best performance of 98.75% accuracy. Experiments are also performed to assess classifier performance using individual articulators, where the best performance of 86.94% is obtained using jaw followed by 84.89% using upper lip. Analysis on convolution filters from the first layer of CNN shows an emphasis on the frequency regions around $\sim$12-16 Hz for all the articulators. Performance results and analysis on learned filters of CNN-LSTM indicates that articulatory motion indeed capture the speaker specific information.

Index Terms: speaker identification, convolution neural networks, LSTM, electromagnetic articulograph

1. Introduction

In day-to-day life communication, identifying a speaker by his/her voice is a usual skill which humans acquire naturally. Due to learned habits of articulation and vocal tract morphology speaker characteristics in speech are unique. In many speech technologies, automatic speaker recognition/identification has emerged as an important application in business and forensics for verifying the identity of a speaker [1]. Briefly, speaker recognition consists of two different tasks, namely, speaker identification and verification. In speaker identification, the task is to identify an unknown speaker from a set of known speakers, whereas in verification an unknown speaker claims an identity and the task is to verify if the claim is true/false [1]. By extracting short-term spectrum features from speech acoustics various approaches are proposed in literature including universal background model with a Gaussian mixture model [2, 3] and i-vector (identity vector) [4, 5]. With advancements in deep neural networks (DNN) [6], various end-to-end approaches are also proposed [7, 8].

Several studies have been carried out in the past for understanding inter-speaker variability in vocal tract morphology. It has been shown that the differences in vocal tract length relates to formant frequency in voiced sounds [9, 10]. Also, the variability of morphology of the hard palate and the posterior pharyngeal wall [11], influences articulation [12, 13]. To capture the morphological variations, a practical feature-level fusion approach has been proposed for combining acoustic and articulatory information in speaker verification task [14, 15]. Specifically, speaker verification performance is enhanced significantly by augmenting mel-frequency cepstral coefficients (MFCCs) with articulatory features obtained from subject-independent acoustic-to-articulatory inversion technique [14, 15]. Recently, speaker verification in a text-dependent manner is performed by articulatory movements using the dynamic time warping (DTW) algorithm [16].

In this work, we focus on speaker identification from the articulation alone in a text independent manner. We hypothesize that due to the learned habits, the articulation of a speaker would induce cues of speaker identity in the articulatory motion. In this work we try to quantify the amount of speaker specific information that articulatory motion conveys. To carry out the experiments, we recorded articulatory data from 31 subjects using Electro-Magnetic Articulography (EMA). To extract the features and identifying speakers from articulatory motion we use convolution neural networks (CNN) [17, 18] and recurrent neural networks (RNN) [19, 20]. Experimental results show that the articulatory motion, when used as a feature for speaker identification task, yields an accuracy of 98.75%, which is on par with the performances obtained using only acoustic (98.56%) and only morphological features (99.72%). Analysis on the filters learned on first layer of CNN shows an emphasis of frequency region around $\sim$7-10 Hz for the jaw and lip articulators. Experiments are also performed by fusing articulatory motion features with MFCC with variable training data conditions. It is observed that, especially with less amount of training data ($\sim$5.5% of total training data) fusion gives an improvement in the accuracy by $\sim$12% with respect to the individual performance by the acoustic and articulatory features. Experiments are also carried out using individual articulators, to examine which articulator contributes more in revealing the speaker identity. It turns that the best performance of 86.94% is obtained using the motion of jaw followed by 84.89% using upper lip motion.

2. Dataset

To capture articulatory motion for experiments, we record the articulatory movement data using Electromagnetic articulograph (EMA) AG501 [21]. 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus [22] are chosen as the stimuli for data collection. We collect data from 31 speakers comprises of 18 males and 13 females in an age group of 20-28 years. All 460 sentences were recorded in a single session for every subject. Prior to the data collection, a consent form is signed by all speakers, as recommended by the institute ethics committee. All speakers are reported to have no speech disorders in the past. For each sentence, we record simultaneous audio and articulatory movement data. We use eight sensors among which six are placed on different articulators, namely Upper Lip (UL), Lower Lip (LP), Jaw, Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD). The remaining two sensors are placed behind the two ears for head

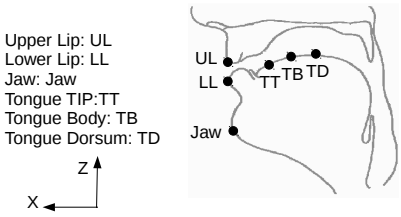

Figure 1: *Schematic diagram indicating the placement of EMA sensors [23].*

movement correction. For attaching the sensors we follow the suggestions provided in [24]. A schematic diagram of the sensor placement is shown in Fig. 1. Each of these eight sensors captures the movements of the corresponding articulators in 3D space. In this study we consider the movements only in the mid-sagittal plane, indicated by X and Z directions in Fig. 1. Once the sensors are attached, sufficient time is given to the speaker to get used to speaking naturally with the sensors attached to different articulators.

A microphone is placed near the subject to record the audio data at 48 kHz synchronously with the articulatory data at 250 Hz. During recording, the sentences are projected onto a screen placed in front of the subject. For all the recordings, manual annotations are done to remove silence in the beginning and the end of each sentence. This results in data with a total duration of 11.72 hours from all subjects with the average duration per subject being 22.66 ($\pm$ 2.6) minutes.

3. Proposed approach

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

Recently, 2-D convolution neural networks gained a great attention and success in the filed of computer vision [17, 18] to extract simple 2D features. In this work, we adopt the convolution network to perform temporal convolution for extracting features in sequence classification task. Let us consider, N_l number of convolution filters in the l^{th} layer with a filter length f_l, and T_l be the number of corresponding time steps. Let us denote the collection of filters by $\mathbf{F}^l$=$\{\mathbf{F}_i^l\}_{i=1}^{N_l}$, where $\mathbf{F}_i^l \in \mathbb{R}^{N_{l-1} \times f_l}$ with a bias vector $\mathbf{b}^l \in \mathbb{R}^{N_l}$. Given an input signal $\mathbf{X}^{l-1} \in \mathbb{R}^{N_{l-1} \times T_{l-1}}$ to the l^{th} layer, we compute output by

$$\mathbf{X}^l = \sigma(\mathbf{F}^l * \mathbf{X}^{l-1} + \mathbf{b}^l) \tag{1}$$

where, σ is a non-linear activation function and $*$ denotes the convolution operation.

RNNs with long short-term memory (LSTM) has been shown to be an effective model for learning problems dealing with sequential data [19]. At time t, let $\mathbf{x}_t$ be the M-dimensional input and N be the number of memory cells in an LSTM layer with output $\mathbf{y}_t \in \mathbb{R}^N$. Then for each LSTM layer we have, input weights $\mathbf{W}_* \in \mathbb{R}^{N \times M}$, recurrent weights $\mathbf{R}_* \in \mathbb{R}^{N \times N}$ and bias weights $\mathbf{b}_* \in \mathbb{R}^N$. The forward pass for LSTM layer can be written as follows [19, 20]:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{y}_{t-1} + \mathbf{b}_i) && \text{input gate} \\
\mathbf{f}_t &= \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{y}_{t-1} + \mathbf{b}_f) && \text{forget gate} \\
\mathbf{c}_t &= \mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{i}_t \odot tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{R}_c \mathbf{y}_{t-1} + \mathbf{b}_c) && \text{cell memory} \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{y}_{t-1} + \mathbf{b}_o) && \text{output gate} \\
\mathbf{y}_t &= tanh(\mathbf{c}_t) \odot \mathbf{o}_t && \text{block output}
\end{aligned} \tag{2}$$

where, σ is a point-wise non-linear activation function and $\odot$ denotes element-wise multiplication of two vectors.

CNNs are known to recognize local patterns in a sequence. On the other hand, RNNs are known to model the temporal dynamics by processing the sequence of input samples and maintaining a state information relative to history. In the proposed approach, we use an architecture by combing CNN and RNN layers. Fig. 2, illustrates the current approach for speaker identification. First, we extract the local temporal structure in the sequential data by performing temporal convolutions using CNN followed by a max-pooling layer. Upon the high level features computed by CNN, an LSTM layer is used to capture the temporal dynamics of sequence. At the output layer we use softmax activation layer to predict the output conditional probability vector (with dimension equal to number of speakers).

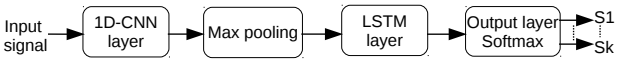

Figure 2: *Illustration of speaker identification setup based on proposed approach.*

The proposed approach has the following hyper-parameters: a) sequential data segmentation: window length (w_l) and window shift (w_s) for processing the entire sequential input into segments. b) CNN parameters: filter length (f_l) and the number of filters in each CNN layer (N_l), max-pooling size (mp), and activation function. c) LSTM parameters: number of memory blocks for LSTM layer.

4. Experimental Setup

The recorded audio data is first down-sampled to 16 kHz from 48 kHz. Then the silence segments before and after every sentence are removed manually in both audio and articulatory movement data. From the audio data we extract 13-dim MFCC [25] feature vector with frame length and shift being 20ms and 10ms respectively. Since, MFCC has been shown as a potential acoustic feature for the task of speaker identification and classification tasks [14], a cepstral mean subtracted MFCC is then used as the acoustic feature vector.

The recorded articulatory movement data has high frequency noise, while most of the energy of the articulatory movement lies below 25Hz [26]. Hence, the recorded articulatory data is low-pass filtered with a cutoff at 25Hz. Then the articulatory data with a sampling rate of 250Hz is down-sampled to 100Hz. Since average position for each sensor could change across sentences [27] we remove the mean sensor position for each articulatory feature in every sentence. This also removes the relative morphological information in the raw articulatory position data and helps in using only the dynamics information in the articulatory motion for the task at hand. We use the obtained articulatory features (AFs) to characterize the articulation of the speaker. Thus AFs results in a 12-dimensional feature vector, which represents the articulatory motion in horizontal and vertical directions, namely, UL_x, UL_z, LL_x, LL_z, Jaw_x, Jaw_z, TT_x, TT_z, TB_x, TB_z, TD_x, TD_z.

The recorded 460 sentences from all the subjects are divided into three sets, 368 sentences for training data, 46 sentences each for test and validation data. The input to the CNN-LSTM is a 12-dimensional sequence of articulatory features. Before passing the input features to the CNN-LSTM, we first chunk the sequential data into a fixed duration 0.85 sec with a shift of 0.05 sec. For training the network we use cross-entropy

as the cost function with Adam optimizer [28]. The implementation was done using Keras library [29].

Evaluation metric: To evaluate the performance of the speaker identification scheme we use identification accuracy as the metric [3] as follows:

$$\text{accuracy}\% = \frac{\#\text{correctly identified segments}}{\#\text{ of total segments}} \times 100. \quad (3)$$

5. Results and Discussion

Baseline comparison: In the first step, we compare the performance of CNN-LSTM with the CNN-CNN and LSTM-LSTM architectures. For CNN-CNN scheme, we choose an architecture with $N_f = 50$ filters in both first and second layer with a window length $w_l = 85$, with a shift $w_s = 5$. For CNN-LSTM and LSTM architectures, we choose an LSTM layer with 150 units. We vary the filter length f_l from 5 to 25 samples with a step size of 5 samples. The best choice of filter length turns out to be f_l =10 on the validation set. To understand how much speaker information AFs capture with respect to acoustics, we also compare the models with the MFCC. Table 1, shows the identification accuracy using AFs and MFCC. Using the entire training data in all the classifier architectures, AFs perform on par with the MFCC, suggesting that AFs capture speaker identity information. Among the three classifiers, CNN-LSTM performs 1.3% better than the second best scheme (CNN). So for further analysis and experiments, we choose the CNN-LSTM architecture.

Table 1: *Comparison of performance of architectures on CNN and LSTM combinations*

	LSTM	CNN	CNN-LSTM
AFs	96.63	97.40	98.75
MFCC	97.54	97.13	98.56

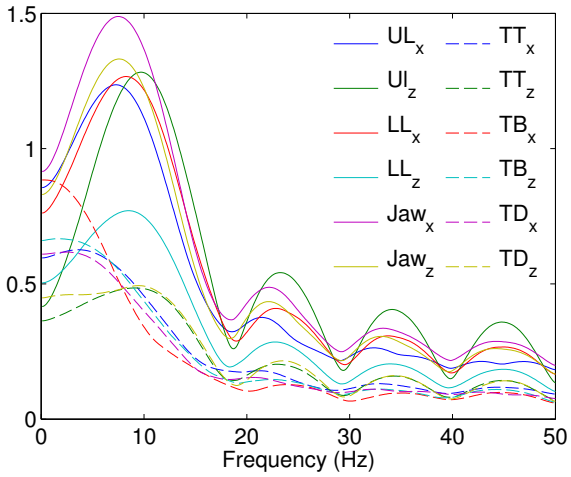

Figure 3: *Cumulative response of CNN filter in the first layer for input AFs.*

Cumulative frequency response of CNN filters: We analyze the filters of the 1st layer of the CNN-LSTM architecture. The average frequency response of all the filters is computed as [18, 30]

$$H_{cum}(z) = \sum_{n_f=1}^{N_f} H_{n_f}(z) \quad (4)$$

where, $H_{n_f}(z)$ is the frequency response of filter $F_{n_f}(n)$ computed using 512-point Discrete Fourier Transform (DFT). Fig. 3 shows the cumulative frequency response for all the articulators. We observe that the filters for jaw and lip articulators are emphasized in the frequency region around ∼7-10 Hz, while those for tongue articulators exhibit a low-pass filter characteristics.

Table 2: *Comparison of performance using individual articulators.*

AFs	UL	LL	Jaw	TT	TB	TD
Accuracy (%)	84.9	81.7	86.95	72.93	64.29	63.17

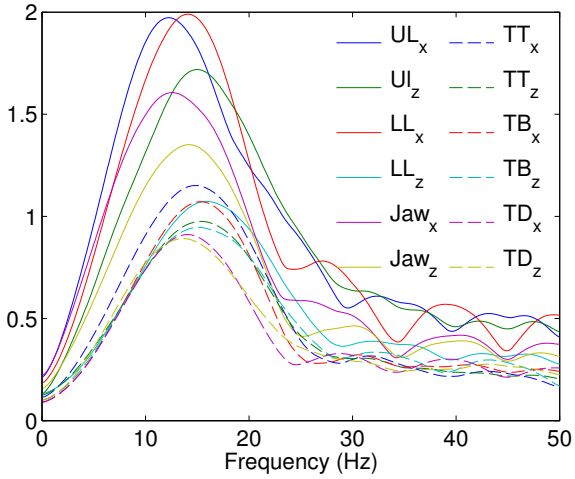

Figure 4: *cumulative response of CNN filters in the first layer with individual AFs.*

Individual Articulatory performance: To know which articulator captures most of the speaker identity characteristics, we conduct experiments with each articulator individually. The motion of horizontal (X) and vertical (Z) motion of individual articulators are considered as features. This results in a 2-dimensional feature for each of the six articulators namely, LL, UL, Jaw, TT, TB and TD. For, each of them a CNN-LSTM classifier is trained separately. Table 2 summarizes the performance of individual articulators. Among all, Jaw and UL articulators results in a relatively better accuracy revealing that jaw and UL encode speaker specific information better than other articulators. Experiments are also performed with all lip articulators (UL+LL gives 95.53%) and all tongue articulators (TT+TB+TD gives 95.42%). The cumulative frequency responses of filters learned by the CNN filters in first layer are shown in Fig. 4. Here, we clearly observe that all the articulators exhibit a band-pass nature primarily in a range of ∼12-16 Hz. Comparing Fig. 3 and Fig. 4, we observe that the frequency characteristics of the filters from the CNN-LSTM trained with individual articulators are different from those from CNN-LSTM trained with all articulators. This particularly true for tongue articulators. This could be due to the fact that tongue articulators have less discriminative performance compared to lips and jaw (from Table 2). So, when all the articulators are considered together for training, tongue articulators may not get emphasized by CNN filters as observed in Fig. 3.

Variable Data Size: To understand the variation in the speaker identification accuracy due to the availability of train-

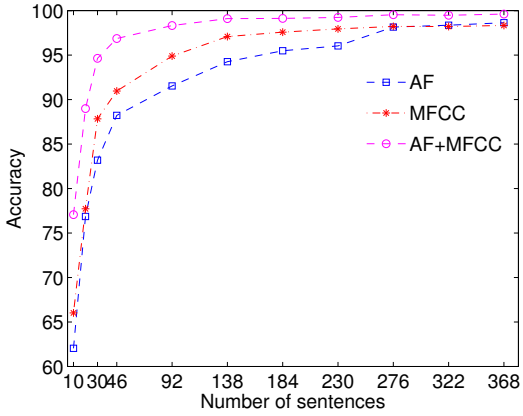

Figure 5: *Performance of CNN-LSTM by varying the training data size.*

ing data, we experiment by varying the amount of training data size from each subject (i.e, number of training sentences are varied as 10, 20, 30, 46, 92, 138, 184, 230, 276, 322, 368). In order to examine if AFs contain complementary speaker information, we also consider fusing AFs with MFCC for evaluating the performance. Fig. 5 shows the accuracy of classifier with AFs, MFCC and fusion of MFCC with AFs (AFs+MFCC) using varying amount of training data. X-axis represents the number of sentences chosen for training- varied from 10 to 368. From Fig. 5 we can observe that as the amount of training data increases the performance of AFs, MFCC and AFs+MFCC increases monotonically. We find that although with less training data MFCC perform better than AFs, with increasing training data, the performance of AFs is comparable to that of MFCC. Fusion of AFs with MFCC helps in improving the preformance to 99.6% accuracy using entire training data suggesting that AFs (98.75%) adds complementary information to MFCC (98.56%). This is particularly observed in low training data, e.g, for 20 training examples fusion shows an improvement in accuracy by $\sim$12% (AFs (76.8%), MFCC (77.72%), AFs+MFCC (89.00%)). It is also interesting to observe that by the fusion of AFs to MFCC with only 25% of total training data, we achieve a performance of 98.32% which is equivalent to the performance using MFCC and AFs individually with entire the training data.

Performance evaluation on Morphological features: Since morphology plays a crucial role in encoding speaker identity, we also experiment with the features which can capture the morphological variations across the subjects. Results in Table 1, 2 and Fig. 5 were presented by ignoring the vocal tract morphological variations by removing the mean from the individual articulatory trajectory within each sentence. In this experiment, to capture the morphological variations across subjects, we compute mean sensor locations of articulators for each sentence, which is considered as the morphological features (MFs). Thus, for each utterance we get one MF (12 dimensional), resulting in a total of 460 MFs per subject. In order to avoid difference in the relative positions of the sensors in recording different subjects, sensor placement was done by a single person following the guidelines in [24]. We, thus, assume that the MFs do not get influenced by the differences in the sensor position across subjects, rather it only captures their morphological variations. To train a classifier with MFs, we choose a DNN. Similar to the experiments with the AFs, we split the data into 80%, 10% and 10% for train, test and validation, respectively. For DNN, we choose first layer with linear activation, followed by a two hidden layers (256 units) with tanh activation function and a last layer with softmax activation. The DNN is trained using cross-entropy as the cost function with an early stopping criteria. The results using MFs with the DNN classifier yield an accuracy of 99.72%. Such a high accuracy using morphological features is expected because the relative distances among articulators will be unique for each speaker and DNN could use this cue well for speaker identification.

Speaker identification results using AFs are found to be comparable and complementary to the those using MFCC. This demonstrate that articulatory motion indeed encodes speaker identity information. The complementary information from AFs is found to be useful particularly under low training data condition. We hypothesize that the additional information from the articulatory motion, could potentially be contributed by the motion of non-critical articulators [31]. For achieving the articulatory goals (linguistically), critical articulators are more carefully controlled and exhibit less variability compared to the non-critical articulators. For example, in producing /p/, it is essential that the lower lip (critical) comes in contact with the upper lip, while tongue articulators (non-critical) can be variable. The information in the position of non-critical articulators [32], while available in a direct measurement, may not be estimated well by the inversion schemes as were done in [14]. However, a direct measurement of articulatory motion may not be feasible in real-life applications. A more detailed investigation is needed to understand the subject specific behavior of articulatory planning of non-critical articulators during speech production and develop techniques to estimate them from acoustics. Note that for all the experiments, we consider articulatory motion along mid-sagittal plane (X and Z direction), whereas the importance of movements of articulators along Y direction needs further investigation.

6. Conclusions

In this work, an investigation is carried to infer the speaker identity using articulatory motion from different articulators namely, tongue, lips and jaw, in a text-independent manner. Experiments are carried out with 31 subjects, and articulatory motion is captured using EMA AG501. Experiments are performed with CNN-LSTM and DNN to build classifiers for speaker identification for AFs and MFs respectively. Experimental results show that, articulatory motion encodes the speaker identity and gives an accuracy of (98.75%) comparable to that using acoustic (98.56%) and morphological features (99.72%). Also, a fusion of AFs with MFCC results in an accuracy of 99.6%. Experiments are also carried out by considering individual articulators, where the best performance is obtained using jaw (86.95%) followed by upper lip (84.89%) among all articulators. Analysis of cumulative frequency response of filters for individual articulators learned from CNN shows that the emphasis is given in the frequency range of $\sim$12-16 Hz. Further investigation is needed to understand the influence of non-critical articulators in characterizing speaker identity. These are parts of our future work.

7. Acknowledgements

Authors thank all the subjects participated in the data collection, and Deep P, Nisha G, Kaustubha NK for helping in recordings. Authors thank Pratiksha Trust for their support.

8. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[6] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[7] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

[8] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform cldnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4860–4864.

[9] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[10] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[11] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 521–530, 2013.

[12] J. Brunner, S. Fuchs, and P. Perrier, "On the relationship between palate shape and articulatory behavior," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3936–3949, 2009.

[13] J. Brunner, Fuchs, and P. Perrier, *The influence of the palate shape on articulatory token-to-token variability*. Universitätsbibliothek Johann Christian Senckenberg, 2013.

[14] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer speech & language*, vol. 36, pp. 196–211, 2016.

[15] M. Li, J. Kim, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on fusion of acoustic and articulatory information." in *INTERSPEECH*, 2013, pp. 1614–1618.

[16] Y. Zhang, Y. Long, X. Shen, H. Wei, M. Yang, H. Ye, and H. Mao, "Articulatory movement features for short-duration text-dependent speaker verification," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 753–759, 2017.

[17] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of IEEE International Symposium*. IEEE, 2010, pp. 253–256.

[18] S. Mallat, "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016.

[19] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

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

[21] "3d electromagnetic articulograph, available online: http://www.articulograph.de/, last accessed:1/9/2016." [Online]. Available: http://www.articulograph.de/

[22] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database, 1999. [Online]. Available: http://sls.qmuc.ac.uk

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

[24] A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[25] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering, 1993.

[26] P. K. Ghosh and S. 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.

[27] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, University of Edinburgh, 2002.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.

[30] D. Palaz, R. Collobert *et al.*, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings of INTERSPEECH*, 2015, pp. 11–15.

[31] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "A kinematic study of critical and non-critical articulators in emotional speech production," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1411–1429, 2015.

[32] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 688–700, 1992.