Low resource acoustic-to-articulatory inversion using bi-directional long short term memory

Aravind Illa, Prasanta Kumar Ghosh

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

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

Abstract

Estimating articulatory movements from speech acoustic features is known as acoustic-to-articulatory inversion (AAI). Large amount of parallel data from speech and articulatory motion is required for training an AAI model in a subject dependent manner, referred to as subject dependent AAI (SD-AAI). Electromagnetic articulograph (EMA) is a promising technology to record such parallel data, but it is expensive, time consuming and tiring for a subject. In order to reduce the demand for parallel acoustic-articulatory data in the AAI task for a subject, we, in this work, propose a subject-adaptative AAI method (SA-AAI) from an existing AAI model which is trained using large amount of parallel data from a fixed set of subjects. Experiments are performed with 30 subjects' acoustic-articulatory data and AAI is trained using BLSTM network to examine the amount of data needed from a new target subject for the SA-AAI to achieve an AAI performance equivalent to that of SD-AAI. Experimental results reveal that the proposed SA-AAI performs similar to that of the SD-AAI with $\sim$62.5% less training data. Among different articulators, the SA-AAI performance for tongue articulators matches with the corresponding SD-AAI performance with only $\sim$12.5% of the data used for SD-AAI training.

Index Terms: acoustic-to-articulatory inversion, BLSTM network, Adaptation

1. Introduction

Estimating articulatory movements from speech acoustic features is known as acoustic-to-articulatory inversion (AAI) [1]. There are several applications of AAI including speech recognition [2, 3], speech synthesis [4], speaker verification [5] and multimedia applications [6, 7, 8]. For subject dependent AAI (SD-AAI), various approaches have been proposed in the literature including codebook [9, 10], Gaussian mixture model (GMM) [11], Hidden Markov Model (HMM) [12], mixture trajectory model [13], Deep Neural Network (DNN) [14, 15, 16]. All these approaches need parallel acoustic-articulatory data for training AAI model, which, in turn, requires recording of speech and simultaneous motion of articulators from a subject of interest. Electromagnetic articulograph (EMA) is a promising technology to record such parallel data. In EMA recording, multiple sensors are glued to articulators of interest. While the presence of sensors has minimal impact on articulation [17] and does not pose any major safety issue, gluing sensor is a time consuming process and a subject requires good amount of time to get used to speaking naturally with sensors attached. Another challenge in the EMA recording is that the sensors could fall off in the middle of recording due to salivation or poor gluing [18]. Re-attaching sensor does not ensure placement of the sensor in its exact original position. This, in turn, causes discomfort to the subject as more time is spent to collect large amount of acoustic-articulatory data required for training a SD-AAI model.

To overcome these hazards during recording of the parallel acoustic-articulatory data, various techniques have been proposed in the literature for adaptation of only acoustic data using target subject with respect to several reference subjects. This methods are referred to as subject independent AAI (SI-AAI), where the articulatory motion predicted from test subject's speech does not belong to the articulatory space of the test subject but it belongs to the training subject only, unlike that in SD-AAI. Various acoustic space transformation techniques used in SI-AAI include vocal tract length normalization [19], cascade Gaussian mixture regression [20, 21], parallel reference speaker weighting [22] and by estimating the acoustic mismatch between training and test subjects using a generic acoustic space (GAS) [23]. Since there is a mismatch in both acoustics and articulatory space between test and training subjects in SI-AAI, these adaptation techniques do not generalize well to match with articulatory space of the test subject. So there is a need for adaptation of AAI model with low resource of parallel acoustic-articulatory data from the target subject.

With a motivation to reduce the amount of time spent on data recording for a new target subject, we, in this work, aim to define a new scope of AAI, referred to as subject adaptive AAI (SA-AAI), where the articulatory motion predicted using SA-AAI lies in the space of the test subject just like in SD-AAI but the parallel acoustic-articulatory data needed from the same subject during training is minimal unlike that of SD-AAI. The SA-AAI, while aims to reduce the demand for parallel data from the test subject during training, requires large amount of parallel acoustic-articulatory data from a fixed set of subjects, referred to as reference subjects. These data from reference subjects are used to train a generalized background AAI model (GBM-AAI) which is further adapted using small amount of parallel data of the target test subject. In the context of SA-AAI, we address the following questions: 1) How much parallel data from target test subject is required for adaptation of AAI model during training to achieve a performance similar to that of SD-AAI? 2) Does this requirement of parallel data vary in an articulator specific manner? To the best of our knowledge, there is no work in the literature that addresses adaptation of acoustic and articulatory space in a data driven manner for AAI. Motivated by the transfer learning paradigm in pattern recognition and computer vision [24], we propose to fine-tune the parameters of the GBM-AAI model instead of learning the SD-AAI model for a new target subject.

Experiments are performed using 11.4 hours of parallel recordings from 30 subjects. Experimental results reveal that the SA-AAI scheme performs on par with the SD-AAI schemes with $\sim$62.5% less data indicating that there is benefit in using the GBM-AAI model followed by the proposed adaptation scheme. When the performance of SA-AAI scheme is compared among different articulators, it is found that the amount

of parallel acoustic-articulatory data from the target subject needed to match the SD-AAI performance for tongue sensors is less than that for lip sensors.

2. Dataset

For this work, 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus [25] are chosen as the stimuli for data collection. We collect data from 30 subjects which comprises 13 females and 17 males subjects in an age group of 20-28 years. All the subjects are native Indians with proficiency in English, and reported to have no speech disorders in the past. Prior to the data collection, a consent form is signed by all subjects, as recommended by the institute's ethics committee. Prior to the recording, all subjects are familiarized with the 460 sentences to avoid any pronunciation error during recording. For each sentence, we record simultaneous audio and articulatory movement data.

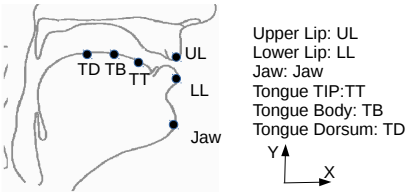

Figure 1: *Schematic diagram indicating the placement of EMA sensors.*

Electromagnetic articulograph (EMA) AG501 [26] is used to record the articulatory movement data for this study. AG501 has 24 channels to measure the horizontal, vertical and lateral displacements and angular orientations of a maximum of 24 sensors. The articulatory movement is collected with a sampling rate of 250Hz. We use 8 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 movement correction. The sensors are carefully placed following the guidelines provided in [27]. A schematic diagram of the sensor placement is shown in Fig. 1. Each of these eight sensors captures the movements of articulators in 3D space. In this study we consider the movements only in the midsagittal plane, indicated by X and Y directions in Fig. 1. Thus, we have twelve articulatory trajectories 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. Before placing the sensors, the subject is made to sit comfortably in the EMA recording setup. Once the sensors are placed for recording, the subjects are given sufficient time to get used to speaking naturally with the sensors attached to different articulators. During recording, subjects are given breaks whenever they felt tired of speaking and recording was resumed only when the subject felt comfortable to continue.

A microphone is placed near the subject to record the audio data at 48kHz synchronously with the articulatory data. During recording, the sentences are projected onto a screen placed in front of the subject. Manual annotation is performed to remove silence at the start and end of the sentences. After removing silence, the total duration of the entire acoustic-articulatory recording turns out to be 11.4 hours, where average durations of recording per subject is 22.8 ($\pm$ 2.5) minutes.

3. Proposed approach

The relation between the acoustic features and articulatory movements is known to be non-linear and non-unique [14, 3]. Also, the relation is not instantaneous, i.e., acoustic feature at a time need not be related to the articulatory position only at that time; instead it could be related to positions before and after that time. This is, for example, captured by using a fixed set of frames before and after the current frame in traditional DNN based AAI model. Even after incorporating such contextual information, the articulatory contours predicted by DNN based AAI model turn out to be jagged in nature, which is further post processed through low pass filtering. Deep recurrent neural networks architecture, namely, BLSTM network has shown to overcome the problems of capturing context and smoothing characteristics and achieves the state-of-art AAI performance [16]. So, in this work, we choose BLSTM network as a choice for AAI model. At frame index t, let $\mathbf{x}_t$ be the M-dimensional input and N be the number of memory cells in a LSTM layer with output $\mathbf{y}_t \in \mathbb{R}^N$. Then, for each LSTM layer there will be different weight vectors of type: input weights $\mathbf{W}_* \in \mathbb{R}^{N\times M}$, recurrent weights $\mathbf{R}_* \in \mathbb{R}^{N\times N}$ and bias weights $\mathbf{b}_* \in \mathbb{R}^N$ (where, '$_*$' corresponds to i, f, c, o). The forward pass for LSTM layer can be written as follows [29, 30]:

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

where, g is a point-wise non-linear activation function and $\odot$ denotes point-wise multiplication of two vectors. Total number of parameters for each LSTM layer are $4 \times (MN + N^2 + N)$. BLSTM layer creates a second separate instance of the LSTM layer to process the input sequences in two directions, namely, chronological and reverse order, which double the number of parameters required for training. Since, there is a large number of parameters to train, it, in turn, requires adequate amount of data.

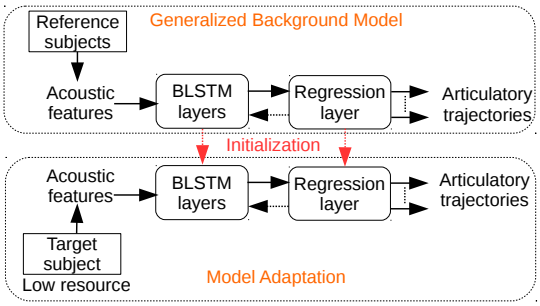

Figure 2: *Block diagram of SA-AAI involving model adaptation scheme.*

Fig. 2, shows the block diagram of the proposed SA-AAI. In the first step, it trains a GBM-AAI model from a set of reference subjects, where initial layers are BLSTM layers and last layer is time-distributed linear regression layer. GBM-AAI captures a mapping between acoustics and articulatory motion across many subjects. The weights of this GBM-AAI model are used as pre-trained weights for adaptation for a new target subject. Thus, the SA-AAI involves using the weights from the GBM-AAI model as the initial weights, and fine-tuning the

weights with the low resource target subject's data. To overcome the problems of over-fitting to the training data, we use validation data for early-stopping. The principle of SA-AAI is similar to the universal background model (UBM) in GMM for modeling the acoustic variability [31, 32] across the speakers and adapt it further to a target speaker.

4. Experimental Setup

The recorded speech from all the subjects is down-sampled from 48kHz to 16kHz. As an acoustic feature, we compute 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) [33] for every 20ms with a shift of 10ms. Further, for each sentence, cepstral mean subtraction is performed. The 12 dimensional articulatory data is post processed to obtain articulatory feature following the steps described next. It is known that the articulatory trajectories are smooth in nature, and most of the energy for all the articulators lies below 25Hz [34]. So, the recorded articulatory position data, is first low-pass filtered at 25Hz to avoid high frequency noise incurred due to EMA measurement error. The articulatory data is then down-sampled from 250Hz to 100Hz. Further, within every sentence we make each dimension of the articulatory feature zero-mean, since the average position for each sensor could change from utterance to utterance [3].

For every subject from the recorded 460 sentences, a fixed set of 368 utterances is chosen as a train set (80%), and the remaining 92 is divided equally for validation (10%) and test (10%) sets. The recorded data from 30 subjects are divided into two sets, namely Set-1 (8 male and 7 female) and Set-2 (9 male and 6 female). The experiments are performed in a two-fold manner, where in each fold we train three types of AAI models:

- Type-1: Pool all the training data (80%) from every subject in Set-1 and train a GBM-AAI model, denoted by GBM-AAI-1.
- Type-2: Consider $p\%$ of the training data from a target subject T in Set-2, and train a SA-AAI model by fine-tuning the weights obtained from GBM-AAI-1.
- Type-3: For a baseline comparison, train SD-AAI using complete training data of subject T in Set-2.

Type-2 and Type-3 AAI models are created separately for all subjects T in Set-2. Finally, a similar set of models are trained by swapping Set-1 with Set-2. This results in two generalized models for AAI, namely GBM-AAI-1 and GBM-AAI-2. In this work, four values of $p\%$ are considered: 12.5%, 37.5%, 62.5% and 100% of training data. This results in 30×4 =120 SA-AAI models and 30 SD-AAI models. Apart from utilizing GBM-AAI for initializing the weights for SA-AAI, we also use GBM-AAI for further assessment of SA-AAI models. We evaluate GBM-AAI-1 on test sets of Set-1 & Set-2 and evaluation results are denoted as GBM-1-M (matched test) and GBM-1-X (mismatch test). Similarly, GBM-2-M and GBM-2-X are computed.

For AAI model architecture during training, we deploy first three as BLSTM layers with 100 units followed by a linear regression layer. We consider, minimizing Minkowski-R error [35] between the original and predicted output as an objective function. Minkowski-R error function is of the form

$$E = \sum_{t=1}^{T}\sum_{i=1}^{12}\left|y_L^i(t) - d^i(t)\right|^{\mathrm{R}} \tag{2}$$

where, $d^i(t)$ and $y_L^i(t)$ are the original and predicted i^{th} articulatory position value at t^{th} test frame. In all the experiments, Adam [36] optimizer is used for training with early stopping using Keras library [37].

To assess the performance of AAI, we choose two evaluation metrics, Root Mean Square Error (RMSE) and Correlation Coefficient (CC) [34] for each articulator separately. RMSE and CC in i^{th} articulatory feature is given by

$$RMSE^i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(d^i(t) - y_L^i(t))^2}, \tag{3}$$

$$CC^i = \frac{\sum_{t=1}^{T}(d^i(t) - \bar{d^i})(y_L^i(t) - \bar{y}_L^i)}{\sqrt{\sum_{t=1}^{T}(d^i(t) - \bar{d^i})^2 \sum_{t=1}^{T}(y_L^i(t) - \bar{y}_L^i)^2}}. \tag{4}$$

where, $\bar{d^i}$ and $\bar{y}_L^i$ are the corresponding mean of $d^i(t)$ and $y_L^i(t)$ across the number of frames T.

5. Results and Discussion

Choice of R in the cost function: For training AAI models, we initially experimented with two different choices for R (R=1 & R=2). For R=1, the Minkowski-R takes the form of Mean Absolute Error (MAE). Similarly for R=2, the cost function reduces to Mean Square Error (MSE), which is the widely used cost function for AAI [14, 28]. The choice of R is based on empirical evaluation of GBM-AAI model. For each subject, we compute the average RMSE and CC over all the articulators. In Table. 1, we report mean (standard deviation (σ)) of average RMSE and CC across all target subjects, for different choices of R. In both matched (GBM-*-M) and mismatch (GBM-*-X) cases, MAE (R=1) performs better in terms of RMSE and CC compared to MSE (R=2). This improvement might be due to the fact that MSE is sensitive to the outliers and receives large contributions from the points which have the largest errors. The choice of R value less than 2 reduces the sensitivity to outliers. Similar results are also observed in SD-AAI models as well. So, R=1 is used for all further experiments.

Table 1: *Comparison of performance for choice of* R=1 *and* R=2 *(MSE vs MAE)*

Model	Metric	choice of R in cost function	
		R=2 (MSE)	R=1 (MAE)
GBM-1-M	CC	0.8703(.0286)	0.8921(.0318)
	RMSE	1.3639(.1301)	1.2978(.1886)
GBM-1-X	CC	0.75294(.0598)	0.7733(.0565)
	RMSE	1.9242(.2974)	1.8716(.2984)
GBM-2-M	CC	0.8512(.0509)	0.8735(.0487)
	RMSE	1.399(.3185)	1.3463(.2828)
GBM-2-X	CC	0.7275(.0606)	0.7433(.0563)
	RMSE	2.004(.3705)	1.9544(.3741)

Performance of SA-AAI: The SA-AAI performance (using CC and RMSE) on target subjects from Set-2 using GBM-AAI-1 model is shown in the first column of Fig. 3. The same is shown for target subjects from Set-1 using GBM-AAI-2 in the second column of Fig. 3. In all plots in Fig. 3, x-axis indicates the p % of data from the target subject for SA-AAI model. The CC and RMSE in all plots are computed following the steps described next. At first, for each subject T, we compute average RMSE and CC over all the articulators. Upon the average RMSE and CC, we compute mean and σ across all the subjects from the corresponding Set. First row represents the performance in-terms of mean CC and second row represents mean RMSE of SA-AAI and the corresponding σ is indicated by errorbars. For comparing with SA-AAI performance, the

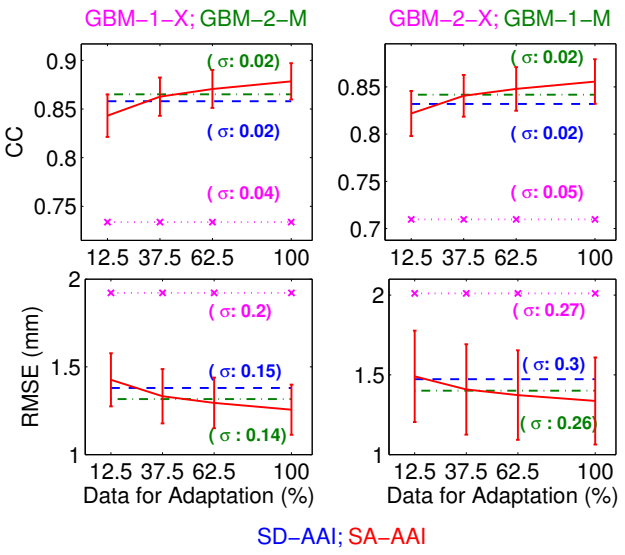

Figure 3: *CC (top row) and RMSE (bottom row) of SA-AAI averaged across all the articulators. First (second) column represents mean (σ) across the target subjects from Set 1 (Set 2). (SA-AAI (—),SD-AAI (---), GBM-*-M (- · -), GBM-*-X (-×-))*

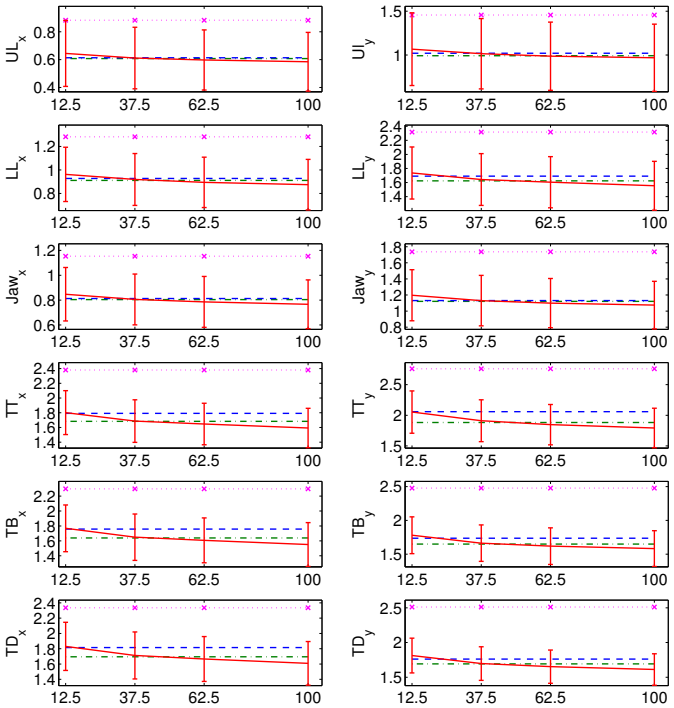

Figure 4: *RMSE (mm) of individual articulators averaged across all 30 subjects. (SA-AAI (—),SD-AAI (---), GBM-*-M (- · -), GBM-*-X (-×-)).*

average RMSE and CC (with σ in written bracket) using SD-AAI, GBM-*-M, GBM-*-X are plotted in the respective figures. While SD-AAI requires entire training data from the target, GBM-*-X does not require any training data from the target subject.

From Fig. 3, we observe that at the 12.5 % data from the target subject, the SA-AAI achieves a better performance than the GBM-*-X model. We also observe that at p=12.5%, the SA-AAI performance is similar to SD-AAI performance. At 37.5%, the SA-AAI model performance is better than the SD-AAI model. The superior performance of SA-AAI with small amount of target subject's data could come from using the GBM-AAI. At 62.5%, the SA-AAI outperforms both SD-AAI and GBM-*-M. It is also interesting to note that, at the availability of full training (p=100%), the performance of SA-AAI is better than the GBM-*-M and SD-AAI by a margin larger than that at p=62.5%. This suggests that instead of training an SD-AAI model using only the training data of a new target subject, better AAI performance could be achieved if a GBM-AAI model trained with a set of reference subjects (different from the target subject) is used for adaptation with target subject's training data. This could be due to the fact that GBM-AAI model captures rich acoustic-articulatory mapping from multiple reference subjects unlike that in SD-AAI model.

Performance of individual articulators: We also examine how the SA-AAI performance for an individual articulator varies with p. Fig. 4 shows the SA-AAI performance in terms of RMSE (mm) for each articulator (similar performance is observed using CC as-well). In Fig. 4, the RMSE is averaged across all 30 subjects (i.e., subjects in both Set-1 and Set-2). The percentage (p) of data from the target subject at which SA-AAI performance becomes superior to the performance of SD-AAI or GBM-*-M varies across articulators. For example, in the case of tongue articulators, namely, TT_x, TT_y, TB_x, TB_y, TD_x, TD_y, similar performance with respect to SD-AAI happens at p= ∼12.5%, while for LL_x and UL_x this happens only at p= ∼37.5%. This could imply that the motion of lower lip are more subject specific and that of tongue are more subject independent.

In summary, training a GBM-AAI model with data from references subjects helps to perform SA-AAI under low resource condition. At 37.5% of training data, we could able achieve a performance on par with SD-AAI. Using the pre-trained weights from GBM-AAI and fine-tuning them in SA-AAI scheme, helps even in the case of availability of full training data. It turns out that an improvement of 0.14 mm in average RMSE for SA-AAI (1.29 mm) is obtained with p=100% compared to that of SD-AAI (1.43 mm).

6. Conclusions

The proposed SA-AAI scheme enables to perform a low-resource AAI for a new target subject by fine-tuning weights learned from GBM-AAI model. Among different articulators, when the performance of SA-AAI scheme is compared with reference to SD-AAI, it is found that the amount of parallel acoustic-articulatory data needed from the target subject for tongue sensors is less than that for lip sensors. Note that in the current approach we used complete training data from reference subjects to train a GBM-AAI model. Further, experiments have to be conducted to verify how the performance of SA-AAI will vary, while training GBM-AAI with low resource of data from reference subjects as-well. Also, in the current experimental setup, Set-1 and Set-2 for training GBM-AAI is gender balanced. It will be interesting to investigate, if there is any gender dependency while performing SA-AAI for a target subject, by training (and choosing) GBM-AAI specific to the gender. These are the part of our future work.

7. Acknowledgements

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

8. References

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

[2] J. Frankel, K. Richmond, S. King, and P. 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] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[4] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. 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] 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.

[6] C. Ding, L. Xie, and P. Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2015.

[7] J. Jia, Z. Wu, S. Zhang, H. M. Meng, and L. Cai, "Head and facial gestures synthesis using pad model for an expressive talking avatar," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 439–461, 2014.

[8] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audio-visual speech synthesis based on pad," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 570–582, 2011.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] Z. Wu, K. Zhao, X. Wu, X. Lan, and H. Meng, "Acoustic to articulatory mapping with deep neural network," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9889–9907, 2015.

[15] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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

[17] N. Meenakshi, C. Yarra, B. Yamini, and P. K. Ghosh, "Comparison of speech quality with and without sensors in electromagnetic articulograph ag 501 recording," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 935–939.

[18] K. Richmond, "A multitask learning perspective on acoustic-articulatory inversion," in *Seventh Annual Conference of the International Speech Communication Association*, 2007, pp. 2465–2468.

[19] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion." in *INTERSPEECH*, 2016, pp. 455–459.

[20] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded gaussian mixture regression," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2246–2259, 2015.

[21] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded gaussian mixture regressions," in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 2013, pp. 2753–2757.

[22] A. Ji, M. T. Johnson, J. J. Berry, A. Ji, M. T. Johnson, J. J. Berry, A. Ji, M. T. Johnson, and J. J. Berry, "Parallel reference speaker weighting for kinematic-independent acoustic-to-articulatory inversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1865–1875, 2016.

[23] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2015.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[25] 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

[26] "3d electromagnetic articulograph, available online: http://www.articulograph.de/, last accessed:1/9/2016." [Online]. Available: http://www.articulograph.de/

[27] 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.

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

[29] 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.

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

[31] 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.

[32] D. Povey, S. M. Chu, and B. Varadarajan, "Universal background model based speech recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4561–4564.

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

[34] 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.

[35] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[36] 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

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