TOWARDS SPEAKER-INDEPENDENT ULTRASOUND TONGUE IMAGING-BASED ARTICULATORY-TO-ACOUSTIC MAPPING

XXXXXX

XXXXXXXX

ABSTRACT

In this paper, an articulatory-to-acoustic mapping task is studied, which aims to predict the mel-spectrogram of the audio signals using midsagittal ultrasound tongue images of the vocal tract. Despite of sustainable efforts been made, most previous attempts have been constrained to the speaker-dependent scenario, and the performance is greatly decreased for unseen speakers. Here, a novel approach is proposed for the speaker-independent scenario, using domain-adaptation and adversarial learning. To validate the performance of proposed method, objective evaluation is conducted to demonstrate the effectiveness of the proposed method. The quantitative results indicate our method can achieve superior performance.

Keywords: Ultrasound tongue imaging, articulatory-to-acoustic mapping.

1. INTRODUCTION

In the natural speech production process, the speech signal is highly related to articulation. Understanding the association between the articulation and the speech signals not only can be helpful to improve our understanding of speech production but also can stimulate the theoretical development of speech recognition. There is an increasing trend that aims to build the articulatory-to-acoustic mapping [1] as the applications of the mapping seem to be evident, such as the silent speech interface (SSI) [2] which has a long-term goal to generate the speech using the soundless articulatory information.

Previous attempts employed statistical methods [3] for the inversion between the articulatory movements and speech, while deep learning has begun to dominate this field [4]. For example, [5] explored a two-layer deep neural network for the inversion task, and satisfying performance can be obtained. Despite the sustainable efforts that have been made [6], the mapping performance can be greatly varied for different speakers and most of the previous methods are constrained for the speaker-dependent scenario. In this paper, we present a novel approach towards speaker-independent mapping, which is inspired by the domain adaptation method. Specifically, we explore decoupling the speaker spectral generation task and the speaker recognition task. Leveraging a novel designed loss function, we can improve the performance under the speaker-independent scenarios, through adversarial learning. Our code is available at: https://github.com/xianyi11/Articulatory-to-Acoustic-with-Domain-Adaptation.

2. RELATED WORK

The early studies on articulatory-to-acoustic mapping usually employ low-order spectral representations of speech as the target, such as [7] using 12 coefficients. Later, [8] used other spectral features which has 25-dimensional to handle the articulatory-to-acoustic mapping task. Recently, [9] used a large-scale pre-trained model WaveGlow [10] to predict more detailed representations (80-dimensional mel-spectrogram) from ultrasound tongue images (UTI), resulting in more accurate models. However, to the best of our knowledge, many of previous attempts try to solve the task in a speaker-dependent manner, while the performance is greatly decreased for the speaker independent way (i.e., unseen speaker). There are just a few studies on the speaker-independent articulatory-to-acoustic mapping. Motivation by the shortcomings, we mainly focus on the speaker-independent scenarios.

3. METHODOLOGY

3.1. Overall design

In Figure 1, we present the overall flowchart of the proposed method. Our approach consists of two main parts: the first part is a convolutional neural network for the feature extraction from the UTI,

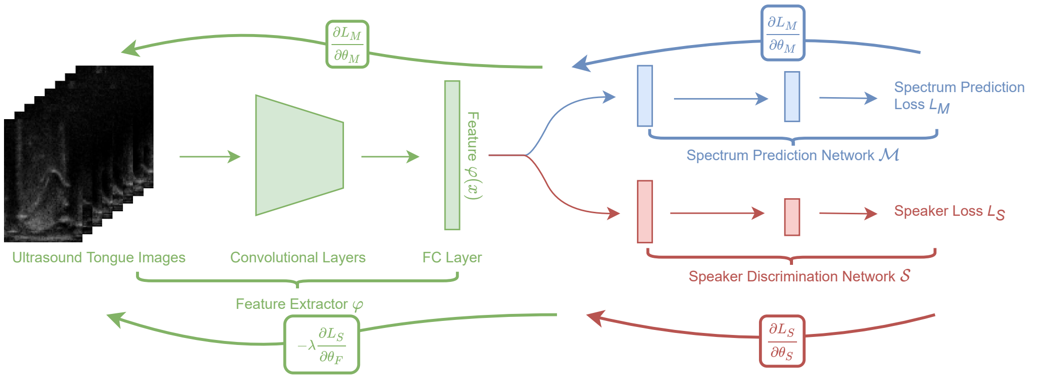
$\frac{\partial L_M}{\partial \theta_M}$
$\frac{\partial L_M}{\partial \theta_M}$
Spectrum Prediction Loss L_M
Spectrum Prediction Network $\mathcal{M}$
Feature $\varphi(x)$
Speaker Loss L_S
Ultrasound Tongue Images
Convolutional Layers
FC Layer
Speaker Discrimination Network $\mathcal{S}$
Feature Extractor φ
$-\lambda\frac{\partial L_S}{\partial \theta_F}$
$\frac{\partial L_S}{\partial \theta_S}$

Figure 1: Overall flowchart of the proposed method for the articulatory-to-acoustic mapping using UTI.

and the latter part is designed for the estimation task. To improve the generalization ability across different speakers, we decouple the latter part into two branches, one branch for the generation task and the other is for the speaker discrimination. For the joint training, a novel designed loss is explored to train the network through adversarial learning. Compared to the method without speaker discrimination and adversarial learning, our method can improve the prediction performance under the speaker-independent scenarios. The components will be explained in more detail subsequently.

3.2. Mel-Spectrogram Prediction

The UTI contains the deformation information of the vocal tract during natural speech production. Intuitively, we could predict the sound from the ultrasound image by the high-dimensional non-linear transformation. Thereby, an ultrasound video could be mapped to the continuous speech spectrogram in Mel-scale through the framewise prediction. Specifically, we build the mapping using a multi-stage deep neural network. The predicted Mel-spectrogram $\hat{y}_i$ could be acquired by:

$$\hat{y}_i = u_i^T \varphi(x). \tag{1}$$

where φ is the backbone neural network which is parameterized by θ_φ, and $x \in R^{H\times W}$ is the ultrasound tongue image. The weights of spectrogram prediction network $\mathscr{M}$ can be noted as u, where u_i is the i_{th} weight vector. The loss function applied for this generation task is the mean-squared error (MSE):

$$\mathscr{L}_M = \frac{1}{M}\sum_{i=0}^{M}(y_i - \hat{y}_i)^2, \tag{2}$$

where M is the number of Mel-bins in the Mel-spectrogram, y is the ground-truth of the spectrogram.

3.3. Speaker Discrimination

The speaker information and the speech content are naturally coupled in the ultrasound image and spectrogram. In order to make our system irrelevant or weakly dependent to the speaker information, it is necessary to decouple the speaker information and the speech content. We set up a shallow speaker discrimination network $\mathscr{S}$ which is parameterized by $\theta_{\mathscr{S}}$ to recognize the speaker from the ultrasound image:

$$\hat{s}_i = v_i^T \varphi(x). \tag{3}$$

where the weights of network $\mathscr{S}$ are denoted as v, and v_i is the i-th weight vector. Note that, during the update of network $\mathscr{S}$, the weights of network $\mathscr{M}$ are fixed, so as to avoid the impact of the cross-modal mapping process. The loss function for learning speaker information is the Cross-Entropy loss:

$$\mathscr{L}_S = -\log\frac{e^{s_i}}{\sum_j^K e^{s_i}}, \tag{4}$$

where i is the speaker index, and K is the number of distinguishable speakers which will be explained in detail in Sec.4.

3.4. Adversarial Training

After the network $\mathscr{S}$ converges, the feature $\varphi(x)$ output from the backbone network φ is speaker-dependent, because after the shallow non-linear transformation, the speaker could be recognized. We hope the feature $\varphi(x)$ has better generalization ability across different speakers, rather than close to the target speaker. We implement this in an adversarial way by decreasing the $-\mathscr{L}_S$ in the speaker discrimination learning process. The final loss for spectral feature prediction is calculated by:

(5) $\mathscr{L} = \mathscr{L}_M - \lambda \mathscr{L}_S,$

where λ is the empirical weight which will be investigated in Sec.4.3. The parameters $\theta_{\mathscr{S}}$ of the speaker discrimination network is frozen in the adversarial training.

4. EXPERIMENTAL RESULTS

4.1. Dataset and Implementation Details

The Ultrasuite [11] dataset is used throughout the experiments. These children suffer from speech sound disorders (SSD) to varying degrees. UXTD is the typically developing subset of the Ultrax dataset which contains 58 individuals (31 females and 27 males). The ultrasound framerate is 121.5 FPS. We down-sample the frames by 10 (for every second, we sample 12 frames due to the constraint of the computation resources). For the backbone neural network, we employ ResNet50, which is a well known CNN architectures for the feature extraction. Behind the last convolutional layer, we explore the Batch Norm Fully Connected (BN-FC) [12] structure to get the final 1024-dimension embedding feature. The spectrogram prediction network $\mathscr{M}$ and speaker discrimination network $\mathscr{S}$ are both BN-FC structures. The output dimensions are the number of Mel-bins (80) and the number of speakers. The networks are trained by a stochastic descent optimizer (SGD) with the batch size of 512.

In the experiment, The 66% samples (about 220,468 images) are used for training, and the rest are used for testing. Both the train set and test set contain the data of all 58 individuals (*Base* in table 1). To test the performance of our system on speaker independent data, the partition of the train set and test set is speaker-based. All the data sampled from 38 individuals among all 58 individuals are used for training, and the rest 20 speakersâ data is used for testing (*Sep* in table 1)

4.2. Compared methods

As most of previous attempts focus the mapping using the speaker-dependent manner, we compare the proposed approach with two baselines, *i.e.*, Source-Only model and ST-Adversarial model. Specifically, the Source-Only model is trained without domain-adaptation (no speaker discrimination branch are incorporated into the framework in Fig. 1). This model achieves the articulatory-to-acoustic mapping by the neural network without considering the speaker information. The ST-Adversarial model employs the source and target domain adversarial training inspired by [13]. Specifically, the recognition task is to identify whether the feature belongs to the source or target domain. In our setting, the train set and test set can be regarded as the source and target domain respectively, and the speaker discrimination loss $\mathscr{L}_S$ optimizes a binary classification task ($K = 2$ for Eq.4). In contrast to these, in the proposed ID-Adversarial model, the recognition task is to identify the speaker accurately, and K is the number of speakers during the training for Eq.4.

Table 1: Comparison results in term of mean MSE (Lower is better), SSIM and CW-SSIM (Both metrics ranges between 0 and 1. Higher value denotes better performance, while 1 represents the predicted one is the same as the ground-truth).

		MSE	SSIM	CW-SSIM
Source-Only	Base	1.88	0.73	0.70
	Seq	1.88	0.74	0.73
ST-Adversarial	Base	1.84	0.71	0.68
	Seq	1.78	0.75	0.72
ID-Adversarial	Seq	**1.62**	**0.76**	**0.74**

4.3. Objective Evaluation

Quantitative evaluation is conducted to demonstrate the effectiveness of the proposed method. Three evaluation metrics are used. 1) MSE; 2) Structural Similarity Index (SSIM) between the reconstructed spectrogram and the original spectrogram. 3) Complex Wavelet Structural Similarity Index (CW-SSIM) which is robust to small rotations and translations compared to SSIM. The evaluation results are shown in Table 1. The results show that the adversarial model performs better than the Source-Only model with MSE metric. Note that, the ST-Adversarial model performs even worse than the Source-Only model under the SSIM and CW-SSIM, which indicates that source-target adversarial training may not improve the structural similarity. While the ID-adversarial model is effective in various evaluation metrics.

We analyze the impact of the hyper-parameters λ in Eq.5. The optimal value of λ is about 0.4 to 0.5. We use 0.4 in our following experiments. It is worthwhile noting that λ should be increased cautiously. In order to assign more weights to the speaker discrimination network automatically, the model ignores the main prediction task, which

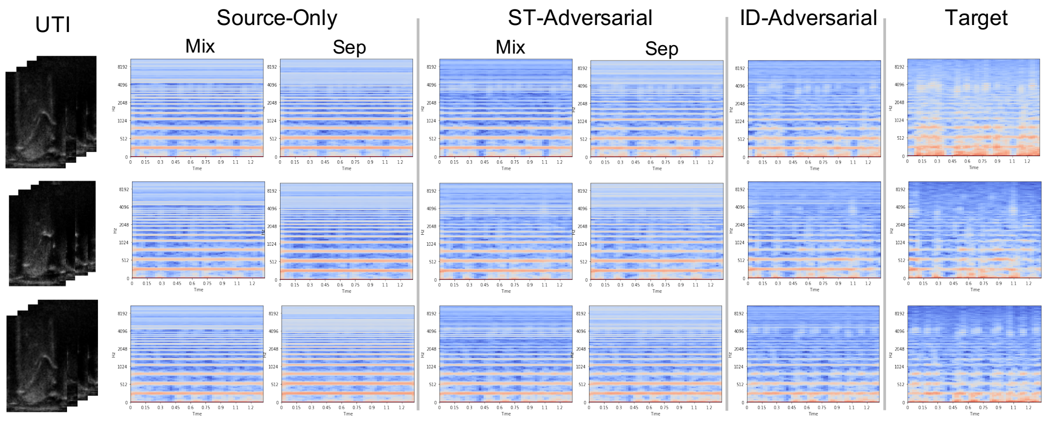

Figure 2: The illustration of generated spectrogram by different models for three different speakers.

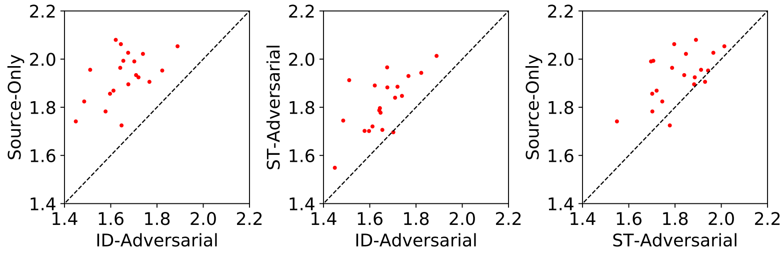

Figure 3: The MSE for 20 speakers with different models. In the middle subplot, all the points are above the diagonal, which indicating that ID-Adversarial model is better than Source-Only model for all the 20 speakers.

makes the performance deteriorate rapidly.

4.4. Qualitative Analyses

We also performed a qualitative analysis to investigate the quality of articulatory-to-acoustic mapping. The generated spectrogram and corresponding target are shown in Fig 2 from three different speakers. We could see that, without domain-adaptation (Source-Only), the spectrogram generated by different speakers are relatively similar. The spectrogram shows multiple horizontal lines along with different frequencies and lacks vocal details. In comparison, the vocal details presented by the ID-Adversarial model are richer.

The formant in the spectrogram is more obvious in the ID-Adversarial model. The formant represents a set of adjacent harmonics which are boosted by resonance in some part of the vocal tract. Obvious formant means that in the subsequent process of generating speech from the spectrogram, the generated speech is clearer. Especially for three different speakers, generated spectrograms are different for the ID-adversarial model.

Fig 3 exhibits the improvement of domain-adaptation approach. It shows the MSE for 20 speakers with different models in the *Sep* setting. In the left subplot, all the points are above the diagonal, which is indicating that the ID-Adversarial model is better than the Source-Only model for all the 20 speakers. We could see that the domain-adaptation approach improves the performance in the speaker-independent scenario.

5. CONCLUSION

This paper proposed a method towards speaker-independent articulatory-to-acoustic mapping, using UTI. Specifically, the domain adaption and adversarial method are applied in our framework, which can decouple the generation and speaker discrimination task. To demonstrate the effectiveness of the proposed method, extensive experiments are conducted. Objective evaluation is conducted to compare the generated spectrograms and ground truth, using three evaluation metrics. The results indicate that our proposed method can achieve superior performance under the speaker-independent scenario. In the future, we plan to conduct subjective listening tests to evaluate the quality of speaker-independent samples.

6. REFERENCES

[1] B. Cao, M. J. Kim, J. R. Wang, J. P. van Santen, T. Mau, and J. Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information." in *INTERSPEECH*, 2018, pp. 3152–3156.

[2] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[3] 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.

[4] H. Li, J. Tao, M. Yang, and B. Liu, "Estimate articulatory mri series from acoustic signal using deep architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4854–4858.

[5] 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)*. IEEE, 2019, pp. 1–8.

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

[7] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "Dnn-based ultrasound-to-speech conversion for a silent speech interface," 2017.

[8] T. G. Csapó, M. S. Al-Radhi, G. Németh, G. Gosztolya, T. Grósz, L. Tóth, and A. Markó, "Ultrasound-based silent speech interface built on a continuous vocoder," *arXiv preprint arXiv:1906.09885*, 2019.

[9] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis," *arXiv preprint arXiv:2008.03152*, 2020.

[10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," 2018.

[11] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," *InterSpeech*, 2018.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.