JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0000738

Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short term memory network

Aravind Illa and Prasanta Kumar Ghosh

Electrical Engineering Department, Indian Institute of Science, Bangalore, 560012, India
aravindi@iisc.ac.in, prasantg@iisc.ac.in

Abstract: Estimating articulatory movements from speech acoustic representations is known as acoustic-to-articulatory inversion (AAI). In this work, a speaker conditioned AAI (SC AAI) is proposed using a bi-directional LSTM neural network, where training is performed by pooling acoustic-articulatory data from multiple speakers along with their corresponding speaker identity information. For this work, 7.24 h of multi-speaker acoustic-articulatory data are collected from 20 speakers speaking 460 English sentences. Experiments with 20 speakers indicate that the SC AAI model performs better than SD AAI model with an improvement of correlation coefficient by 0.036 (absolute) between the original and estimated articulatory movements.

[Editor: Douglas D. O'Shaughnessy] Pages: EL171–EL176

Received: 30 October 2019 **Accepted:** 27 January 2020 **Published Online:** 13 February 2020

1. Introduction

Speech articulation involves movements of articulators including tongue, lips, jaw, and velum which form constriction in the vocal tract. The estimation of the articulatory movement from the acoustics is known as acoustic-to-articulatory inversion (AAI). Various models were proposed in the literature to learn AAI in a speaker dependent manner, e.g., Gaussian mixture model,[1] hidden Markov model (HMM),[2] and neural network based approaches.[3] Among the neural network based approaches, the long short term memory (LSTM), which is a recurrent neural network, achieves the state-of-art performance.[3]

Articulatory inversion learned with multiple speakers' acoustic-articulatory data have been shown to benefit (i) unknown speaker evaluation, (ii) in terms of improving the accuracy in a closed-set speaker condition. In unknown speaker evaluation, it has been shown that[4] speaker independent AAI benefits from transforming acoustic features of different speakers to a target speaker acoustic space using vocal tract length normalization. In a closed-set speaker condition (train and test with the same speakers), multi-speaker articulatory trajectory formation based on HMM and speaker-adaptive training were proposed and showed no statistically significant difference from that for speaker-dependent models.[5,6] To overcome the limitation on the amount of acoustic-articulatory data, a low resource AAI model was proposed using transfer learning and a generic AAI (GM AAI) model[3] with the LSTM network, which is trained by pooling the data from all speakers. It has been shown that the performance of the GM AAI is better than that of an individual speaker dependent AAI (SD AAI) model, which is trained using acoustic-articulatory data from a single speaker. This suggests that LSTMs can be used to learn acoustic-articulatory mappings of multiple speakers through a single AAI model rather than building separate speaker specific models. The GM AAI model can be further improved for a specific speaker by fine-tuning the GM AAI model with the acoustic-articulatory data specific to a speaker.[3] An alternative approach to fine-tuning GM AAI is to provide auxiliary features, which carry speaker specific information, along with the acoustic features for learning rich acoustic-to-articulatory mappings of multiple speakers. We hypothesize that conditioning GM AAI with speaker specific information using auxiliary features would be efficient and it would provide a more compact way of learning multiple AAI mappings. In this work, we propose an AAI model using the LSTM network by conditioning with the speaker specific information known as the speaker conditioned AAI (SC AAI) model. Its performance is compared with SD AAI, GM AAI, and GM AAI with speaker specific fine-tuning models.

For this work, we collected a new multi-speaker acoustic-articulatory database from 20 speakers speaking 460 English sentences, with an average duration of 21.72 ($\pm$2.19) minutes per subject. These multi-speaker acoustic-articulatory data can be used to study speaker specific articulatory signature, inter-speaker variability analysis, and learning multi-speaker AAI. In this

work, we aim to learn a single model using the SC AAI approach with multiple speakers' acoustic-articulatory data that could improve the accuracy of the AAI model in a closed-set speaker evaluation.

2. Dataset

We recorded acoustic-articulatory data using an AG501 electro-magnetic articulograph (EMA).[7] Acoustic-articulatory data were recorded from 20 speakers, comprising 10 male (M1–M10) and 10 female (F1–F10) speakers. All the speakers were proficient in English and reported to have no speech disorders in the past. For speech stimuli, we chose 460 MOCHA TIMIT sentences.[8] During recording, each sentence was projected on a computer screen placed in-front of the speaker, and a slide changer was provided to the speaker to navigate through all the sentences. We recorded simultaneous acoustic-articulatory data for each sentence.

Acoustic data were collected at a sampling rate of 48 kHz using the t.bone EM9600 shot-gun microphone[9] placed in front of the speaker. Articulatory data were collected by gluing (using "Epiglu"[7]) sensors of AG501 on six articulators, namely, upper lip (UL), lower lip (LL), jaw (Jaw), tongue tip (TT), tongue body (TB), and tongue dorsum (TD), following recommendations by Pattem *et al.*[10] particularly for the tongue sensors. Two additional sensors were glued on the mastoids for head movement correction. The sensors capture the articulatory movements in horizontal and vertical directions indicated by X and Y, respectively, in the midsagittal plane.[3] This results in a 12-dimensional articulatory feature vector whose elements are indicated 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. For the recorded acoustic-articulatory data we performed manual annotations to remove the start and end silence segments in each sentence.

3. Proposed approach

The acoustic to articulatory mapping function is known to be non-unique and nonlinear in nature.[11] Also, the articulatory movement trajectories are known to be smooth in nature.[11] In order to preserve the smoothness characteristics in estimated articulatory trajectories, these are further post-processed by either low-pass filtering[11] or using dynamic features with HMM[2] or Kalman filtering.[12] Neural networks have been shown to perform well in learning non-linear and complex functions.[11] In particular, bi-directional long short term memory (BLSTM) networks are known to model the time series data well and, hence, are used as the state-of-art modelling technique for the AAI task. BLSTM networks have also been shown to preserve the smoothness in the estimated articulatory trajectories.[3] In this work, we propose a speaker conditioned (SC) AAI model using BLSTM networks. We pool the acoustic-articulatory data from all speakers to train a SC AAI model together with the corresponding speaker identity information as an auxiliary feature. Although acoustic features implicitly carry speaker identity, we hypothesize that providing speaker identity explicitly would enable the network to learn the acoustic-articulatory mappings better.

Figure 1 illustrates the proposed approach for SC AAI. In SC AAI, the speaker identity information is provided as a one-hot encoded vector, the dimension of which is equal to the number of speakers in the training set. This one-hot representation of speaker identity information is fed to the embedding layer (indicated by "Embed layer" in Fig. 1) as an auxiliary information, while the acoustic features are fed to the dense layer. The outputs of the embedding and dense layers are concatenated as shown in Fig. 1. The concatenated vectors across all time frames of a sentence, which carry acoustic and speaker identity information, are fed as an input to the BLSTM layers. Note that the speaker identity information component remains same in the concatenated vectors across all time frames in a sentence. The output layer (indicated as "Regression layer" in Fig. 1) is a time distributed dense layer with linear activation function.

4. Experimental setup

The recorded acoustic-articulatory data using 460 sentences were divided into a training set 80% (364), validation set 10% (46), and testing set 10% (46) for each speaker. The recorded speech signal was down-sampled to 16 from 48 kHz. 13-dimensional Mel-frequency cepstral coefficients (MFCCs) were used as the acoustic features as they have been shown to be optimal for the AAI

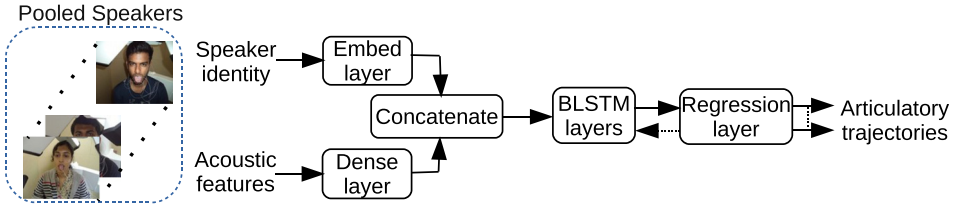

Fig. 1. (Color online) Block diagram of the proposed SC AAI model.

https://doi.org/10.1121/10.0000738

task.[11] The MFCCs were computed using a window size of 20 ms and a frame shift of 10 ms. To avoid high frequency noise, the articulatory position data were low-pass filtered with a cutoff frequency of 25 Hz, since most of the energy of the articulatory data lies below 25 Hz for all the articulators. Furthermore, the articulatory data were down-sampled to 100 from 250 Hz to obtain frame synchronization with acoustic features. For each dimension of the acoustic and articulatory feature vector we performed mean removal and variance normalization at an utterance level.

For SC AAI experiments, we chose 200 units for the dense layer which takes 13-dimensional MFCC as an input. As there are 20 speakers in the training set, the speaker identity information was represented by a 20-dimensional one-hot encoded vector. Note that one-hot vector is a sparse encoding, where only one index corresponding to a speaker is represented by integer value "one" and rest of the indices with "zero" integers. Therefore, an embedding layer was used to map the sparse binary valued one-hot encoded vector into the 32-dimensional continuous vector space. This also matches with the continuous vector representation of the dense layer output and thereby appropriate for concatenation. The outputs of the dense layer and embedding layer were concatenated and fed to the BLSTM hidden layers, comprising three hidden layers with 256-dimensional output units in each. We grouped acoustic-articulatory features of all the frames in an utterance to perform utterance by utterance training. In the train and test sets, we performed zero padding to obtain a fixed length sequence of 4 s (400 samples), which allows us to use a batch size of 50 to accelerate the speed of training. The mean squared error was chosen as the objective function to minimize the training loss and early stopping was performed based on the validation loss.

During testing of the SC AAI model, the speaker identity information may not be known. In order to estimate this information during test time, we also trained a closed-set speaker identification (SID) network using acoustic features, i.e., MFCCs. So, during testing, the speaker identity information was predicted using the SID network, having two LSTM units of 150 units each followed by a time distributed dense layer of 100 units and softmax layer. We chose categorical cross entropy as the loss function. All the experiments were performed using Keras with Tensorflow as back-end.

In order to compare with the performance of SC AAI, we considered baselines corresponding to different AAI models. The first column of Table 1 lists different AAI models, while the second, third, and fourth columns have yes/no entries to indicate whether pooling, fine-tuning, and speaker-conditioning were performed while training the AAI-models, respectively. Multiple speakers' acoustic-articulatory data were pooled to train GM AAI and SC AAI. Motivated from transfer learning approach, we further perform fine-tuning on GM AAI model with speaker specific acoustic-articulatory data, which results in speaker specific models indicated by GM-FSD. When speaker identity is unknown, SC AAI is evaluated with estimated speaker identity (ESI) from SID network indicated by SC-ESI. The last column indicates the number of AAI models required for all the speakers in the training set. To evaluate the performance of different AAI-models, we chose root mean square error (RMSE) and correlation coefficient (CC)[3,11] as evaluation metrics computed for each articulator separately.

5. Results and discussions

In this section, we present the results of the experiments which compare the performance of SC AAI with that of the baseline AAI models in a speaker specific manner, following which the AAI performance is also presented in an articulator specific manner. Finally, we present the results of SC AAI in the case of mismatched speaker identity information.

The accuracy of SID network used in the SC-ESI scheme is 95.54% and 95.00% on the validation and test sets, respectively. Figure 2 shows the performance of five different AAI models using boxplots for each speaker, where each box represents the first quartile (bottom edge of the box), median (horizontal line inside the box), and third quartile (upper edge of the box) of the distribution of CC computed between the predicted and the original articulatory trajectories and averaged across all the articulators. We perform a t-test on the CC values from all test sentences of all speakers to examine statistical significance in the difference between the mean CC

Table 1. Different AAI models used for comparison in this work.

AAI-model	Pooling	Fine-tuning	Speaker identity	# Models
Speaker dependent (SD)	no	no	no	# speakers
Generic model (GM)	yes	no	no	single
Fine-tuning GM (GM-FSD)	yes	yes	no	# speakers
Speaker conditioned (SC)	yes	no	yes (direct)	single
SC-ESI	yes	no	yes (estimated)	single

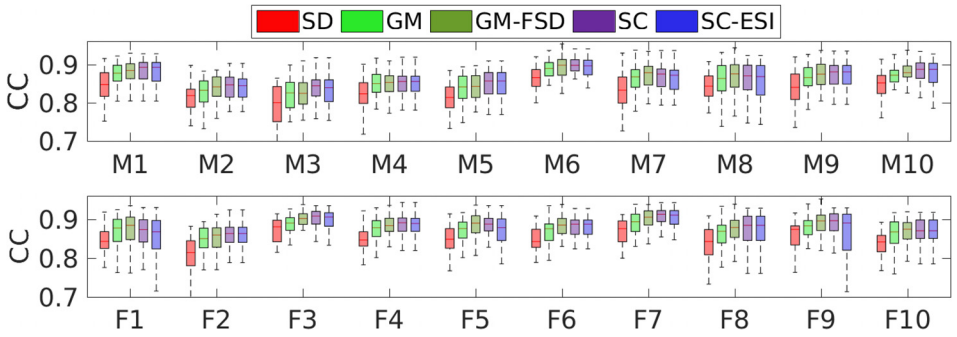

Fig. 2. (Color online) CC averaged across all articulators for each of 20 speakers (male and female speakers in top and bottom row, respectively).

obtained using SC AAI and that using the baseline models. We observe that CC values improves significantly ($p < 0.05$) with SC AAI compared to SD AAI, GM AAI, and GM FSD models. We further perform t-test in a speaker specific manner. We observe that the SC AAI performs significantly ($p < 0.05$) better than the SD AAI model consistently across all the speakers. Similarly, performance improvements are observed while comparing SC AAI model with GM AAI, and found to be statistically significant ($p < 0.05$) across all the speakers except for M4, M7, M8, and F1. While comparing GM-FSD and SC AAI models in a speaker specific manner, significant ($p < 0.05$) difference is observed only for three speakers, namely, M3, M5, and F12. For rest of the subjects, the performance of SC AAI is on par with GM-FSD, which suggests that a single SC AAI model captures the multiple speakers' acoustic-to-articulatory mappings what 20 models do in the case of GM-FSD. We also observe that there is no significant drop in the performance when SC-ESI AAI used is compared to when SC AAI is used, except for F5 and F9. This could be because individual speaker classification accuracy using SID network indicates that F5 and F9 are the subjects which yield the least classification accuracies of 80.43% and 82.61%, respectively, followed by M6 with 89.13%. The performance of SC-ESI AAI suggests that in the absence of speaker information, speaker identity information estimated using SID network can be utilized to perform speaker conditioned AAI without any drop in performance.

Figure 3 reports the CC from different AAI models for individual articulator averaged across all the speakers. There is a significant ($p < 0.05$) improvement in CC using SC AAI compared to that using SD AAI for all articulators. An analysis of the improvements achieved by the individual articulators reveals that the maximum relative improvement is observed for lip articulators (in particular 9.47% for LL_y) and the minimum for tongue articulators (in particular 3.15% for TD_y). Similarly, a comparison of SC AAI with GM AAI reveals that there is consistent improvement across all the articulators with the maximum relative improvement in CC for UL_y (2.53%) and minimum for TT_y (0.65%). This implies that the prediction of lip position benefits from the speaker identity information and could be because lip articulators are more speaker dependent compared to tongue articulators. These are consistent with previous findings,[3] where lip articulators demand more speaker specific acoustic-articulatory data compared to tongue articulators while fine-tuning the GM AAI model.

Table 2 reports the performance of different AAI models in terms of RMSE and CC averaged across all articulators and speakers. Note that the RMSE is computed on articulatory trajectories which are mean and variance normalized. So RMSE does not have any units. We find that among all the AAI models, SC AAI has the best performance followed by GM-FSD AAI model. From Table 2, we observe that instead of training a SD AAI model using the data only from a single speaker, training a GM AAI model by pooling data from multiple speakers results in an improvement in the AAI performance. This indicates that BLSTM networks are

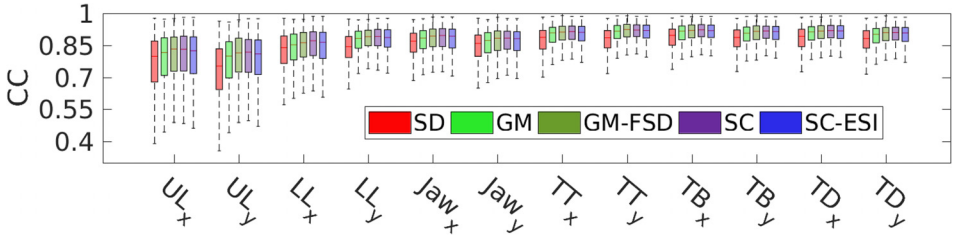

Fig. 3. (Color online) Articulatory specific CC averaged across all speakers.

https://doi.org/10.1121/10.0000738

Table 2. RMSE and CC averaged across all the articulators and speakers with different AAI models.

	SD	GM	GM-FSD	SC	SC-ESI
RMSE	1.178	1.096	1.067	**1.049**	1.072
	(0.075)	(0.087)	(0.085)	(0.077)	(0.070)
CC	0.836	0.860	0.869	**0.872**	0.865
	(0.020)	(0.018)	(0.020)	(0.019)	(0.019)

able to learn multiple acoustic-articulatory mappings in a single model resulting in an improvement in the performance compared to that using the SD AAI model. Thus, the GM AAI model captures a rich acoustic-articulatory representation for all the speakers, which is further improved by fine-tuning with speaker specific data, i.e., by GM-FSD AAI models. Although the GM AAI model could learn the speaker specific information implicitly from acoustic features, thereby speaker specific acoustic-to-articulatory mappings, there is a scope for improvement in AAI for each speaker by doing speaker specific fine-tuning as is evident from the GM-FSD performance. By conditioning speaker specific identity information explicitly, the proposed SC AAI model could leverage that scope for improvement resulting in the best AAI performance using a single model. Note that we further fine-tuned SC AAI in a speaker specific manner which resulted in a RMSE of 1.0487 ($\pm$0.084) which is similar to that using the SC AAI model reported in Table 2. This indicates that a single SC AAI model is optimal for speaker dependent AAI without any further need for fine-tuning, unlike the GM-FSD AAI model. All the experiments are performed with mean and variance normalized articulatory trajectories, which minimize the morphological variations across speakers and reduce the inter speaker variability. The magnitude of improvements using SC-AAI over GM AAI could be more if we perform AAI without normalizing the articulatory trajectories. Experiments are also performed using acoustic features from a particular speaker fed along with speaker identity information from another speaker. This results in performance drop in mismatched speaker identity cases compared to matched ones, which is due to the inter-speaker variability in the articulatory motion. Although the SC AAI network performs better than the GM AAI network in seen speaker case evaluation, the current SC AAI approach does not generalize to the unseen speaker cases. This limitation of the current approach could be solved by incorporating speaker embedding from the SID networks[13] instead of one-hot representations of speaker identity information. These are parts of our future investigation.

6. Conclusion

In this work we proposed an SC AAI network that utilizes auxiliary information of the speaker identity along with the acoustic features to estimate articulatory movements. Experiments with 20 speakers indicate that speaker conditioning is the key to achieving a better AAI performance with the proposed SC AAI model compared to the baseline AAI models. Comparing the SC AAI model with the GM AAI model results in an improvement in performance. Comparison with the GM-FSD AAI model indicates that there is no significant difference in performance for all subjects, but it provides a compact way of learning multiple speakers' acoustic-articulatory mappings within a single SC AAI model. In the future, we would investigate different approaches to extend the SC AAI to benefit speaker independent AAI and focus on the applications including speech recognition and synthesis tasks.

Acknowledgments

We thank all subjects for participating in EMA data collection and the Pratiksha Trust and the Department of Science and Technology (DST), Government of India, for their support.

References and links

[1]T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Commun. **50**(3), 215–227 (2008).
[2]S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. Speech Audio Proc. **12**(2), 175–185 (2004).
[3]A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *Proceedings of Interspeech* (2018), pp. 3122–3126.
[4]G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," J. Acoust. Soc. Am. **146**(1), 316–329 (2019).
[5]S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," Speech Commun. **48**(12), 1677–1690 (2006).
[6]S. Hiroya, "Acoustic-to-articulatory inversion using a speaker-normalized HMM-based speech production model," in *International Seminar on Speech Production* (2008).

[7]"3D electromagnetic articulograph," http://www.articulograph.de/ (Last viewed 07/10/2019.)
[8]A. Wrench, "MOCHA-TIMIT," speech database, Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh (1999), http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (Last viewed 8 February 2020).
[9]EM 9600 shotgun microphone, http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/ (Last viewed 07/10/2019.)
[10]A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," Comput. Speech Lang. **47**, 157–174 (2018).
[11]P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," J. Acoust. Soc. Am. **128**(4), 2162–2172 (2010).
[12]S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proceedings of the 5th Seminar on Speech Production* (2000), pp. 237–240.
[13]D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE (2018), pp. 5329–5333.