COMPLEXITY-PERFORMANCE TRADE-OFF IN ACOUSTIC-TO-ARTICULATORY INVERSION

Aravind Illa and Prasanta Kumar Ghosh

Electrical Engineering Department, Indian Institute of Science, Bangalore

ABSTRACT

Estimating articulatory motion from speech acoustics is known as acoustic-to-articulatory inversion (AAI). The knowledge of position information of articulators along with the acoustics have shown to benefit in various applications like speech recognition, speech synthesis, accent conversion etc. Various methods have been proposed in the literature for AAI, namely codebook based, Gaussian Mixture Model (GMM) [1], Deep Neural Networks (DNN) [2] and Bidirectional Long Short Term Memory (BLSTM) network architecture [3, 4]. Among all these approaches, BLSTM has shown to perform well and achieved state-of-art performance. Although BLSTM model has shown to perform well, there is no systematic comparison among these models with respect to their complexities.

In this work, we systematically compare AAI performance across different models, namely GMM, DNN, convolution neural network (CNN) and BLSTM. Articulatory movements are known to vary slowly in nature, in order to preserve these characteristics in the predicted articulatory trajectories, these are further post-processed using different techniques like low-pass filtering, Kalman filtering and maximum likelihood parameter generation (MLPG). We also compare these post-processing techniques, since to the best of our knowledge no comparison has been made before on these post-processing techniques.

To carry out experiments, 460 MOCHA TIMIT sentences were chosen as speech stimuli to record acoustic-articulatory data using EMA AG501[1]. Six sensors were glued to speech articulators namely, upper lip (UL), lower lip (LL), jaw (Jaw), tongue tip (TT), tongue body (TB) and tongue dorsum (TD). Two more sensors were glued behind the ears for head movement correction. We considered articulatory movements in the horizontal and vertical direction in the midsagittal plane, which results in 12 dim articulatory features. Acoustic-articulatory data from a total of 20 subjects was recorded, out of which 10 were male and 10 were female. All the subjects were fluent speakers of English with no record of speech disorders in the past and from an age group of 21-28 years. We performed manual annotations for the recorded acoustic-articulatory data to remove start and end silence segments in each sentence. Further for every sentence, at each dimension of the articulatory feature we perform mean and variance normalization. As an acoustic feature, we computed 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) vector for every 20ms with a shift of 10ms.

With the 20 subjects' acoustic-articulatory data we performed AAI experiments in a subject dependent manner. For each subject, the recorded 460 utterances of acoustic-articulatory data were divided into three sets for: train 80% (364), validation 10% (46) and test 10% (46). As input acoustic features, we considered 13-dim MFCC along with the delta and delta-delta coefficients. The target variables were 12-dimensional articulatory features along

[1] http://www.articulograph.de/

with their velocity and acceleration coefficients. For the GMM based AAI model, the GMMs were trained with full co-variance matrix with 64, 40 and 32 mixtures components. In DNN, we chose 3-hidden layers with last layer as linear regression layer. Experiments were performed with three different choices of model parameters by varying number of hidden units as 512, 256 and 126. In CNN, we chose 3-hidden layers as DNN but we replaced fully connected layers with 1-d convolution filters of length 5, and number of filters in each layer was varied from 256, 128 and 64, which were followed by a linear regression layer. In BLSTM, we chose 3-hidden layers by varying the number of hidden units in each layer as 128, 64 and 32 followed by a linear regression layer.

Table 1. Comparison of AAI post-processing techniques in terms of average CC.

	Direct	Kalman	LPF	MLPG
GMM	0.7047	0.7799	0.7862	**0.8297**
DNN	0.7533	0.7862	0.7897	**0.7932**
CNN	0.816	0.8268	0.8274	**0.8405**

For GMM, DNN and CNN based articulatory movement predictions we performed post-processing using three techniques, namely Kalman filtering, lowpass filtering and MLPG. To assess the performance, we report average pearson correlation coefficient (CC) between the predicted and original articulatory trajectories. Table 1, reports the performance of AAI with and without post-processing techniques. We observe that all the techniques shows improvements compared to predictions without any post-processing. Among the three post-processing methods, we observe that MLPG performs better than Kalman and Low-pass filtering. We do not observe any benefit of using post-processing techniques for BLSTM, since the predicted articulatory movements preserve smoothness characteristics due to the network architecture.

Fig. 1 reports AAI performance (in y-aixs) with varying complexity, i.e., the number of learnable parameters (in x-axis) across different models. Fig. 1 shows the performance of BLSTM without any post-processing and, GMM, DNN and CNN predictions with and without MLPG. '$\circ$' with model labels in black color denotes the performance of AAI without any post-processing and '$*$' with blue color model labels denotes the performance with MLPG. It is interesting to observe that even when the number of parameters is reduced, the BLSTM performance does not drop drastically. Among all the AAI models, BLSTM yeids the best complexity performance trade-off, which are followed by CNN, GMM and DNN with MLPG. Interestingly, we observe that with MLPG post-processing GMM outperforms the DNN performance.

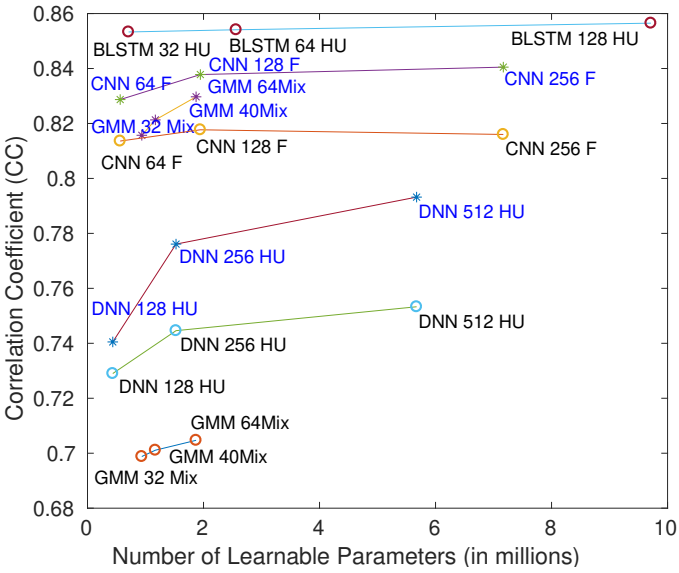

Fig. 1. AAI performance vs model complexity

REFERENCES

[1] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[2] Korin Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping.," in *Proceedings of the ICSLP, Pittsburgh*, 2006, pp. 577–580.

[3] Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, and Lianhong Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.

[4] Aravind Illa and Prasanta Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," *Proc. Interspeech*, pp. 3122–3126, 2018.