INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Transfer learning of articulatory information through phone information

Abdolreza Sabzi Shahrebabaki[1], *Negar Olfati*[1], *Sabato Marco Siniscalchi*[2], *Giampiero Salvi*[1,3], *Torbjørn Svendsen*[1]

[1] Department of Electronic Systems, NTNU
[2] Department of Computer Engineering, Kore University of Enna
[3] KTH Royal Institute of Technology, Dept. of Electrical Engineering and Computer Science

{abdolreza.sabzi, olfati, giampiero.salvi, torbjorn.svendsen}@ntnu.no,
marco.siniscalchi@unikore.it

Abstract

Articulatory information has been argued to be useful for several speech tasks. However, in most practical scenarios this information is not readily available. We propose a novel transfer learning framework to obtain reliable articulatory information in such cases. We demonstrate its reliability both in terms of estimating parameters of speech production and its ability to enhance the accuracy of an end-to-end phone recognizer. Articulatory information is estimated from speaker independent phonemic features, using a small speech corpus, with electromagnetic articulography (EMA) measurements. Next, we employ a teacher-student model to learn estimation of articulatory features from acoustic features for the targeted phone recognition task. Phone recognition experiments, demonstrate that the proposed transfer learning approach outperforms the baseline transfer learning system acquired directly from an acoustic-to-articulatory (AAI) model. The articulatory features estimated by the proposed method, in conjunction with acoustic features, improved the phone error rate (PER) by 6.7% and 6% on the TIMIT core test and development sets, respectively, compared to standalone static acoustic features. Interestingly, this improvement is slightly higher than what is obtained by static+dynamic acoustic features, but with a significantly less. Adding articulatory features on top of static+dynamic acoustic features yields a small but positive PER improvement.

Index Terms: Articulatory inversion, transfer learning, speech recognition, deep learning

1. Introduction

Parameters related to the position and movement of the articulators involved in speech production can be of use in numerous applications. Examples include automatic speech recognition (ASR) [1, 2], speech synthesis [3, 4], pronunciation training [5] and description of the speech production mechanism. The articulatory parameters can be derived by measuring the articulators' kinematics through different methods, such as magnetic resonance imaging (MRI) [6], X-ray microbeam [7], ultrasound [8] and electromagnetic articulography (EMA) [9, 10, 11]. Among these methods EMA is most frequently adopted as it allows using higher sampling rates and simple pre-processing is sufficient to extract the articulatory features from the measurements.

However, measuring the articulatory trajectories directly is not applicable in most real world applications since it requires instrumentation not available outside laboratories, and imposes heavy burdens on the subjects. Thus, in order to utilize articulatory parameters in speech processing applications, we need to estimate them from more accessible information. The most obvious information source is the speech acoustic waveform, and the task to be accomplished is acoustic-to-articulatory inversion (AAI). AAI is challenging from several aspects. The first problem is the one-to-many mapping problem because several articulator gestures may produce the same acoustic speech signal. A common approach to address this problem is to employ trajectory based deep neural networks [12, 13, 14, 15]. The next problem is insufficient amounts of data for adequate modeling of the acoustic space, leading to inferior performance for speaker independent (SI) scenarios compared to the speaker dependent (SD) scenarios, or matched speakers compared to mismatched speakers in SI scenarios. For the articulatory space, lack of data is also important, but the articulatory domain exhibits in general less variation compared to the acoustic space, which makes it less speaker dependent.

In scenarios where the textual content of the spoken utterance is known linguistic information, e.g. the predicted phone sequence for that utterance, can be used. Indeed, to cope with scarcity of input data for modeling the acoustic space in the AAI task, augmenting the acoustic features with linguistic information has been shown to improve the performance [16, 13, 15] for SD scenarios. Systems utilizing the linguistic information alone have also been reported to work quite well [17, 15] even when using binary features, e.g. one-hot encoded phonemic features (PHN, phone identity) or binary articulatory feature vectors, where multiple features can be active simultaneously [15]. The performance of linguistic information based articulatory inversion (AI) is in line with the reported results in [18], which confirms that front articulators in the vocal tract are related to the linguistic content and the back cavity articulators are more speaker specific. We report in [19] that utilizing linguistic features improves both SD and SI cases significantly. That performance boost is due to less variation between speakers in the linguistic space that is built from a limited set of discrete binary value vectors, in contrast with the acoustic space that is a continuous valued space. In fact, the speaker variability in the linguistic space is limited to the phone duration in the uttered speech sequence.

The advancement in deep neural networks for the task of AI and the positive effect of exploiting PHN features in this task motivate us to propose a new transfer learning approach for AI. We extract articulatory knowledge from a speech corpus providing articulatory measurements, e.g., the "Haskins production rate comparison" (HPRC), and use transfer learning to convey the knowledge to a scenario where articulatory measurements are not available, e.g., the TIMIT [20] phone recognition task. To this end, a teacher model is trained to perform phone-to-articulatory inversion (PAI) on HPRC. The trained teacher provides articulatory targets needed to build a student model that performs acoustic-to-articulatory inversion

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-1139

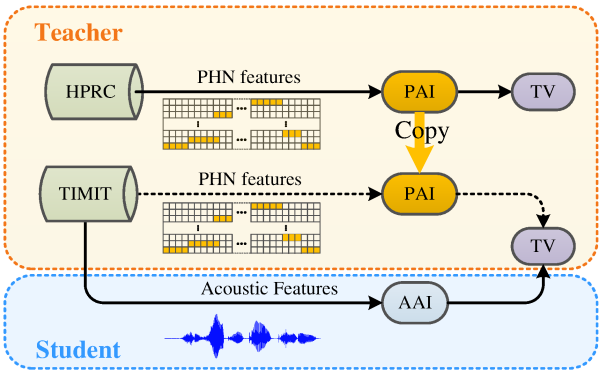

Figure 1: *Block diagram of the proposed transfer learning method from the HPRC to the TIMIT database, and knowledge distillations from phonemic features to acoustic features through articulatory space. Dashed arrows correspond to no training.*

(AAI) on TIMIT. Finally, we use the articulatory information that we estimate on TIMIT through AAI, as input features to perform phone recognition, demonstrating that articulatory features boost phone recognition accuracy.

The rest of paper is organized as follows. The proposed transfer learning method is described in Section 2. Corpora and evaluation methods are in Sections 3 and 5, respectively. Experiments and results are described in Section 5 followed by Section 6 to conclude our work.

2. Teacher-student approach to articulatory information transfer

The proposed approach is motivated by the following observation: Articulatory information can be useful for various speech processing tasks, such as ASR. However, such information is not usually available in corpora for speech recognition. Moreover, it may not be possible to estimate articulatory parameters from the speech signal (AAI) with a satisfactory level of accuracy, and speaker adaptive AAI suitable for typical ASR scenarios is a challenging task. To overcome this, we propose to use phonemic to articulatory inversion (PAI), which is speaker independent by design, as a bridge between scenarios where AAI can be estimated, and speech technology applications where this is usually not the case.

To put forth our solution, we define the following feature sets, and models. The acoustic features, $\mathbf{x} \in \mathbb{R}^n$, the articulatory features, $\mathbf{y} \in \mathbb{R}^m$, and the phone features, $\mathbf{p} \in \mathbb{B}^l$, where $\mathbb{R}$ is the field of real numbers, and $\mathbb{B}$ is the Boolean field. A teacher neural architecture is built on HPRC data to perform the mapping $f_{\text{PAI}} : \mathbb{B}^l \rightarrow \mathbb{R}^m$, from phonemic to articulatory features. This mapping is shown in the upper part in Figure 1. The teacher model not only performs PAI for the HPRC task, but it also provides the articulatory targets for performing PAI with TIMIT data. This process is shown in the middle part in Figure 1, where the PAI architecture is copied to be used with TIMIT phone features at its input and generates articulatory feature estimates at its output. Finally, a student neural architecture is built to perform the mapping $f_{\text{AAI}} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ on the TIMIT task. The inputs are acoustic features extracted from the TIMIT waveforms; the outputs are articulatory targets, provided by the teacher neural networks. This step is shown in the bottom part in Figure 1.

With the above feature sets and models, we are ready to use f_{AAI} in order to recover the articulatory features directly from the speech signal without using any annotations. Those articulatory features can be used e.g. as supplemental information in an ASR task, with the goal of improving the overall system performance. In sum, we have built a framework to transfer the knowledge embedded into the articulatory parameters available in the HPRC task to the TIMIT task by using f_{PAI} and f_{AAI} systems, avoiding to address the mismatch between different recording settings and speaker characteristics through a adaptation stage, which is the conventional solution.

The two neural architectures used for articulatory estimation and shown in Figure 1 were trained by minimizing the mean square error (MSE) between estimated values and the ground truth. Those two neural architectures accomplish the following tasks:

Phone-to-articulatory inversion - PAI: This model is trained to estimate the output articulatory features, $\mathbf{y}$, from the input PHN features, $\mathbf{p}$. The PAI neural architecture consists of two bi-directional long short-term memory (BLSTM) layers having 128 cells for each forward and backward directions.

Acoustic-to-articulatory inversion - AAI: The AAI neural structure is a combination of five stacked 1-D convolutional layers of kernel size [1,3,5,7,9], followed by two BLSTM layers with 128 cells in each direction. The convolutional layers extract features from the input acoustic features, $\mathbf{x}$, and the BLSTM layers model temporal dynamics in the system and estimate the articulatory features, $\mathbf{y}$.

3. Corpora

3.1. HPRC

The "Haskins Production Rate Comparison"(HPRC) [11], is a multi-speaker EMA corpus with data from four female and four male native American English speakers. Sampling rates for the speech signal and the EMA recordings are 44.1kHz and 100Hz, respectively. Eight sensors were used to measure the articulators' trajectories. Those eight sensors are placed at the tongue rear (TR), tongue blade (TB), tongue tip (TT), upper and lower lip (UL and LL), mouth left (ML), jaw or lower incisors (JAW) and jaw left (JAWL). The sensors movements are measured in the midsagittal plane in X, Y and Z direction, which denote movements of articulators from posterior to anterior, right to left and inferior to superior, respectively. In the HPRC corpus, sensors do not record significant movements in Y direction; we therefore generate information related to the articulatory movements by employing the geometrical transformations defined in [21] on the X and Z directions. Nine tract variables (TVs) are obtained, namely: Lip Aperture (LA), Lip Protrusion (LP), Jaw Angle (JA), in addition to Constriction Degree and Location for Tongue Rear (TRCD, TRCL), Tongue Blade (TBCD, TBCL) and Tongue Tip (TTCD, TTCL). The sampling rate of the articulatory features was maintained. The HPRC speech signals were resampled to 16kHz to match the TIMIT sampling rate.

3.2. TIMIT

The TIMIT database [22] consists of 6300 sentences spoken by 630 speakers from 8 major dialect regions of the United States. There is a predefined portion for training consisting of all the SX and SI sentences from 462 speakers with a total of 3696 sentences. The sentences from the remaining 168 speakers are meant for development and testing purposes. We will follow

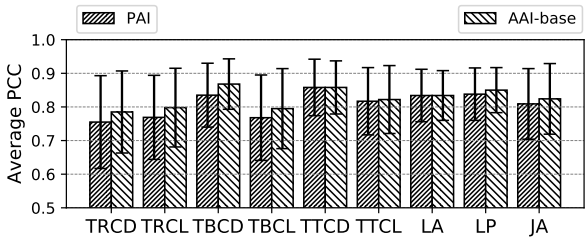

Figure 2: *Averaged PCC and standard deviation for different tract variables of the HPRC test set.*

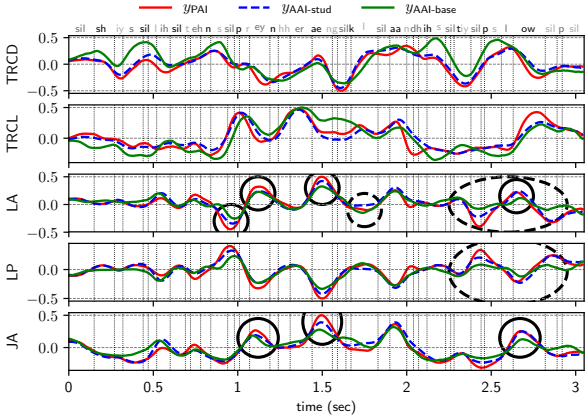

Figure 3: *TV trajectories from f_{PAI}, $f_{AAI\text{-}base}$, and $f_{AAI\text{-}stud}$ for utterance "She slipped and sprained her ankle on the steep slope."*

[23] and use the core test set spoken by 24 speakers for testing and the development set spoken by 50 speakers for validation. The core test set consists of 192 utterances and the development set consists of 400 utterances.

4. Evaluation methods

We used two evaluation methods to assess the proposed technique. The first method computes the Pearson's correlation coefficient explicitly on the target articulatory parameters. The second method is implicit and aims at demonstrating the effectiveness of our approach by inspecting the effects of using estimated articulatory features on the TIMIT phone recognition task.

4.1. Pearson's correlation coefficient

To measure the performance of the articulatory inversion methods, the Pearson's correlation coefficient (PCC) [24] is adopted. The PCC measures the similarity between estimated and ground truth trajectories and is defined as:

$$\mathrm{PCC} = \frac{\sum_i (y(i) - \bar{y})(\hat{y}(i) - \bar{\hat{y}})}{\sqrt{\sum_i \left(y(i) - \bar{y}\right)^2 \sum_i \left(\hat{y}(i) - \bar{\hat{y}}\right)^2}}, \qquad (1)$$

where $y(i)$ and $\hat{y}(i)$ are the ground truth and estimated parameters value of the i^{th} frame respectively and $\bar{y}$ and $\bar{\hat{y}}$ are mean values of $y(i)$ and $\hat{y}(i)$.

4.2. End-to-end phone recognizer

There is no actual ground truth articulatory measurements for TIMIT; therefore, we verify the performance of the proposed approach through the phone error rate (PER) of a phone recognizer built on TIMIT data. In particular, the ESPnet recognizer [25] is used in this work. This phone recognizer is based on (i) an end-to-end encoder-decoder with hybrid connectionist temporal classification (CTC), and (ii) an attention mechanism [26]. The encoder part contains four layers of BLSTM with 320 cells, one layer of LSTM for the decoder with 300 cells, location-aware attention mechanism with 10 convolution filters of length 100, and the same weight, 0.5 for the CTC and attention losses. The interested reader is referred to [26] for more details.

5. Experiments & Results

We evaluate two different types of AI systems, namely PAI- and AAI-based systems. The PAI and AAI systems trained on HPRC material are referred to as f_{PAI} and $f_{\text{AAI-base}}$, respectively, and validated using the PCC measure. In order to assess the f_{PAI} accuracy for TIMIT data, the estimated TVs are visualized and discussed with regards to the speech production mechanism. The student model, which is referred to as $f_{\text{AAI-stud}}$, trained on the TIMIT acoustic data, is assessed from the inversion performance point of view, with the average PCC measure computed using the f_{PAI} as ground truth. An example of estimated TVs for $f_{\text{AAI-stud}}$ and $f_{\text{AAI-base}}$ are visualized. In addition, a comparative ASR performance test is carried out for the TIMIT corpus in terms of PER, to compare efficiency of the $f_{\text{AAI-base}}$ and $f_{\text{AAI-stud}}$ systems and their complementary information for ASR task. Implementations of AI systems are performed using Keras [27] with TensorFlow backend [28].

5.1. Articulatory, phonemic, & acoustic representations

The TVs are calculated for the HPRC data at a rate of 100Hz. In order to have the same 100Hz rate for the acoustic and phonemic feature, a 25ms sliding analysis window and 10ms frame shift are used for acoustic feature extraction. The spoken utterances in HPRC corpus were labeled with the Penn phonetics lab forced aligner [29]. There are 61 phone categories which are folded onto 39 categories [30] to match the conventional 39 phones used in TIMIT [20]. Each phone is represented as a one-hot 39-dimensional vector (PHN) [17]. For TIMIT, we use PHN features for estimating the TVs with the teacher network. For AAI accomplished through the student network, we use the feature vectors consisting of 13 Mel frequency cepstral coefficients (MFCCs). Finally, 23-dimensional Mel filter bank log energies (FBE) are employed along with 3 estimated pitch and voicing features as 26-dimensional static acoustic features in the ESPnet phone recognizer. We also consider first and second derivatives of the FBEs in the phone recognition task.

5.2. Phone-to-articulatory inversion on HPRC

The f_{PAI} input is a 39-dimensional phonemic feature vector, including silence. It should be noted that starting and ending silences have been removed with an energy based threshold speech activity detection (SAD) procedure. Moreover, the 9-dimensional TV features are utterance-based z-score normalized and scaled to be in range $(-0.5, +0.5)$. Training data from the all eight speakers is used to build the f_{PAI} system; whereas validation data is employed with the goal of preventing overfitting. In Fig. 2, we observe that the f_{PAI} is able to predict the articulators in the front vocal cavity akin to the $f_{\text{AAI-base}}$ system. This is inline with what reported in [18, 31], namely that the front articulators capture the linguistic content. The back cav-

Table 1: *PER for acoustic features and their combinations with the estimated TVs from $f_{AAI\text{-}stud}$ and f_{PAI}. D denotes feature dimensionality.*

feature type	D	Dev PER	Test PER
x	26	25.6%	27.9%
x, $y_{\text{AAI-base}}$	35	20.9%	23.3%
x, $y_{\text{AAI-stud}}$	35	**19.6%**	**21.2%**
x, Δx, $\Delta^2 x$	78	19.8%	21.4%
x, Δx, $\Delta^2 x$, $y_{\text{AAI-base}}$	87	19.8%	22.8%
x, Δx, $\Delta^2 x$, $y_{\text{AAI-stud}}$	87	**19.1%**	**20.8%**

Table 2: *Lower bound of PER for the estimated TVs from f_{PAI} combined with the FBEs.*

feature type	D	Dev PER	Test PER
y_{PAI}	9	12.3%	13.3%
x, y_{PAI}	35	8.8%	9.5%
x, Δx, $\Delta^2 x$, y_{PAI}	87	8.2%	9.1%

ity articulators relate closely to speaker specific properties as it is mentioned in [31], and this is reflected by the less precise prediction capability of the PAI system than the AAI system.

5.3. Acoustic-to-articulatory inversion on HPRC

The performance of $f_{\text{AAI-base}}$ system in terms of PCC is shown in Fig. 2. As discussed before, PCC values are comparable for f_{PAI} and $f_{\text{AAI-base}}$ systems for front vocal cavity. For the back cavity, the $f_{\text{AAI-base}}$ system performs better. We can attribute the better performance of the AAI in comparison with the PAI, to the matched speaker independent training style.

5.4. Teacher-student approach to AAI on TIMIT

In the proposed teacher-student approach to perform transfer learning and extract articulatory estimates from acoustic information, we use the f_{PAI} system previously trained on HPRC as the teacher. Articulatory parameters are estimated in terms of TV for TIMIT by feeding TIMIT phonemic transcriptions into the f_{PAI} system. In Fig. 3, we can observe (inside the solid ellipses) that for production of the stop sound /p/, the LA is decreasing and LP is increasing, vowel /æ/ has wider LA or JA than vowels /eɪ/ or /oʊ/, which is inline with dropping of the jaw in production of vowel /æ/ while the jaw is slightly open in /eɪ/ or closed in /oʊ/. Evaluation of the student model ($f_{\text{AAI-stud}}$) is carried out by the average PCC measure, which is 0.929 for the core test set of TIMIT. The PCC distribution is shown in Fig. 4 for each TVs. Estimations from $f_{\text{AAI-stud}}$ and $f_{\text{AAI-base}}$ are visualized in Fig. 3. We can observe that at the end of the utterance (inside the dashed ellipses), the values of the $f_{\text{AAI-base}}$ estimation do not decrease or increase for lip separation or protrusion, respectively, when the stop sound /p/ is present and it is expected to have lowest values for the LA compared to the other phones in this sequence of phones. We can see the $f_{\text{AAI-base}}$ estimation of the LA for /l/ is less than the estimated value for /p/ which is wrong because for production of /p/ lips are closed and for production of /l/ lips are separated. That implies the $f_{\text{AAI-base}}$ model does not provide correct information with respect to speech production constraints.

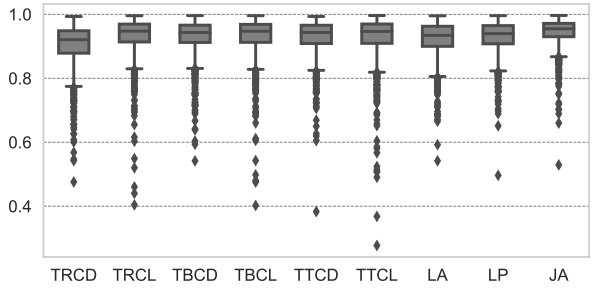

Figure 4: *Distribution of PCC between estimated TV trajectories from $f_{AAI\text{-}stud}$ and f_{PAI}.*

5.5. Exploiting TV estimates in phone recognition

We now explore the role of articulatory information in the task of phone recognition. The ESPnet recognizer in Section 4.2 is employed to build all of our phone recognizers. Several experiments are conducted in order to gain insights on the role of the TV estimates in speech recognition. In the initial experiment, we train the phone recognizer on static acoustic features, (x), only. In the second experiment, we include dynamic features to x and denote it as $(x, \Delta x, \Delta^2 x)$. The phone recognizers based on acoustic features only serve as baseline systems. The PER for different input features is reported in Table 1. $y_{\text{AAI-stud}}$ combined with x, significantly improves the recognition accuracy, and reduce the PER by 6.7% on the test set. Interestingly, a slightly better PER, +0.2%, is obtained by replacing the 52-dimensional dynamic acoustic features $(\Delta x, \Delta^2 x)$ with the 9-dimensional $y_{\text{AAI-stud}}$. Moreover, we can observe that employing the $y_{\text{AAI-stud}}$ obtains better performance than the $y_{\text{AAI-base}}$. The combination of $y_{\text{AAI-stud}}$ with $x, \Delta x, \Delta^2 x$ reduces the PER by 0.6%.

Finally, we used the TV features y_{PAI} (obtained from the phonemic transcriptions) alone and combined with $x, \Delta x, \Delta^2 x$ to calculate the lower bound of PER in this problem. The results are shown in table. 2.

6. Conclusions

This work proposes a new teacher-student method to transfer articulatory knowledge from the HPRC corpus through phonemic features onto the TIMIT corpus, which is purely acoustic. We exploit the transferred knowledge to build an acoustic to articulatory inversion (AAI) system for TIMIT with the goal of improving ASR performance. In this way, we obtained 0.6% improvements compared to the baseline system for PER when the mixed acoustic and estimated articulatory representations are used. Similarly we obtain better PER combining static acoustic and articulatory features (35 dim.) compared to dynamic acoustic features (78 dim.) proving that articulatory features are a more efficient representation of the dynamics of speech production. We also show that our method performs better than transferring AAI models trained on the HPRC corpus with acoustic adaptation. In the future, we will work on transfer learning of both acoustic and phonetic features to improve the performance of our AI system and getting closer to the PER lower bound.

7. Acknowledgements

This work has been supported by the Research Council of Norway through the project AULUS, and by NTNU through the project ArtiFutt. The third author is supported by the PRIN 2007 project nr. JNKCYZ_002.

8. References

[1] J. Frankel and S. King, "ASR-articulatory speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[2] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[3] K. Richmond and S. King, "Smooth talking: Articulatory join costs for unit selection," in *ICASSP*, 2016, pp. 5150–5154.

[4] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, 2013.

[5] T. Hueber, A. Ben Youssef, G. Bailly, P. Badin, and F. Elisei, "Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory HMM for pronunciation training," in *Interspeech*, 2012, pp. 783–786.

[6] S. Narayanan, K. N. S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[7] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

[8] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.

[9] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[10] R. Korin, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the MNGU0 articulatory corpus," in *Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[11] M. Tiede, C. Y. Espy-Wilson, D. G. V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[12] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[13] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information," in *Interspeech 2016*, 2016, pp. 1497–1501.

[14] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *ICASSP*, 2015, pp. 4450–4454.

[15] A. S. Shahrebabaki, N. Olfati, A. S. Imran, S. M. Siniscalchi, and T. Svendsen, "A Phonetic-Level Analysis of Different Input Features for Articulatory Inversion," in *Interspeech*, 2019, pp. 3775–3779.

[16] P. Zhu, X. Lei, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *Interspeech*, 2015, pp. 2192–2196.

[17] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated RNN," in *Interspeech*, 2018, pp. 3112–3116.

[18] D. J. Broad and H. Hermansky, "The front-cavity/f2′ hypothesis tested by data on tongue movements," *The Journal of the Acoustical Society of America*, vol. 86, no. S1, pp. S113–S114, 1989. [Online]. Available: https://doi.org/10.1121/1.2027307

[19] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals," in *press, Interspeech*, 2020.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[21] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2014.

[22] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recpgnition Workshop*, L. S. Baumann, Ed., Feb 1986, pp. 100–109.

[23] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *EUROSPEECH*, 1997, pp. 401–404.

[24] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech 2018*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[26] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, Dec 2017.

[27] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[28] M. Abadi and et al., "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[29] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[30] K. . Lee and H. . Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[31] A. Illa and P. K. Ghosh, "An Investigation on Speaker Specific Articulatory Synthesis with Speaker Independent Articulatory Inversion," in *Proc. Interspeech 2019*, 2019, pp. 121–125. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2664