



Unsupervised Acoustic-to-Articulatory Inversion with Variable Vocal Tract Anatomy

Yifan Sun, Qinlong Huang, Xihong Wu

Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

{yifan_sun, huangqinlong}@pku.edu.cn, wxh@cis.pku.edu.cn

Abstract

Acoustic and articulatory variability across speakers has always limited the generalization performance of acoustic-to-articulatory inversion (AAI) methods. Speaker-independent AAI (SI-AAI) methods generally focus on the transformation of acoustic features, but rarely consider the direct matching in the articulatory space. Unsupervised AAI methods have the potential of better generalization ability but typically use a fixed morphological setting of a physical articulatory synthesizer even for different speakers, which may cause nonnegligible articulatory compensation. In this paper, we propose to jointly estimate articulatory movements and vocal tract anatomy during the inversion of speech. An unsupervised AAI framework is employed, where estimated vocal tract anatomy is used to set the configuration of a physical articulatory synthesizer, which in turn is driven by estimated articulation movements to imitate a given speech. Experiments show that the estimation of vocal tract anatomy can bring both acoustic and articulatory benefits. Acoustically, the reconstruction quality is higher; articulatorily, the estimated articulatory movement trajectories better match the measured ones. Moreover, the estimated anatomy parameters show clear clusterings by speakers, indicating successful decoupling of speaker characteristics and linguistic content.

Index Terms: acoustic-to-articulatory inversion, vocal tract anatomy

1. Introduction

Acoustic-to-articulatory inversion (AAI) intends to estimate movements of articulators, e.g., tongue and lips, from acoustic signals. The estimated articulatory movements can be used for data augmentation when developing text-to-speech (TTS) and automatic speech recognition (ASR) systems, especially in low resource cases [1]. And articulatory information is also useful for Computer-Assisted Pronunciation Training (CAPT) and Computer-Aided Language Learning (CALL) [2, 3].

Traditionally, parallel acoustic-articulatory data are used to train statistical models, such as codebook approaches [4], Kalman filtering [5] and Hidden Markov Model (HMM) [6]. In recent years, the introduction of deep neural networks has significantly improved the estimation accuracy of articulatory movements [7, 8, 9]. However, due to the inter-speaker variability, the cross-speaker generalization performance of supervised AAI approaches is still unsatisfactory. Several speaker independent AAI (SI-AAI) methods have been proposed to solve this problem, such as vocal tract length normalization (VTLN) [10, 11], parallel reference speaker weighting [12], cascade Gaussian mixture regression [13] and generic acoustic space (GAS) [14]. These methods apply acoustic transformations to bridge from unseen speakers to known speakers. However, the matching in the acoustic space can not guarantee the

matching in the articulatory space [15], as a result, the generalization performance of these methods is still not satisfactory.

Unsupervised AAI methods imitate how human beings learn to speak and generally work in an analysis-by-synthesis manner. Such methods typically iteratively adjust control parameters of a physical articulatory synthesizer to reproduce a target utterance. To this end, several optimization methods have been proposed, such as random search [16], distal learning [17] and genetic algorithms [18]. Recently, Shibata et al. proposed to train an articulatory inference network with deterministic policy gradients (DPG) [19, 20]. Theoretically, unsupervised AAI methods have the potential to generalize well to unseen speakers. However, existing methods typically fix the vocal tract configuration of articulatory synthesizers with measured anatomy parameters (e.g., the vocal tract length) from a specific speaker, even when applied to different speakers [17, 19]. In this case, the inter-speaker variability is mapped into the intra-speaker variability during the speech imitation process and this implicit speaker normalization has been shown to limit the reconstruction performance of F_0 and formant frequency [17].

Considering the process of speech production, the inter-speaker variability lies in the variability of vocal tract anatomy and articulation habits [21], where the latter can be characterized by the trajectories of articulatory movements. Differences in vocal tract anatomy affect both the acoustics and the articulation of speech. Acoustically, vocal tract length is closely related to the formant frequency [22, 23], and evidence from theory and simulation suggests that the anatomy of the hard palate and posterior pharyngeal wall can also affect the resonant properties of the vocal tract [24, 25]. Articulatorily, speakers appear to compensate for the physical difference in the vocal tract by adjusting lingual articulation, so as to reduce acoustic variations [25]. For example, during the articulation of coronal fricatives, changes of the palate shape will lead to changes of jaw height and tongue body position [26, 27]. As a result, the variability of the vocal tract anatomy will bring about the variability of articulation.

In this work, we propose that jointly estimate the vocal tract anatomy and articulatory movements for unsupervised AAI can bring about benefits. We adopt the VocalTractLab (VTL) [28] as the articulatory synthesizer, of which the anatomy of a modeled vocal tract can be specified by several parameters. The Embodied Joint Embedding framework (EmJEm) [29] is adopted to train a neural network to infer not only time-varying articulatory trajectories but also the anatomy parameters. Experiments show that compared with fixing anatomy parameters, higher acoustic reconstruction quality and a more accurate estimation of the underlying articulatory movements can be obtained in this way, potentially benefit from less compensatory efforts. Besides, the estimated anatomy parameters show clear speaker-specific clusterings, which implies that the network learns to model the

inter-speaker variability directly in the articulatory space and disentangle speaker-related features from linguistic content.

2. Approach

2.1. Articulatory synthesizer

We adopt the VocalTractLab 2.3 (VTL) [28] as the physical articulatory synthesizer. The VTL model takes 19 tract parameters and 11 glottis parameters as input to synthesize utterances. The tract parameters include the positions of hyoid, tongue body center, tongue tip, tongue blade, tongue root, jaw, and lip, and the shape and opening of velum. For the glottis model, we use the recommended ‘‘Geometric Glottis’’. The glottis parameters include F_0 , subglottal pressure, bottom and top displacement of the vocal cord, chink area, relative amplitude, double pulsing, pulse skewness, flutter, and aspiration strength.

Conventionally, a fixed speaker file is required by VTL for speech synthesis, where the vocal tract anatomy is specified. We argue that imitating different speakers’ speech with a fixed vocal tract model would inevitably affect the accuracy of the estimation of articulatory movements, as articulation needs to be adjusted to compensate for the mismatch in vocal tract anatomy. Therefore, it is necessary to adjust the vocal tract anatomy to match different speakers. To this end, we modify the synthesis pipeline of VTL so that we can specify 13 anatomy parameters for each utterance. The definitions and ranges of the anatomy parameters are listed in Table 1. The anatomy parameters describe the physical size and shape of different components of the vocal tract and they are fixed for each utterance, while the tract and glottis parameters are time-varying. Hereinafter, we collectively refer to the tract, glottis and anatomy parameters as articulatory parameters.

Table 1: Anatomy parameters of VTL.

Articulatory dimensions	Range
Lip width (W0)	[0.5, 1.5] cm
Mandible height (H5)	[0.9, 1.8] cm
Lower molars height (H4)	[0.3, 0.7] cm
Upper molars height (H3)	[0.3, 0.7] cm
Palate height (H2)	[0.8, 2.0] cm
Palate depth (D0)	[2.6, 5.0] cm
Hard palate length (W1)	[2.2, 5.2] cm
Soft palate length (W2)	[2.3, 3.1] cm
Pharynx length (H0)	[3.5, 8.0] cm
Larynx length (H1)	[2.0, 4.0] cm
Larynx width (W4)	[2.0, 3.5] cm
Vocal fold length (W3)	[0.5, 2.0] cm
Oral-pharyngeal angle (A0)	[-105.0, -90.0] °

2.2. Embodied Joint Embedding (EmJEm)

We adopt the EmJEm framework [29] to train an articulatory inference network with alternating data sampling steps and model training steps. At the sampling step, the partly trained inference network estimates the underlying articulatory parameters of each training utterance. Then, following [30], articulatory parameters are sampled from a Multivariate Gaussian which is centered at the estimated articulatory parameters and fed to the VTL model to synthesize utterances. The standard deviation σ of the Multivariate Gaussian decays exponentially with the number of iterations at a decay rate γ . Such that we obtain a synthetic dataset of parallel articulatory-acoustic data. At the

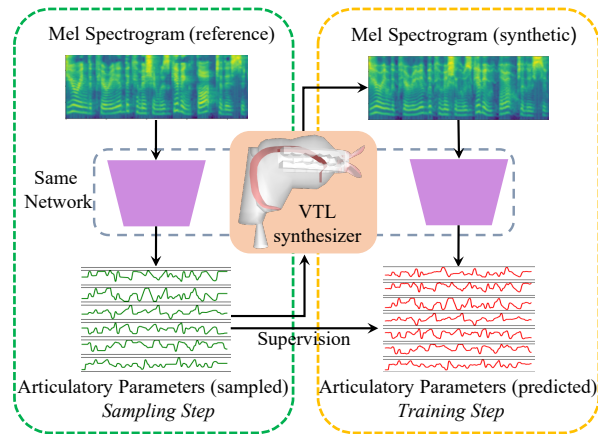


Figure 1: An iteration of the EmJEm framework, consisting of a sampling step and a training step. At the sampling step, we sample from a Multivariate Gaussian which is centered at the articulatory parameters estimated by the inference model.

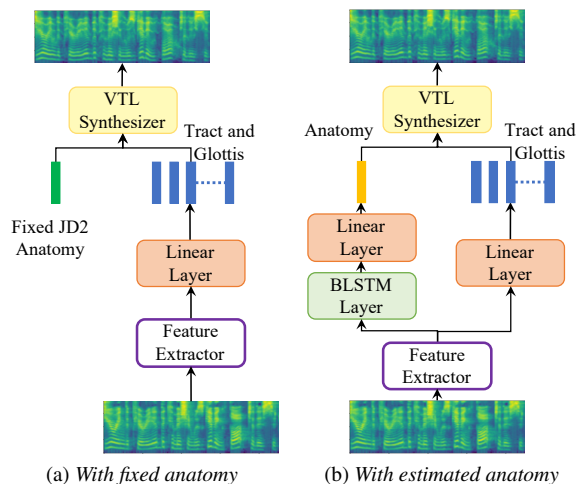


Figure 2: Resynthesize workflow comparison: (a) the anatomy is fixed; (b) the network jointly estimates the articulatory movements (tract and glottis parameters) and the anatomy.

training step, the inference network is further trained with the newly synthesized paired data. Figure 1 illustrate the workflow from the sampling step to the training step during an iteration.

2.3. Fixed anatomy modeling

As a baseline, we train an inference network under the EmJEm framework with the anatomy parameters of VTL fixed. A speaker file of an adult male speaker ‘‘JD2’’ is used, which is given as an example by the authors of the VTL model.

As is illustrated in Figure 2a, the inference network is composed of a feature extractor (details in Section 3.2) and an output layer. The feature extractor extracts features from the mel-spectrogram of a given utterance and the output linear layer maps from the extracted features to 30-dimensional tract and glottis parameters frame-by-frame. Tanh activation is used for the output linear layer to restrict the output within the range of $[-1, 1]$. After rescaling, the time-varying tract and glottis pa-

parameters are then fed the the VTL model to resynthesize the utterance. During the process, the anatomy parameters of the VTL model is specified by the “JD2” speaker file. We call this method as *FIXed Anatomy Modeling (FIX-AM)*.

At the training step, a mean squared error (MSE) loss $L_{tract,glottis}$ is calculated between the sampled and predicted tract and glottis parameters, and the training objective L is

$$L = L_{tract,glottis} \quad (1)$$

2.4. Utterance-level anatomy modeling

To adjust the vocal tract anatomy for different speakers, the inference network need to estimate both the time-varying tract and glottis parameters and the anatomy parameters.

To this end, in addition to the output layer that estimates tract and glottis parameters, a parallel output stream is added to estimate anatomy parameters, as is illustrated in Figure 2b. A single layer of bidirectional LSTM (BLSTM) with a hidden size of 128 is used to gather the information of the entire utterance and its last cell state is then mapped to 13-dimensional anatomy parameters with a linear layer. The anatomy parameters are estimated independently for each utterance, thus we call this method as *Utterance-level Anatomy Modeling (UT-AM)*.

The network is optimized in a multi-task learning manner, and the training objective L is composed of the anatomy parameters prediction loss $L_{anatomy}$ and the tract and glottis parameters prediction loss $L_{tract,glottis}$,

$$L = L_{tract,glottis} + \lambda L_{anatomy} \quad (2)$$

where λ is a hyperparameter. Both of the $L_{anatomy}$ and $L_{tract,glottis}$ are MSE losses.

2.5. Speaker-level anatomy modeling

Utterance-level anatomy modeling makes it possible to adjust the vocal tract anatomy from utterance to utterance. However, it is possible that the anatomy parameters estimated from different utterances of a same speaker differ from each other, which is unreasonable. Moreover, using only a single utterance to estimate the anatomy of the vocal tract could be difficult, especially when considering the compensatory effect of articulation. To obtain speaker-specific and more robust anatomy estimation, we further utilize the speaker labels to unify the anatomy parameters estimated from the same speaker.

To do this, at the sampling step, we firstly estimate the articulatory parameters utterance by utterance, and then for each speaker, we collect the estimated anatomy parameters of that speaker and calculate the averaged anatomy parameters as the final estimation of that iteration. Except for this speaker unification operation, everything else of the model training is the same with *utterance-level anatomy modeling*, and we call this method as *SPeaker-level Anatomy Modeling (SP-AM)*.

3. Experimental setup

3.1. Dataset

We use the Haskins Production Rate Comparison database (HPRC) [31] as the data set. HPRC contains synchronized EMA trajectories and speech data recorded from eight native speakers of American English (4 males and 4 females). Each speaker reads 720 sentences at 2 speech rates. The EMA trajectories were recorded with 8 sensors at 100 Hz and have been filtered with a 20 Hz Butterworth lowpass filter. The database

is randomly split into training, validation, and testing sets, with 10294, 1292 and 1284 samples respectively. We removed the silent segments at the beginnings and endings of the utterances with the provided time stamps. Only the speech data is used for the unsupervised training process, while the EMA data is only used for articulatory evaluations.

3.2. Feature extractor

A feature extractor that extracts features from mel-spectrograms for the output mappings is shared in Section 2.3, 2.4 and 2.5. We adopt a feature extractor proposed in [30] that is a stack of a convolutional neural network (CNN) module and a Conformer [32] module. The CNN module extracts local features from the mel-spectrograms with 2 parallel streams, one is a 7-layer U-net [33] and the other is a stack of 3 1d-convolutional layers. The output dimensions of the U-net and the 1d-convolutional layers are 114 and 30, respectively. The output of the two streams are concatenated to 144-dimensional features and then fed to the Conformer blocks. The Conformer block is powerful in sequence modeling and we stack 3 Conformer blocks with a feature dimension of 144, with 4 attention heads of 36 dimensions.

3.3. Training details

144-dimensional log magnitude mel-spectrograms with 25 ms frame length and 10 ms frame shift are extracted as the acoustic inputs. The inference network is randomly initialized and the loss weight λ is set to 0.1. The network is optimized with the Adam optimizer, with a learning rate of 5×10^{-4} , $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and we clip the gradient norm to 20. The sampling noise σ decays exponentially at a rate $\gamma = 0.98$, with an initial noise of 0.5. We train the inference network for 10 epochs with a batch size of 80 in each iteration.

3.4. Evaluation metrics

We measure the acoustic similarity with mel-cepstrum distortion (MCD) [34]. We extract 25-dimensional mel-cepstral coefficients with Speech Signal Processing Toolkit¹ (SPTK) and ignore the first dimension.

Pitch is an important acoustic clue. We use the root mean squared error of $\log F_0$ ($\log F_0$ RMSE) to measure the similarity of F_0 between the reference and re-synthesized utterances. We extract pitch with SPTK, with a frame shift of 10 ms.

We evaluate the intelligibility of re-synthesized utterances with STOI [35], which scores from 0 to 1. A higher STOI score indicates higher intelligibility.

In addition, we use pearson correlation coefficient (CC) for articulatory evaluation. To do this, we first transform the estimated tract parameters by VTL to match the dimensions of the recorded EMA data. 12 dimensions are obtained, including TTX, TTY, TDY, TDY, TBX, TBY, ULX, ULY, LLX, LLY, LIX and LIY (denoting the horizontal and vertical position of tongue tip, tongue dorsum, tongue body, upper lip, lower lip and lower incisor) and then CC between the measured and the estimated articulatory trajectories are calculated.

4. Results

The averaged testing results of acoustic evaluations and the articulatory evaluations are summarized in Table 2 and Table 3. It can be observed that the introduction of the anatomy modeling, whether utterance-level or speaker-level, can effectively

¹<http://sp-tk.sourceforge.net/>

Table 2: Acoustic evaluation results for method comparison.

Method	MCD	$\log F_0$ RMSE	STOI
<i>FIX-AM</i>	4.79	0.49	0.81
<i>UT-AM</i>	4.70	0.43	0.82
<i>SP-AM</i>	4.63	0.46	0.81

Table 3: Mean correlation coefficient (CC; standard deviation in bracket) of articulators for different modeling methods.

Articulator	<i>FIX-AM</i>	<i>UT-AM</i>	<i>SP-AM</i>
TTX	0.23(0.23)	0.29(0.23)	0.24(0.24)
TTY	0.00(0.23)	0.04(0.25)	0.08(0.23)
TDX	0.40(0.20)	0.46(0.20)	0.42(0.21)
TDY	0.19(0.20)	0.30(0.23)	0.30(0.21)
TBX	0.37(0.19)	0.47(0.18)	0.42(0.19)
TBY	0.23(0.22)	0.17(0.24)	0.19(0.24)
ULX	0.31(0.20)	0.32(0.23)	0.28(0.22)
ULY	0.20(0.23)	0.25(0.22)	0.25(0.22)
LLX	0.37(0.22)	0.41(0.24)	0.40(0.21)
LLY	0.50(0.17)	0.49(0.17)	0.50(0.16)
LIX	0.26(0.25)	0.43(0.23)	0.35(0.22)
LIY	0.36(0.20)	0.43(0.20)	0.41(0.17)
Averaged	0.28(0.25)	0.34(0.26)	0.32(0.24)

improve the model performance. Acoustically, the anatomy modeling enables a better imitation of utterances from different speakers, resulting in lower MCD and $\log F_0$ RMSE. Articulatorily, a better estimation of articulatory trajectories can be obtained by anatomy modeling. When comparing the *utterance-level anatomy modeling* with *fixed anatomy modeling*, the averaged correlation coefficient relatively increases by about 21%. This result supports our hypothesis that imitating different speakers’ speech with a fixed vocal tract model would inevitably affect the accuracy of the estimation of articulatory movements, due to the compensatory effect of articulation. So that it is better to jointly model the vocal tract anatomy and articulatory movements.

The comparison between the *utterance-level anatomy modeling* and *speaker-level anatomy modeling* shows that the unification of the anatomy parameters of each speaker during the sampling step, as is described in Section 2.5, can further reduce the acoustic reconstruction distortion. An example from *speaker-level anatomy modeling* is illustrated in Figure 3. However, such a constraint may also reduce the flexibility of articulation so that the articulatory estimation accuracy is slightly worse.

Since the anatomy parameters are related to speaker characteristics, we check the distribution of the anatomy parameters estimated from the *utterance-level anatomy modeling* and *speaker-level anatomy modeling* and illustrate the results with t-distributed stochastic neighbor embedding (t-SNE) algorithm [36] in Figure 4. From Figure 4a, the anatomy parameters are generally distributed by speakers, although the boundaries between speakers are not that clear. Such a result shows that *utterance-level anatomy modeling* can successfully train a model to distinguish between different speakers, even without the help of speaker labels. As a comparison and shown in Figure 4b, with the speaker unification operation, the anatomy parameters show clear clusterings by different speakers and the diversity of the anatomy parameters of each speaker is small. Such a result demonstrates the effectiveness of *speaker-level*

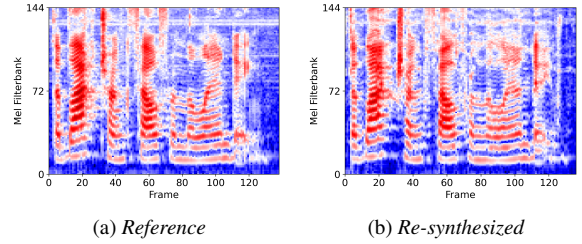


Figure 3: Mel-spectrogram comparison: (a) the reference utterance and (b) the utterance re-synthesized by speaker-level anatomy modeling (SP-AM). The text is “The black trunk fell from the landing.”.

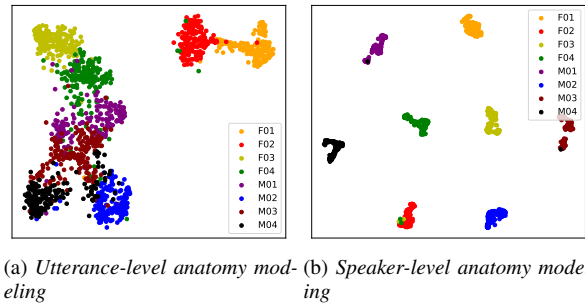


Figure 4: Visualization of anatomy parameters using t-SNE for speakers (represented by different colors) in the testing set of HRPC dataset.

anatomy modeling and it also reflects that speaker-related features is disentangled from linguistic content, so that the estimated anatomy parameters can serve as a speaker representation.

5. Conclusions

In this paper, we propose to jointly estimate articulatory movements and the vocal tract anatomy during the unsupervised inversion of speech. Experiments are conducted under the EM-JEM framework and the results show that the anatomy modeling can effectively reduce the acoustic distortion of speech reconstruction and improve the accuracy of the estimation of articulation, possibly due to the less articulatory compensation. The estimated vocal tract anatomy is shown to be distributed by speakers. And a method of speaker-level anatomy modeling is further proposed to obtain discriminative speaker-specific anatomy, which can be viewed as a speaker representation with explicit physical meanings. Future work includes investigating the correlation between the estimated anatomy parameters with measured ones and exploring the potential of applying the estimated anatomy parameters in areas such as speaker identification.

6. Acknowledgements

The work is supported in part by the National Natural Science Foundation of China (No. 11590773), the Key Program of National Social Science Foundation of China (No. 15ZDB111).

7. References

- [1] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.
- [2] A. Neri, C. Cucchiari, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, vol. 15, no. 5, pp. 441–467, 2002.
- [3] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and blstm-based deep tone models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, 2019.
- [4] 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.
- [5] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proc. 5th Seminar on Speech Production*. Citeseer, 2000, pp. 237–240.
- [6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.
- [7] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *International Conference on Nonlinear Speech Processing*. Springer, 2007, pp. 263–272.
- [8] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.
- [9] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5931–5935.
- [10] 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.
- [11] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019.
- [12] 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*, vol. 24, no. 10, pp. 1865–1875, 2016.
- [13] 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*, vol. 23, no. 12, pp. 2246–2259, 2015.
- [14] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2015.
- [15] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *INTERSPEECH*, 2018, pp. 3122–3126.
- [16] A. Xu, P. Birkholz, and Y. Xu, "Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation," in *Proceedings of The 19th International Congress of Phonetic Sciences*, 2019.
- [17] S. Prom-on, P. Birkholz, and Y. Xu, "Training an articulatory synthesizer with continuous acoustic data," in *INTERSPEECH*, 2013, pp. 349–353.
- [18] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, no. 1, pp. 19–48, 1994.
- [19] H. Shibata, M. Zhang, and T. Shinozaki, "Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 530–537.
- [20] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*. PMLR, 2014, pp. 387–395.
- [21] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
- [22] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.
- [23] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.
- [24] A. Lammert, M. Proctor, A. Katsamanis, and S. Narayanan, "Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact," in *INTERSPEECH*, 2011, pp. 2813–2816.
- [25] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language, and Hearing Research*, vol. 56, pp. 1924–1933, 2013.
- [26] M. Honda, A. Fujino, and T. Kaburagi, "Compensatory responses of articulators to unexpected perturbation of the palate shape," *Journal of Phonetics*, vol. 30, no. 3, pp. 281–302, 2002.
- [27] M. Thibeault, L. Ménard, S. R. Baum, G. Richard, and D. H. McFarland, "Articulatory and acoustic adaptation to palatal perturbation," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2112–2120, 2011.
- [28] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS one*, vol. 8, no. 4, p. e60603, 2013.
- [29] Y. Sun and X. Wu, "Embodied self-supervised learning by coordinated sampling and training," *arXiv preprint arXiv:2006.13350*, 2020.
- [30] Y. Sun, Q. Huang, and X. Wu, "Unsupervised inference of physiologically meaningful articulatory trajectories with vocaltractlab," in *Submitted to INTERSPEECH*, 2022.
- [31] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, 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.
- [32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020, pp. 5036–5040.
- [33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.
- [34] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.
- [35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4214–4217.
- [36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.