DNN-BASED SPEAKER-ADAPTIVE POSTFILTERING WITH LIMITED ADAPTATION DATA FOR STATISTICAL SPEECH SYNTHESIS SYSTEMS

Miraç Göksu Öztürk[1], Okan Ulusoy[1], Cenk Demiroğlu[2]
Bogazici University[1], Ozyegin University[2], Istanbul, TURKEY

Abstract

Deep neural networks (DNNs) have been successfully deployed for acoustic modelling in statistical parametric speech synthesis (SPSS) systems. Moreover, DNN-based postfilters (PF) have also been shown to outperform conventional postfilters that are widely used in SPSS systems for increasing the quality of synthesized speech. However, existing DNN-based postfilters are trained with speaker-dependent databases. Given that SPSS systems can rapidly adapt to new speakers from generic models, there is a need for DNN-based postfilters that can adapt to new speakers with minimal adaptation data. Here, we compare DNN-, RNN-, and CNN-based postfilters together with adversarial (GAN) training and cluster-based initialization (CI) for rapid adaptation. Results indicate that the feedforward (FF) DNN, together with GAN and CI, significantly outperforms the other recently proposed postfilters.

Introduction

- SPSS methods are typically speaker-dependent and generally generate muffled audio samples.
- Postfiltering is one way of alleviating the muffled speech problem.
- DNN-based postfiltering models outperform the conventional ones.
- Neural Network models require a large amount of data.
- Data collection for speaker-dependent systems is both time consuming and costly.
- Adaptation to an unseen speaker with limited data is necessary but a challenging task.

Methods

- **Speaker-Independent Text-To-Speech(SPSS) Model**
  - The model predicts acoustic features (MGC, LF0, BAP and VUV) given the text and linguistic features.
  - Three FF layers followed by one Long-Short Term Memory (LSTM) layer and one output layer is trained where an FF layer, the LSTM layer and the output layer have 512, 256 and 154 (the dimension of the output) units, respectively.
- **Speaker-Independent Postfiltering Models**
  - DNN-, RNN-, and CNN-based postfilters are applied after maximum-likelihood parameter generation (MLPG).

Network Architectures

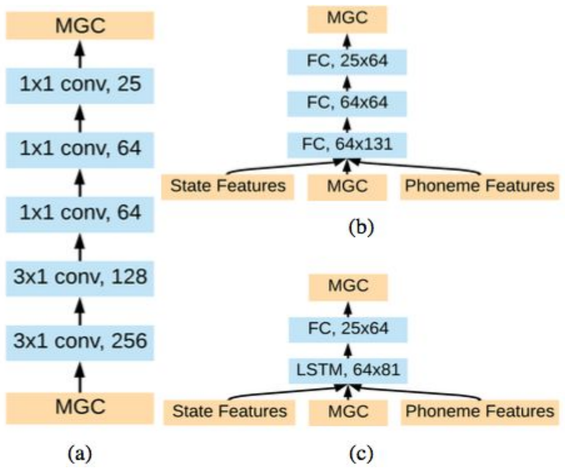

Figure 1: (a) CNN-based postfilter, (b) Feedforward postfilter, (c) RNN-based postfilter.

Methods

- **Cluster-Based Initialization**
  - Since adaptation is performed with very limited data, the optimization algorithm can quickly fall into a nearby local optima with low chances of escaping it. In order to prevent this, we clustered the reference speakers into 5 groups using i-vectors with the k-means method.
  - For each cluster, one model is generated by adapting the SI model with the utterances of the speakers belonging to that cluster.
  - While adapting for a target speaker, the model is initialized with one of these 5 pre-trained models that is closest to the target speaker.
- **Adversarial Training**
  - We fine-tuned the DNN-, RNN-, and CNN-based postfilters using adversarial loss in addition to MSE loss.

Subjective Results

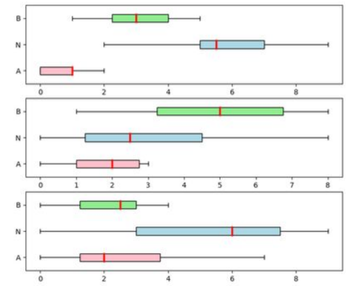

Figure 2: In the top figure, SI-baseline system (A) is compared with the RNN-SI postfilter (B) using the AB test. Significance (p-value) is 0.01. In the middle figure, SI-baseline system (A) is compared with the RNN-SI postfilter (B) using the AB test. Significance (p-value) is 0.01. In the bottom figure, RNN- SI posttfilter (A) is compared with the CNN-SI postfilter (B) using the AB test. Significance (p-value) is 0.55.

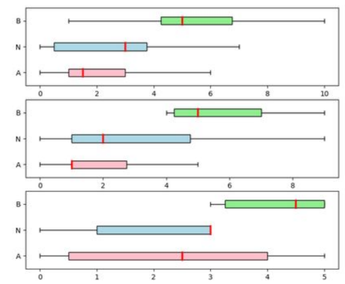

Figure 3: RNN postfilter (A) is compared with the FF+CI+ADV postfilter (B) using the ABX test. Results are shown when the adaptation data is 5 sec (top figure), 10 sec (middle figure), 15 sec (bottom figure). Significance (p-value) of the 5 and 10 sec cases are 0.01 whereas significance of the 15 sec case is 0.03.

Database

- **Wall Street Journal Speech Database**
  - 154 Features including 25 MGC, 1 LF0, 25 BAP together with their delta and delta-delta features.
  - Voiced/Unvoiced binary information.
  - Sampling rate of 16 KHz and 5 msec frame rate.
  - Among the total of 156 speakers, 135 of them were used for training, the remaining 21 speakers' data for test and adaptation.

Objective Results

POSTFILTER	MCD
SI-Baseline	5.19
FF-SI-PF	5.89
RNN-SI-PF	**5.16**
CNN-SI-PF	5.45
FF-SI-PF-CI	5.60
RNN-SI-PF-CI	5.23
CNN-SI-PF-CI	5.47
FF-SI-PF on CNN-SI-PF	6.15
RNN-SI-PF on CNN-SI-PF	5.34

Table 1: Mel cepstral distortion (MCD) scores of the speaker-independent postfilters with and without cluster-based initialization (CI) are shown. Scores for tandem use of FF- and RNN-based postfilters with the CNN-based postfilter are also shown.

POSTFILTER	5 sec	10 sec	15 sec
FF	5.74	5.68	5.65
FF+CI	5.45	5.35	5.31
FF+CI+ADV	5.40	5.16	5.11
FF on CNN-PF	5.69	5.60	5.52
RNN	**5.15**	**5.14**	5.15
RNN+CI	5.21	5.20	5.20
RNN+CI+ADV	5.35	5.15	**5.07**
RNN on CNN-PF	5.30	5.31	5.31
CNN	5.45	5.31	5.27
CNN+CI	5.45	5.29	5.25
CNN+CI+ADV	7.24	7.01	6.99

Table 2: Mel cepstral distortion (MCD) scores of the speaker-adapted postfilters with 5, 10, and 15 seconds of adaptation data.