Unsupervised Inference of Physiologically Meaningful Articulatory Trajectories with VocalTractLab

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

Recently, the introduction of reinforcement learning methods and the Embodied Joint Embedding (EmJEm) approach has made it feasible to unsupervisedly infer articulatory movements from arbitrary utterances. However, the quality of re-synthesized utterances is still unsatisfactory and there is a lack of direct evaluation of the inferred articulatory movements to see if they are physiologically meaningful. In this work, we extend the EmJEm approach to tackle these problems of unsupervised acoustic-to-articulatory inversion (AAI). The VocalTractLab is adopted as the articulatory synthesizer and a novel architecture of the articulatory inference network is proposed. To obtain physiologically meaningful articulatory trajectories, a smoothness constraint is introduced as an articulatory prior during training. Experiments show that the proposed approach is able to re-synthesize utterances with state-of-the-art quality while effectively smooth the articulatory trajectories. We directly compare the unsupervisedly obtained articulatory trajectories with the recorded articulatory data from the HPRC database and it turns out that the inferred articulatory trajectories have a relatively high correlation with the recorded trajectories. This encouraging result shows the practical potential of unsupervised AAI methods.

Index Terms: acoustic-to-articulatory inversion

1. Introduction

Acoustic-to-articulatory inversion (AAI) aims to infer the movements of articulators, e.g., tongue and lips, from acoustic signals. Information of articulatory movements is useful for phonetics studies and benefits the development of speech technologies like text-to-speech (TTS) and automatic speech recognition (ASR), especially in low resource cases [1]. Articulatory information also play a key role in computer assisted pronunciation training systems for second language learning [2].

Supervised AAI approaches train statistical models with parallel acoustic-articulatory data. Several methods have been proposed, including codebook approaches [3], Kalman filtering [4] and Hidden Markov Model (HMM) [5]. In recent years, deep learning methods have achieved significant success on the AAI task [6, 7, 8, 9]. Despite that, mainly due to the limited size of articulatory datasets and the significant variability among speakers, the generalization performance of supervised approaches is still unsatisfactory [10].

Unsupervised AAI methods generally works in an analysis-by-synthesis manner to control an articulatory synthesizer to reproduce a target utterance. To this end, several optimization methods, such as random search [11], distal learning [12] and genetic algorithms [13], have been proposed. However, these methods are typically time consuming so that they only suit the inversion of short speech segments, e.g., a word or a syllable. Recently, several unsupervised AAI methods have been developed for arbitrarily long utterances. Shibata et al. [14] built a hybrid auto-encoder with an articulatory inference neural network and the VocalTractLab (VTL) [15] synthesizer. Since the articulatory synthesizer is non-differentiable, they trained the inference network with deterministic policy gradients (DPG) [16]. Sun and Wu [17] proposed an Embodied Joint Embedding (EmJEm) framework to train an articulatory inference network with alternating sampling and training steps. For these methods, once trained, an articulatory inference network needs a single time-efficient feed-forward pass to do inference. However, the reconstruction quality of these methods still has room for improvement and more importantly, these reported works are short of direct articulatory evaluations.

In this work, we extend the EmJEm approach in the following aspects to realize high reconstruction quality and obtain physiologically meaningful articulatory parameters:

- To realize high quality reconstruction, the VTL model is adopted as the articulatory synthesizer and a novel architecture of the inference network is proposed.
- We propose to sample articulatory parameters from a Multivariate Gaussian at the sampling step of EmJEm so that the common mode collapse problem can be avoided.
- A smoothness prior of articulation is introduced to the articulatory parameters sampling process so that the inferred articulatory trajectories can be smoothed while the reconstruction quality can be maintained.
- The inferred articulatory parameters are transformed to match the dimensions of electromagnetic articulography (EMA) measurements and then directly articulatory evaluations are conducted. The qualitative and quantitative evaluation results reflect that the inferred articulatory parameters are physiologically meaningful.

2. Approach

2.1. Articulatory synthesizer

The VocalTractLab 2.3 (VTL) [15] uses a 3D vocal tract model to simulate the human vocalization process. It takes 19 tract parameters and 11 glottis parameters as input to synthesize utterances. In addition, a speaker file is required where the morphology of the modeled vocal tract is specified. However, according to [18], jointly estimating the vocal tract morphology and the articulatory movements can improve the speech reconstruction quality and make the estimation of articulation more accurate. Following [18], we modify the synthesis pipeline of the VTL to specify 13 anatomy parameters for each utterance.
The anatomy parameters include the physical size of lip, mandible, molars, hard palate, soft palate, pharynx, larynx and the vocal fold. The tract parameters describe the time-varying positions of the articulators such as the hyoid, jaw, lip, tongue and the state of the velum during articulation. The “Geometric Glottis” model is chosen to model the glottis, whose parameters include $F_0$, subglottal pressure, displacement of the vocal cord, chink area, flutter and aspiration strength and so on. More details of the parameters can be found in [15]. During our experiments, the anatomy parameters are fixed for each utterance, while the tract and glottis parameters are time-varying.

### 2.2. Training framework: EmJEm

We train an articulatory inference network under the EmJEm framework [17] which works in an iterative manner. Each iteration can be split into two steps: a sampling step and a training step. At the sampling step, for each given reference utterance, the partly trained inference network is used to estimate the underlying articulatory parameters which can be viewed as the center of a posterior distribution. Then articulatory parameters are sampled from the posterior and fed to an articulatory synthesizer. In this way, a dataset of articulatory-acoustic data is obtained, where each paired data has physical mapping relationship coming from the articulatory synthesizer. Then, at the training step, the synthesized data is used to continue the training of the inference network. The sampling and training step of an iteration is illustrated in Figure 1.

### 2.3. Sampling strategy

At the sampling step, the posterior distribution is modeled with a Dirac-Delta distribution in [17], so that the deterministic outputs of the network are directly adopted as articulatory samples. However, we noticed mode collapse in our experiments with this sampling strategy. That is, the inference network can cheat by always predicting similar articulatory parameters regardless of the input utterances, such that the training loss diminishes to zero and the optimization process ends. To overcome this problem, we choose to sample articulatory parameters $z$ from a Multivariate Gaussian $\mathcal{N}(\mu; R(\mu), \sigma^2 I)$, which is centered at the estimated articulatory parameters $R(\mu)$, where $R$ is the inference network and $\mu$ is a given utterance. The noise from the sampling process guarantees that there is always gradients to optimize the network. To balance the exploration of the articulatory space and exploitation of the partly trained model, the sampling noise $\sigma$ exponentially decays with iterations at a decay rate $\gamma$.

### 2.4. Model architecture

As is illustrated in Figure 2, we propose an architecture for the articulatory inference network which is composed of a novel feature extractor and two output streams. The feature extractor is composed of a convolutional neural network (CNN) based local feature extraction module and a Conformer [19] based sequence modeling module. For the CNN module, a U-net [20] extracts local features at different scales and 3 layers of 1d-convolutional modules (composed of a pointwise convolutional layer and a depthwise convolutional layer) are used to gather information from all feature dimensions. The outputs of the U-net and the 1d-convolutional layers are mapped to 114 dimensions and 30 dimensions of features, respectively. Then the features are concatenated and fed to a stack of 3 Conformer blocks with a feature dimension of 144, with 4 attention heads of 36 dimensions. Following the feature extractor are two output streams: one is a single layer of bidirectional LSTM (BLSTM) with the hidden size of 128 whose last cell state is mapped to 13-dimensional anatomy parameters. The other is a linear layer with a size of $144 \times 30$ mapping the output of the Conformer blocks to 30-dimensional tract and glottis parameters frame by frame. Tanh activation is used for the output linear layers to restrict the output within the range of [-1, 1]. Hereinafter, we refer to the proposed architecture as MIX-net.

Following [18], the weighted sum of the anatomy estimation loss $L_{\text{anatomy}}$ and the tract and glottis estimation loss $L_{\text{tract,glottis}}$ is used as the training objective $L$.

$$L = L_{\text{tract,glottis}} + \lambda L_{\text{anatomy}}$$  \hspace{1cm} (1)

where $\lambda$ is the weighting hyperparameter. $L_{\text{anatomy}}$ and $L_{\text{tract,glottis}}$ are both mean squared errors (MSE).
2.5. Smoothness prior

Subject to physical constraints, the movement trajectories of the articulators are generally smooth. It has been revealed that the energy spectrum of the articulatory trajectories is lowpass [21]. And various studies have shown that smoothing the articulatory trajectories can reduce the non-uniqueness problem and benefit the inversion performance [22]. We sample the articulatory parameters from a Multivariate Gaussian so that the trajectories are actually noisy and not smooth. To produce smooth articulatory movement trajectories, we perform lowpass filtering on the sampled time-varying tract parameters before they are fed to the VTL synthesizer. In this way, we introduce the smoothness prior to the training process. Once trained, the inference network can predict smooth trajectories without post-filtering.

3. Experimental setup

3.1. Dataset

We conduct experiments on the Haskins Production Rate Comparison database (HPRC) [23] which contains EMA data and the synchronized utterances from eight native speakers of American English (4 males and 4 females). Each speaker reads 720 sentences with 2 speech rates. We randomly split the data into 3 sets: 10294 samples for training, 1292 samples for validation, and 1284 samples for testing. Silent segments at the beginnings and endings of the utterances were removed. During the unsupervised training process, only the speech data is used, while the EMA data is only used for articulatory evaluations.

3.2. Training details

As the acoustic features, we extract 144-dimensional log magnitude mel-spectrograms with 25 ms frame length and 10 ms frame shift. The inference network is randomly initialized and trained with the cost weight $\lambda = 0.1$. We use Adam optimizer with learning rate of $5 \times 10^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and the gradient norm was clipped to 20. The sampling noise $\sigma$ decays from 0.5 at a decay rate $\gamma = 0.98$. During each iteration, we train the inference network for 10 epochs with a batch size of 80. Following [17], we maintain a training set which contains historically sampled data: during each iteration, we drop the samples in the training set with a fix rate of 0.33, and then update the training set by taking in the newly sampled data.

3.3. Evaluation metrics

For acoustic evaluation, Mel-cepstrum distortion (MCD) [24] is used to measure the acoustic similarity between the reference and re-synthesized utterances. 25-dimensional mel-cepstral coefficients are extracted with Speech Signal Processing Toolkit1 (SPTK) and the first dimension is ignored. The root mean squared error of $\log F_0$ ($\log F_0$ RMSE) [25] is used to measure the $F_0$ difference between the reference and re-synthesized utterances. We extract $F_0$ with SPTK, with a frame shift of 10 ms. Besides, STOI [26] is used to evaluate the intelligibility of the re-synthesized utterances, with a score range of [0, 1]. The higher the STOI score, the higher the intelligibility.

For articulatory evaluation, we try to directly compare the estimated articulatory trajectories with the recorded EMA data from the HPRC database. To this end, we firstly transform the estimated trajectories into the dimensions of the EMA data with VTL, and then calculate Correlation Coefficient (CC) for quantitative evaluation. The dimensions used in evaluation include the horizontal (X) and vertical (Y) positions of six articulators, namely tongue tip (TT), tongue dorsum (TD), tongue body (TB), upper lip (UL), lower lip (LL) and lower incisor (LI).

4. Results

4.1. Inversion without smoothness prior

Methods for Comparison We compare the performance of the proposed approach which we call EmJEm + VTL_MIX with existing unsupervised AAI methods that are suitable for arbitrary utterances, which include 1) DPG [14], a reinforcement learning based method. Following Shibata et al. [14], we adopt the VTL synthesizer and implement the actor (inference network) and the critic (synthesis network) with five and three layers of BLSTM, respectively, with 256 hidden units in each direction. The actor takes acoustic features as input and has two linear output layers, one maps the BLSTM output to the tract and glottis parameters with size 512 × 30, the other maps the last cell state of the BLSTM to anatomy parameters with size 512 × 13. Correspondingly, the critic has two linear input layers, one embeds the tract and glottis parameters with size 30 × 226 and the other embeds the anatomy parameters with size 13 × 30. The embeddings are concatenated before fed into the BLSTM network. The MSE between the reference and the re-synthesized acoustic features is adopted as the reward. The actor and the critic networks are optimized alternately with deterministic policy gradients; 2) EmJEm + TRM_Unet [17], an EmJEm based unsupervised AAI method that adopts the Tube Resonance Model (TRM) [27] as the synthesizer and uses an inference network composed of a U-net and a single-layer BLSTM; 3) EmJEm + VTL_Unet, we conduct ablation experiments to demonstrate the effectiveness of the proposed MIX-net architecture by replacing the MIX-net of the proposed approach with the inference network of EmJEm + TRM_Unet. For all the methods, we use the mel-spectrograms described in Section 3.2 as acoustic features.

Table 1: Acoustic and articulatory evaluation results for method comparison. $\log F_0$ denotes $\log F_0$ RMSE.

<table>
<thead>
<tr>
<th>Approach</th>
<th>MCD</th>
<th>$\log F_0$</th>
<th>STOI</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPG</td>
<td>4.79</td>
<td>1.38</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>EmJEm + TRM_Unet</td>
<td>4.96</td>
<td>0.61</td>
<td>0.77</td>
<td>-</td>
</tr>
<tr>
<td>EmJEm + VTL_MIX</td>
<td>4.70</td>
<td>0.43</td>
<td>0.82</td>
<td>0.34</td>
</tr>
<tr>
<td>EmJEm + VTL_Unet</td>
<td>5.29</td>
<td>0.42</td>
<td>0.76</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Acoustic evaluations are conducted for all the four methods, and for DPG, EmJEm + VTL_MIX and EmJEm + VTL_Unet, articulatory evaluations are conducted. Table 1 summarizes the averaged evaluation results from the testing set. It could be observed that the proposed method outperforms the EmJEm + TRM_Unet method in terms of all the acoustic metrics. The DPG method performs well in MCD and STOI, while the pitch reconstruction is not satisfactory. And more importantly, the articulatory correlation of the DPG method is far worse than that of the proposed method. The comparison between the EmJEm + VTL_MIX and EmJEm + VTL_Unet demonstrates that the inference architecture has become the bottleneck of the acoustic reconstruction performance, and the proposed MIX-net architecture can effectively alleviate this problem. To visualize the performance of the proposed method, we illustrate a pair of reference and re-synthesized mel-spectrograms in Figure 3. As demonstrated by the example, the proposed method can faith-

1http://sp-tk.sourceforge.net/
4.2. Inversion with smoothness prior

4.2.1. Empirical analysis of EMA trajectories

We conduct empirical frequency analysis to reveal the lowpass nature of the articulatory parameters and to choose an suitable lowpass cutoff frequency for trajectories smoothing. Following Ghosh and Narayanan [21], for each of the chosen 12 dimensions of EMA trajectories, we firstly perform discrete Fourier transform (DFT) with a DFT order of 16384; and then calculate the cutoff frequency \( f_c \) below which a certain percentage (say \( \alpha \)) of the total energy of the spectrum is contained. It turns out that \( f_c \) is around 5Hz for \( \alpha = 95\% \) and \( f_c \) is around 10Hz for \( \alpha = 99\% \), which confirms the low-pass nature of articulation.

4.2.2. Smoothing with different strength

As is described in Section 2.5, we perform lowpass filtering on the sampled time-varying tract parameters with a cutoff frequency of 5 Hz or 10 Hz before they are fed to the VTL synthesizer. The filtering operation is conducted only during training.

Table 2: Acoustic and articulatory evaluation results at different cutoff frequencies. \( \log F_0 \) denotes \( \log F_0 \) RMSE.

<table>
<thead>
<tr>
<th>Cutoff Frequency</th>
<th>MCD</th>
<th>( \log F_0 )</th>
<th>STOI</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o cutoff</td>
<td>4.70</td>
<td>0.43</td>
<td>0.82</td>
<td>0.34</td>
</tr>
<tr>
<td>5 Hz</td>
<td>5.71</td>
<td>0.61</td>
<td>0.74</td>
<td>0.24</td>
</tr>
<tr>
<td>10 Hz</td>
<td>4.96</td>
<td>0.49</td>
<td>0.81</td>
<td>0.36</td>
</tr>
</tbody>
</table>

The acoustic and articulatory evaluation results are shown in Table 2. It can be seen that limiting the frequency of articulatory movements has more or less negative impact on the speech reconstruction quality. With 5 Hz cutoff frequency, the acoustic and intelligibility drop significantly and it can be heard from the re-synthesized utterances that some consonants are not articulated clearly. The results demonstrate that certain phonemes, especially the consonants, require rapid movements of the articulators. As a comparison, smoothing with 10 Hz cutoff frequency does not severely harm the speech reconstruction quality, which is consistent with results of the empirical analysis. As for the articulatory evaluation, 10 Hz lowpass filtering helps improve CC, while 5 Hz lowpass filtering is strong enough to have negative effects on the articulatory inference.

To provide a better insight into the estimated articulatory parameters and show the effectiveness of smoothing, we compare the estimated articulatory parameters under different smoothing settings with the measured articulatory trajectories in Figure 4. It is clear to see that, compared with the no-smoothing (w/o cutoff) experiment, the smoothness operation effectively reduces the high frequency components of the inferred articulatory parameters. And the trends of the unsupervisedly inferred articulatory trajectories and the measured trajectories are generally consistent. Besides, considering the fact that those trajectories can faithfully reconstruct the reference utterance, the diversity among the estimated trajectories indeed reflects the non-uniqueness nature of the AAI problem. Moreover, it can be observed that at several frames (e.g., the 28th and the 150th frame), the trajectory trends are in high consistency, potentially indicating that those specific trends (movements) are essential for the articulation of specific phonemes.

5. Conclusions

In this work, we have extended the EmJEm approach to unsupervisedly estimate physiologically meaningful articulatory trajectories. The VTL is adopted as the articulatory synthesizer and a novel articulatory inference architecture has been proposed. Experiments show that the proposed approach is able to re-synthesize utterances with state-of-the-art quality. What’s more, a smoothness constraint is introduced as an articulatory prior. The qualitative and quantitative results demonstrate that the inferred trajectories are physiologically meaningful. For future work, we will continue to improve the reconstruction quality and develop more flexible smoothing methods.

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


