INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

An investigation on speaker specific articulatory synthesis with speaker independent articulatory inversion

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 speech representations from articulatory movements is known as articulatory-to-acoustic forward (AAF) mapping. Typically this mapping is learned using directly measured articulatory movement in a subject-specific manner. Such AAF mapping has been shown to benefit the speech synthesis applications. In this work, we investigate the speaker similarity and naturalness of utterances generated by AAF which is driven by the articulatory movements from a subject (referred to as cross speaker) different from the speaker (target speaker) used for training AAF mapping. Experiments are performed with directly measured articulatory data from 9 speakers (8 target speakers and 1 cross speaker), which are recorded using Electromagnetic articulograph AG501. Experiments are also performed with articulatory features estimated using speaker independent acoustic-to-articulatory inversion (SI-AAI) model trained on 26 reference speakers. Objective evaluation on target speakers reveal that the articulatory features estimated from SI-AAI result in a lower Mel-cepstrum distortion compared to that using directly measured articulatory features. Further, listening tests reveal that the directly measured articulatory movements preserve the speaker similarity better than estimated ones. Although, for naturalness, articulatory movements predicted by SI-AAI perform better than the direct measurements.

Index Terms: articulatory-to-acoustic mapping, acoustic-to-articulatory inversion, voice conversion

1. Introduction

Speech is produced as a result of temporal overlap of articulatory gestures namely, lips, tongue tip, tongue body, tongue dorsum, velum, and larynx, which regulate constriction in different parts of the vocal tract [1]. Knowledge of articulatory kinematics together with acoustic information have shown benefit in various applications like, speech recognition [2, 3], speech synthesis [4, 5], speaker verification [6] and multimedia applications [7, 8, 9]. With the advancements in deep learning techniques, articulatory information has also shown success in silent speech interfaces (benefit patients who have lost their voice due to laryngectomy or diseases affecting the vocal folds) such as in speech recognition [10] and speech synthesis directly from articulatory position information alone [11, 12].

Certain modifications in articulation style or accent is difficult to perform in acoustic domain [13], which are easily descibed via the position and dynamics of the articulators [14]. In [15, 16], it has been shown that accent conversion applications benefit by driving the articulatory synthesizer of a non-native speaker (L2) with articulatory movements of native speaker (L1). However such a system depends on both L1 and L2 acoustic-articulatory data to learn a transformation from L1 to L2 articulatory space or L1 articulatory to L2 acoustic space [16]. In [15], it has also been observed that articulatory features (AFs) are more speaker independent in nature compared to acoustic features. In [17], authors report that the linguistic content is captured in the vocal tract's front cavity while speaker-dependent information in the back cavity. These findings motivate us to investigate the performance of articulatory-acoustic forward mapping (AAF) of a target speaker (TS) when driven with the AFs of cross-speaker (CS) (without performing any transformation), in-terms of speaker similarity and intelligibility with respect to the TS. The objective is to understand the degree to which the AFs are speaker independent while preserving linguistic information. These findings could help to avoid the need for acoustic-articulatory data prior to learning transformation from CS to TS and benefit in applications such as accent and voice conversion, and personalizing the voices in silent speech interfaces.

The articulatory measurements obtained directly from electromagnetic articulograph (EMA) have been shown to be effective in predicting spectrum of speech and benefit the articulatory based speech synthesis applications [14, 18]. However, collecting a large amount of acoustic-articulatory data for each pair of source and target speaker. In the absence of parallel acoustic-articulatory data, techniques are proposed in the literature to estimate articulatory movements from speech acoustic features [19]. It is known as acoustic-to-articulatory inversion (AAI). Interestingly, in [20], it has been shown that using AFs from AAI has performed better than the direct AFs in articulatory synthesis. However this AAI is trained in a speaker-dependent (SD) manner on a single speaker from whom parallel acoustic-articulatory data is required to train SD-AAI. Various methods are proposed in the literature to estimate the articulatory features from speech acoustics in a speaker independent manner known as speaker independent AAI (SI-AAI), which does not require parallel acoustic-articulatory data from the test speaker [21, 22] during training. With the advancement of the SI-AAI methods, in this work, we would also like to compare the performance of AAF performed with articulatory features directly measured as well as estimated from SI-AAI. To the best of our knowledge no work has been reported on articulatory synthesis with AFs estimated in a subject independent manner.

The focus of the work is to investigate AAF on two aspects: a) In the absence of direct acoustic-articulatory data, can estimated AFs from SI-AAI be utilized to learn AAF, and how the performance of AAF alters with respect to direct AFs? b) If a TS specific AAF is driven by the AFs from a CS, does the synthesized speech sound similar to that of the TS speaker?

The rest of the paper is organized as follows. We begin with the explanation of the data collection process followed by the proposed approach. Section 4 presents the experimental setup followed by the results and discussion.

http://dx.doi.org/10.21437/Interspeech.2019-2664

2. Data Collection

For this work, we recorded synchronous acoustic-articulatory data using EMA AG501 [23]. We collected data from 35 speakers comprising 17 males and 18 females in an age group of 20-28 years. No speaker were reported to have any speech disorders in the past. Prior to the data collection, a consent form was signed by all speakers, as recommended by the institute ethics committee. 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus [24] were chosen as the stimuli for the data collection.

Fig. 1 illustrates the placement of sensors on articulators for recording [19]. The sensors were placed on the articulators such that the inter-sensor distances closely match with the recommendation provided following the guidelines reported in [25]. Sensors were glued on six articulators, namely, upper lip (UL), lower lip (LL), jaw (Jaw), tongue tip (TT), tongue body (TB), tongue dorsum (TD). We also glued two sensors behind the ears for head movement correction [26]. After all the sensors were glued to the respective articulators, enough time was given to the subject to get adjusted and comfortable to speak with sensors attached. Stimuli sentences were projected on the slides in a computer screen in front of the subject. All 460 sentences were recorded in a single session for a subject, resulting in a recording duration of 2 to 3 hours.

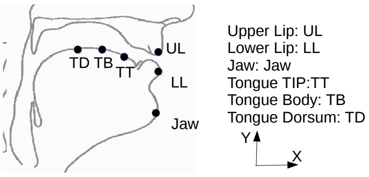

Figure 1: *Schematic diagram indicating the placement of EMA sensors.*

From each of the articulators we obtained horizontal (X) and vertical (Y) movements, which resulted in 12 dimensional articulatory features, namely UL_x, UL_y, LL_x, LL_y, Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, TD_y. This 12 dimensional articulatory data was post-processed to obtain AFs. It is known that most of the energy for all the articulators lies below 25Hz, and the articulatory trajectories are smooth in nature [27]. Hence to avoid high frequency noise incurred due to EMA measurement error, the recorded articulatory position data, was first low-pass filtered with a cut-off frequency of 25Hz. The articulatory data was then down-sampled from 250Hz to 100Hz. Further for every sentence, we normalize each dimension of the articulatory feature to zero-mean and unit variance, since the average position for each sensor could change from utterance to utterance [3] and to reduce the effect of morphological variations across all the speakers. On the other-hand, the recorded speech was down-sampled from 48kHz to 16kHz. As an acoustic feature, we computed 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) [28] vector for every 20ms with a shift of 10ms. Further, for each sentence, cepstral mean subtraction was performed. Manual annotations were done to remove the begin and end silence segments from all the recordings.

3. Proposed Approach

The proposed approach for investigating articulatory-to-acoustic model with direct and estimated articulatory movements is shown in Fig. 2. At first, we extract acoustic features from speech signal using WORLD vocoder [29], such as spectral envelope, pitch and aperiodicity. Further, on spectral envelope we perform mel-scale frequency wrapping and discrete cosine transform to obtain 36-dim mel-cepstrum (MCEP). The objective of speaker specfic AAF model is to estimate the MCEP from the AFs. The AFs can be either acquired directly from EMA or estimated using SI-AAI. A brief review on SI-AAI & AAF and the details of the models deployed for the proposed approach are presented below.

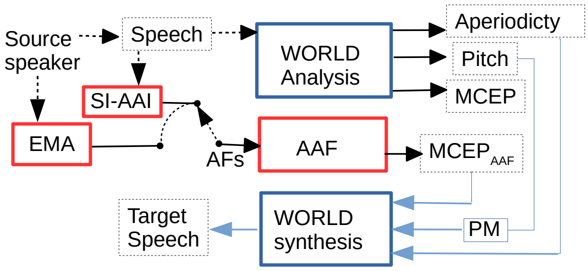

Figure 2: *Illustration of articulatory-to-acoustic forward mapping setup based on the proposed approach.*

Acoustic-to-articulatory inversion: Due to the unavailability of parallel acoustic-articulatory data, various techniques have been proposed in the literature to estimate articulatory data for test speaker using the acoustic-articulatory data from reference speakers [21, 22] . These methods are referred to as subject independent AAI (SI-AAI). With this approach the articulatory motion predicted from the test speaker's speech is represented from the generic articulatory space of the reference subjects used for training SI-AAI. Among the existing approaches, deep recurrent neural networks architecture, namely, bi-directional long short term memory (BLSTM) network has been shown to achieve the state-of-the-art AAI performance [30, 19]. We deploy the AAI model proposed in [19] as the SI-AAI model, where BLSTM networks are used to learn inverse mapping from acoustic features (MFCC) to articulatory features obtained from the reference subjects. Initial layers of the SI-AAI model are BLSTM layers and last layer is a time-distributed linear regression layer. SI-AAI captures rich AAI mapping learned from many reference subjects' acoustic-articulatory data. Thus, an SI-AAI model predicts articulatory movements which are represented from the reference subjects' articulatory space.

Articulatory-to-acoustic forward mapping: In AAF, forward mapping function is learned from articulatory movements to estimate acoustic features in a subject specific manner. Different methods have been proposed in literature for AAF, such as Gaussian mixture models (GMM) [31], hidden Markov models (HMM) [32], and deep neural networks (DNN) [14] and recurrent neural networks [33]. In [18], a comparison has been made across all the methods and it was shown that BLSTM performs better among the existing statistical methods. Hence, we use BLSTM networks for AAF as well, where it takes input features to feed BLSTM layers followed by a time-distributed linear regression layer to predict acoustic features (MCEP). We learn AAF models separately for direct and estimated AFs from SI-AAI.

For both SI-AAI and AAF, we use mean squared error as the loss function between the predicted output and original trajectory (AFs/MCEP) to train the BLSTM network. AAF models are trained in a speaker dependent manner separately for all TS speakers, while in testing phase, the source AFs could be from either TS or CS. During reconstruction, we transform original pitch from source (x) to target speaker (y) statistics with a linear transform: $\hat{F}^y = \frac{\sigma^y}{\sigma^x}(F^x - \mu^x) + \mu^y$, where F denotes pitch (in log-scale) [34] and μ, σ represent mean and standard deviation of feature F respectively. The reconstructed MCEP, transformed pitch and original aperiodicity are fed to the WORLD vocoder synthesizer to obtain target speech. For all the experiments, we use Adam [35] optimizer for training with early stopping. All implementations are done using Keras library [36].

4. Experimental Setup

In this work we are investigating the robustness of articulatory features in driving AAF with AFs from TS and CS, where AFs are obtained by (i) direct articulatory measurements from EMA, (ii) articulatory features estimated from acoustics using SI-AAI. For this purpose, we split the 35 subjects' acoustic-articulatory data collected in the following manner. We consider 26 reference speakers' (13 male and 13 female) acoustic-articulatory data to train SI-AAI model. For experiments with AAF model we consider the remaining 9 speakers, out of which 8 speakers (4 male M1 to M4; and 4 female F1 to F4) were considered as TS and 1 female speaker as CS. CS is used to evaluate the AAF in the mismatched condition.

For SI-AAI model architecture, we deploy the first three layers as BLSTM layers with 256 units followed by a linear regression output layer with 12 units to predict AFs. We use 20 speakers' acoustic-articulatory data for training SI-AAI and other 6 speakers as a validation set for early stopping to avoid over-fitting on the training speakers. As acoustic feature we use MFCC, as it has been shown to be the best for AAI using maximal mutual information criterion between acoustic and articulatory features [27].

While choosing input AFs for AAF model, we add voiced/unvoiced (VUV) information to 12 dimensional AFs, since AFs obtained from EMA does not carry voicing information. This results in a 13 dimensional input AFs to AAF model which predicts 36 dimensional MCEP. Similar to SI-AAI, we choose an architecture for AAF, where first three layers are BLSTM layers with 256 units followed by a 36 units linear regression output layer. For each speaker, from the recorded set of 460 sentences, a fixed set of 368 sentences is chosen as the train set (80%), and the remaining 92 is divided equally for validation (10%) and test(10%) sets. AAF models were trained separately for all eight TS speakers in a speaker dependent manner. During testing, for each TS speaker there are four different AFs based test cases depending on the source speaker and the type of articulatory features. a) TS-EMA: AFs measured directly with EMA from TS, b) TS-AAI: AFs estimated using SI-AAI from TS's MFCC, c) CS-EMA: AFs measured directly with EMA from CS, and d) CS-AAI: AFs estimated using SI-AAI from CS's MFCC. (a) and (c) are evaluated on AAF trained with direct AFs from EMA; (b) and (d) are evaluated on AAF trained with AFs predicted by SI-AAI.

For a baseline comparison, we also train a model which directly maps MFCC to MCEP (M2M) without going to articulatory space (as it is done by inverse mapping MFCC to AFs by SI-AAI and then forward mapping with AFs to MCEP by AAF). The architecture for M2M is similar to that of AAF, which is learned by replacing AFs with MFCC to predict MCEP directly. This results in two additional test case evaluations (e) TS-MFCC: MFCC of TS evaluated on M2M of TS, (f) CS-MFCC: MFCC of CS evaluated on M2M of TS. These together with four earlier test cases result in three matched case test evaluations ((a), (b) and (e)) and three mismatched case test evaluations ((c), (d), and (f)).

Evaluation metric: For evaluation we use both objective and subjective evaluation metrics. For matched case evaluation where AAF is trained and tested with same TS, we use Mel-cepstral distortion (MCD) [34] as an objective measure, which is computed between the original MCEP and predicted MCEP. For subjective evaluation, listening tests are performed to asses the performance of the proposed approach in terms of naturalness and speaker similarity. For evaluating the quality of naturalness, we perform mean opinion score (MOS) test. In this test, listeners are asked to evaluate the naturalness of synthesized utterance from AAF with reference to a natural sentence spoken by TS on a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For naturalness test, we randomly select three sentences from the test set for each TS speaker in every test case from (a) to (f), which results in comparing all 18 (=6x3) pairs of sentences per TS. For evaluating the speaker similarity, we choose the ABX test [34]. In the ABX test, A= "AFs" from (a) to (d), and B= "MFCC" from (e) to (f) of TS/CS based synthesized speech are compared with respect to the X= "original speech" of the TS speaker. This consists of 4 different sets, (a) vs (e), (c) vs (e), (b) vs (f), and (d) vs (f) with reference to natural sentence spoken by TS. Here also, we randomly choose three sentences for each TS speaker which results in 12 (=4x3) triplet sentences for testing TS similarity within each TS's AAF model. We randomize the order in which we present A and B sentences to avoid bias. We performed listing tests with 12 listeners. Each listener was presented with synthesized sentences for two TS speakers (1 male and 1 female) for both naturalness and ABX test in a single session.

5. Results and Discussion

In this section, we first present the results of SI-AAI. Next, we will report the results of matched case and mismatched case evaluation of AAF with TS and CS speaker's AFs respectively.

5.1. SI-AAI

To assess the performance of SI-AAI model we report pearson correlation coefficient (CC) [19] between the predicted and the original articulatory trajectories. Table 1 reports the average CC (standard deviation (SD) in bracket) computed across all the articulators estimated with SI-AAI model using MFCC as acoustic features. We observe that, on average, we obtain 0.77 (0.10) CC across all the TS subjects using SI-AAI. Among all the articulators, we observe that the performance is a worse with UL and LL compared to TT, TB and TD. In particular, a minimum CC of 0.54 (0.10) is obtained for UL_y and a maximum CC of 0.88 (0.03) is obtained for TT_y. This could be due to the fact that kinematics of lips are more subject specific while those of tongue are relatively more subject invariant, which are found to be consistent with the previous literature [19].

Table 1: *Performance of SI-AAI in terms of cc averaged across all the articulators for CS and all TS speakers*

speakers	M1	M2	M3	M4	F1	F2	F3	F4	CS
Avg CC (SD)	0.81 (.09)	0.73 (.14)	0.71 (.22)	0.78 (.10)	0.79 (.10)	0.77 (.12)	0.79 (.11)	0.76 (.09)	0.72 (.14)

5.2. Matched case evaluation:

In the matched case evaluation we trained and tested AAF with the same TS speaker. For each TS speaker we trained two AAF models: i) direct AF obtained from EMA to MCEP , ii) estimated AF estimated with SI-AAI model to MCEP. Table 2,

Table 2: *MCD values obtained from the matched case evaluation of AAF with all TS speakers and CS.*

Subjects	M1	M2	M3	M4	F1	F2	F3	F4	CS
TS-MFCC	3.40 (.15)	3.35 (.14)	3.01 (.14)	3.54 (.15)	3.54 (.18)	3.48 (.17)	3.44 (.31)	3.42 (.17)	3.41 (.14)
TS-EMA	5.28 (.31)	5.40 (.29)	4.98 (.39)	5.75 (.30)	5.58 (.36)	5.50 (.22)	5.51 (.31)	5.84 (.35)	5.50 (.28)
TS-AAI	5.01 (.23)	4.97 (.24)	4.55 (.25)	5.48 (.26)	5.32 (.29)	5.29 (.22)	5.27 (.26)	5.43 (.27)	5.39 (.30)

Table 3: *MOS values from the listening test on naturalness in matched evaluation*

	TS-MFCC	TS-EMA	TS-AAI
MOS	4.11 (.31)	2.97 (.51)	3.37 (.76)

reports the MCD of test utterances for each TS speaker, where row 2, 3 and 4 correspond to matched test case evaluation of TS-MFCC, TS-EMA and TS-AAI, respectively. We observe that MFCC performs better than the AFs, at both directly measured and estimated AFs. Interestingly, we observe that SI-AAI features perform better than the direct EMA AFs. Apart from the objective results we also perform subjective listening tests. The naturalness of synthesized speech is reported in Table 3 in terms of MOS. Results obtained with MOS are found to be consistent with MCD i.e, MFCC performs better followed by estimated AFs compared to direct AFs. Note that the MOS obtained with WORLD vocoder is 4.20 (0.78). The best performance of MFCC in both objective and subjective evaluations of M2M could be due to the fact that both MFCC and MCEP are representations of the same spectrum. The improvement of MOS and MCD with AFs estimated using SI-AAI compared to the direct AFs are found to be consistent with the results with SD-AAI results [20].

5.3. Mismatched case evaluation:

In the mismatched case evaluation, we use AFs from CS to drive the AAF model of each TS speaker. The subjective evaluation in terms of MOS in the mismatched case are reported in Table 4. We observe that, in the mismatched case, the naturalness scores are similar between the MFCC and AFs estimated using SI-AAI. And we observe that due to mismatch in test and train subjects, the performance using MFCCs drops from 4.11 (0.31) to 3.34 (0.53) MOS, but performance using estimated AFs remains similar to the matched case results.

To asses the amount of speaker similarity preserved, we report ABX preference score results in Fig. 3. The first two bars in Fig. 3 indicate the ABX results with test case evaluation of (a) vs (e) and, (b) vs (e) respectively. The last two bars correspond to the ABX results with test case evaluation of (c) vs (f) and, (d) vs (e). We can observe that, in the matched case, MFCC performs better than the AFs. But, in the mismatched case, AFs yield better preference compared to MFCC.

Table 4: *MOS results from the listening test on naturalness in mismatched evaluation*

	CS-MFCC	CS-EMA	CS-AAI
MOS	3.34 (.53)	2.51 (.74)	3.30 (.53)

The mismatched performance of AFs estimated from SI-AAI with CS, suggests that the AFs carry more linguistic information and less speaker-dependent factors. This could help to use AFs in accent or voice conversion application. This could be helpful to synthesize target speaker speech in an unknown language (for target speaker) as well. In this aspect, we carried out preliminary experiments with native language (Telugu) of a CS speaker, where we have acoustic-articulatory data collected in a manner similar to that described in Section 2 for English. The AFs predicted with SI-AAI in case of CS speaker result in a CC of 0.71 (0.16) across all the articulators. AFs estimated with SI-AAI from MFCC of CS in Telugu are used to drive the AAF of TS speakers which are trained with English. Experiments and listening tests reveal that synthesized speech has a naturalness of 3.01 (0.52) MOS. For speaker similarity, ABX test shows that AFs obtain a preference of 64.58% over CS MFCC in Telugu. Further investigation is needed to synthesis the speech with multiple languages from more number of CSs to validate the cross speaker and cross language articulatory speech synthesis. Sample synthesized files from this work are available online[1].

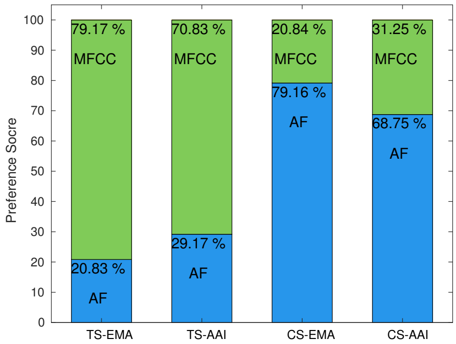

Figure 3: *ABX preference test results with MFCC and articulatory features.*

6. Conclusion

We performed experiments with direct and estimated AFs using SI-AAI for driving the AAF model. Interestingly, objective and subjective measures reveal that estimated AFs perform better than direct AFs in matched case evaluation. This results will enable us to deploy articulatory speech synthesis even in the absense of acoustic-articulatory data from target speaker using SI-AAI. Experiments with mismatched case evaluation, by driving TS specific AAF with AFs from CS, reveal that the synthesized speech samples by AAF are similar to TS without significant drop in MOS for naturalness unlike MFCC. Further investigation is needed to understand what aspects in the AFs estimated from SI-AAI are beneficial. Also, in future, we would like to investigate on synthesizing target speakers speech in multiple languages. These are parts of our future work.

Acknowledgements: Authors thank all the volunteers for their participation in the data collection and listening tests. We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

[1] https://spire.ee.iisc.ac.in/IS19_AAFdemo/

7. References

[1] L. Goldstein and C. 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] 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] B. Cao, M. J. Kim, J. P. van Santen, T. Mau, and J. Wang, "Integrating articulatory information in deep learning-based text-to-speech synthesis." in *INTERSPEECH*, 2017, pp. 254–258.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2323–2336, 2017.

[11] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[12] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[13] H. Hermansky and D. Broad, "The effective second formant f2'and the vocal tract front-cavity," in *International Conference on Acoustics, Speech, and Signal Processing,*. IEEE, 1989, pp. 480–483.

[14] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[15] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301–2312, 2012.

[16] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7694–7698.

[17] D. J. Broad and H. Hermansky, "The front-cavity/f 2 hypothesis tested by data on tongue movements," *The Journal of the Acoustical Society of America*, vol. 86, no. S1, pp. S113–S114, 1989.

[18] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion using blstm-rnns with augmented input representation," *Speech Communication*, vol. 99, pp. 161–172, 2018.

[19] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," *Proc. Interspeech 2018*, pp. 3122–3126, 2018.

[20] S. Aryal and R. Gutierrez-Osuna, "Articulatory inversion and synthesis: towards articulatory-based modification of speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7952–7956.

[21] 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.

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

[23] "3d electromagnetic articulograph, available online: http://www.articulograph.de/, last accessed:21/10/2018." [Online]. Available: http://www.articulograph.de/

[24] 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

[25] 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.

[26] C. Kroos, "Using sensor orientation information for computational head stabilisation in 3d electromagnetic articulography (EMA)," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2009, pp. 776–779.

[27] 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.

[28] 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.

[29] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[30] 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.

[31] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 31–36.

[32] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda, "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[33] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks." in *Interspeech*, 2016, pp. 1502–1506.

[34] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[35] 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

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