UNSUPERVISED SPEAKER ADAPTATION FOR DNN-BASED TTS SYNTHESIS

Yuchen Fan[1] *Yao Qian*[2] *Frank K. Soong*[1] *Lei He*[1]

[1] Microsoft, China
[2]Educational Testing Service Research, USA
{v-yufan, frankkps, helei}@microsoft.com, yqian@ets.org

ABSTRACT

Multi-speaker TTS trained with a general DNN has outperformed individually modelled baseline [1]. Multi-speaker DNN takes advantages of larger amount of training data from multiple speakers to find robust transformations in the hidden layers and covers more speaker variability in the output regression layer. In this paper, we propose a new approach to unsupervised speaker adaptation with multi-speaker DNN. It takes advantage of shared hidden transformation to search for the labels of unlabelled acoustic frames and the found labels are used for speaker adaption. Experimental results show that the new approach of unsupervised adaptation can achieve comparable performance with supervised adaptation both objectively and subjectively. We further extend it to cross-lingual adaptation. It can remove non-native accent and improve the naturalness while keep the same speaker's characteristics.

Index Terms— statistical parametric speech synthesis, deep neural networks, speaker adaptation

1. INTRODUCTION

Employment of Deep Neural Networks (DNNs) leads the research of parametric Text-to-Speech (TTS) synthesis to a new stage [1, 2, 3, 4, 5, 6, 7, 8, 9]. Zen et al. [2] comprehensively addressed some intrinsic limitations of the conventional HMM-based speech synthesis, e.g. decision-tree based contextual state clustering and showed that, on a rather large training corpus ($\sim$ 35,000 sentences), DNN can improve TTS performance over that of GMM-HMM with similar number of parameters. Qian et al. [7] examined various aspects of DNN-based TTS training with a moderate size corpus ($\sim$ 5,000 sentences), which is more commonly used for parametric TTS training. Fan et al. [8] introduced LSTM-based RNN into parametric TTS synthesis, which uses deep structure for state transition modeling and performs the acoustic modelling from a frame to sequence.

DNN can model the corpus of multiple speakers with a multi-speaker structure [1]. The shared hidden layers can encode linguistic diversities from multiple speakers, and get a more robust transformation from linguistic features to acoustic features to benefit synthesized voice quality. Meanwhile, speaker adaptation can also be achieved with limited speech by keeping the speaker-independent or pooled-speaker hidden layers and re-training only the output layer.

However, correct linguistic transcriptions are not always available for adaptation. Unsupervised adaptation needs to be performed without linguistic transcriptions. If working well, it can be used for synthesizing arbitrary speaker's voice with only limited amount of speech from a target speaker, especially useful for a cross-lingual scenario.

For HMM-based speech synthesis, unsupervised adaptation is mostly achieved by getting the linguistic labels from a speech recognizer. King et al. [10] took the triphone labels from recognizer to estimate adaptation transformation. Gibson et al. [11] built a cross-lingual state mapping with the results from a speech recognizer. It should be noted in the recognizer based approach, the accuracy of recognition results cannot be guaranteed for short units like senone. The correctness of long units like phone or word is also generally conditioned by how good the language model is. Limited contextual information, such as tied tri-phone state, i.e., senone, is commonly used in speech recognition while TTS synthesis generally need very rich or full contextual information to obtain good synthesized voice. To obtain correct full contextual labels from recognition results is not easy, especially for the long-span linguistic features.

In this paper, we propose a label search method based on a multi-speaker DNN to find the best matching label which has the closest DNN prediction for unlabeled speech frame. Multi-speaker DNN, trained with many different speakers to cover speaker variability adequately, is expected to make the search more robust. Then, multi-speaker DNN is adapted to a new speaker with the resultant found labels in supervised adaptation. Context information, sample weight and iterative training are examined for further improving the performance of unsupervised adaptation. Our approach does not need language relevant input information, so it can be extended to any other language or even cross-lingual adaptation.

2. MULTI-SPEAKER MODELING AND SUPERVISED SPEAKER ADAPTATION

In DNN-based TTS synthesis, DNN is used as a regression model to map input linguistic features into output acoustic features. DNN is a layer-structured model, which learns jointly a complicated linguistic feature transformation in hidden layers and a speaker-specific acoustic space in output layer. DNN can be decomposed into two stages: linguistic transformation; and acoustic regression. DNN-based TTS synthesis can benefit from multiple speakers' data and solve the adaptation problem by sharing the hidden layers among different speakers.

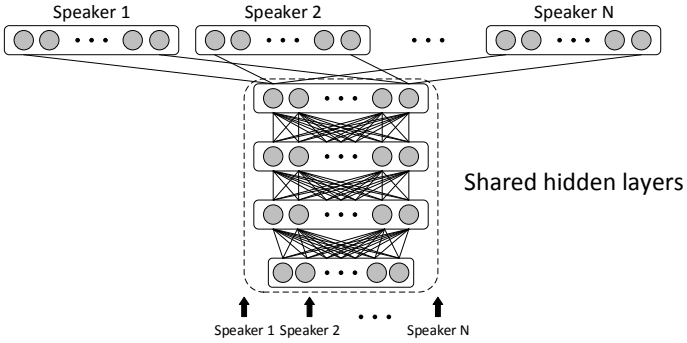

Fig. 1. Multi-speaker DNN Architecture in DNN-based TTS Synthesis [1].

Figure 1 shows the architecture of the proposed multi-speaker DNN. In multi-speaker DNN, hidden layers are shared across all the speakers in the training corpus, as a global linguistic feature transformation acrosses and serves all training speakers. Conversely, each speaker has his own output layer, in the ultimate regression layer, to model the specific acoustic space for a particular speaker. Multi-speaker DNN is jointly optimized with multiple speakers data. It can synthesize each speaker's voice with the knowledge of other speakers.

The shared hidden layers, in the multi-speaker DNN, can be treated as a global linguistic feature transformation universal to all training speakers. The shared hidden layers can also be used to transform the linguistic feature for a new speaker. By fixing the shared hidden layers and only updating the speaker-specified regression layer, supervised adaptation can be achieved with only limited adaptation data and the corresponding linguistic information.

3. UNSUPERVISED SPEAKER ADAPTATION

In speech recognition, unsupervised speaker adaptation can be achieved with adaptation data and its speaker independent recognition result. Similar technique which uses recognizer to get labels is also used in HMM-based TTS synthesis. However, the transcription obtained by recognition usually covers only part of all context labels used in TTS synthesis. Although full context labels can be predicted from recognition result, recognition errors can cause problems. In cross-lingual scenario, phone sets and linguistic characteristics in different languages make the predictions even more difficult.

In our proposed unsupervised speaker adaptation, label search is performed at the frame level. For multi-speaker DNN-based TTS, based on the minimal square error training criterion, given the acoustic feature $\boldsymbol{o}$, the best-match linguistic label $\boldsymbol{l}$ is

$$\boldsymbol{l} = \arg\min_{\boldsymbol{l}} \mathrm{D}(\min_{s} \mathscr{F}_s(\boldsymbol{l}), \boldsymbol{o})$$

where $\mathscr{F}_s(\cdot)$ denotes DNN transformation for speaker s and $\mathrm{D}(\cdot)$ is the distance metric for acoustic features. Based on this criterion, when there are rich enough speech data for a multi-speaker DNN, the label can be more accurately found by traversing all possible linguistic and speaker combinations.

However, it's apparently very difficult, even impossible, to traverse the whole linguistic set in linear time complexity. Some trade-off between search precision and complexity must be made. To achieve fast adaptation for practical applications, we reduce the search space to a smaller but still linguistic-rich label set.

On the other hand, the number of speakers in multi-speaker DNN is also a crucial parameter for label search. Because the number of speakers is usually limited in corpus for TTS, we build multi-speaker DNN with a speech recognition corpus to include many speakers to enrich its variability.

With the full context linguistic labels for each frame, the problem is simplified and identical to the supervised adaptation. Thus the output regression layer of DNN can be estimated by the least squares approach.

Although the proposed unsupervised adaptation is quite straightforward, some issues still need to be addressed.

3.1. Distance Metric

Following the training criterion of DNN, the distance metric for label search is designated as square error. Note that all features for DNN training are normalized into the zero-mean and unit-variance, so the distance metric is also computed in the same normalized scale.

3.2. Context Information

The contextual information, which can become more discriminative with long-span linguistic features and speaker variability, is widely used in speech recognition and speaker verification. Context frames are concatenated to a new high-dimensional vector for distance measure in label search. Since context information takes additional computational cost in the search process, no more than 5 contextual frames was used in this work.

3.3. Sample Weight

Not all the frames can find their corresponding labels exactly, given a limited data set. As a result, certain mappings between linguistic labels and acoustic features are imprecise. Based on the search criterion, pairs with closer distance are more reliable and then should play more crucial roles in adaptation. Therefore, the weighted least square error method is employed in estimating the output layer of adapted network where inverse of distance between acoustic features is chosen as the sample weight.

3.4. Iterative Training

Although sufficient number of speakers can cover most of the speaker variability and make label search feasible, the adapted model, which yields predictions closer to the target speaker, can improve label search precision. The proposed unsupervised adaptation can be optimized iteratively, i.e., label search is repeated iteratively with the adapted model and output layer estimation is performed with resultant labels.

4. EXPERIMENTS

4.1. Experimental Setup

To increase speaker variability so as to obtain better label search result, the multi-speaker DNN is built on the "long" part of WSJ1 corpus, used in speech recognition. Each speaker has 1,200 sentences for training and 40 sentences for testing. In this corpus, there are totally 25 native American English speakers, in which 23 speakers' voices are used for training multi-speaker DNN, and the other two, i.e., one female and one male, speakers' voices are used for evaluating the speaker adaptation performance. Additionally, another bilingual Mandarin and English speaker's voice is used for cross-lingual adaptation.

Speech signals are sampled at 16 kHz, windowed by a 25-ms window, and shifted every 5-ms. An LPC of 40th order is transformed into static LSPs and their dynamic counterparts. The phonetic and prosodic contexts include quin-phone, the position of a phone, syllable and word in phrase and sentence, the length of word and phrase, stress of syllable, POS of word.

In the multi-speaker DNN, the input feature vectors contain 353 dimensions, where 326 are binary features for categorical linguistic contexts and the rest are numerical linguistic contexts. The output feature vector contains a voiced/unvoiced flag, log F0, LSP, gain, their dynamic counterparts, totally 127 dimensions. Voiced/unvoiced flag is a binary feature to indicate the current frame is voiced or not. DNN is set with 3 hidden layers and 1024 nodes for each layer. An exponential decay function is used to interpolate F0 in unvoiced regions. 80% of the silence frames are removed from the training data to balance the training data and to reduce computational cost. Both input and output features of training data are normalized to zero mean and unity variance. DNN training is based on the computational network toolkit (CNTK) [12].

For testing, DNN outputs are fed into a parameter generation module to generate smooth feature parameters with the dynamic constraints. Then formant sharpening based on LSP frequencies is used to reduce the over-smoothing problem in modeling. Finally speech waveforms are synthesized by an LPC synthesizer with generated speech parameters.

Objective and subjective measures are used to evaluate the performance of TTS systems on testing data. Synthesis quality is measured objectively in terms of distortions between natural test utterances of the original speaker and the synthesized speech frame-synchronously where oracle state durations (obtained by forced alignment) of natural speech are used. The objective measures are F0 distortion in the root mean squared error (RMSE), voiced/unvoiced (V/U) swapping errors and normalized spectrum distance in log spectral distance (LSD). The subjective measures were done on speech naturalness and speaker similarity. In the naturalness section, each subject is asked to compare natural and synthesized speech and give 5-point scale scores, from 1 as "bad" to 5 as "excellent". Speaker similarity is done similarly, from 1 as "very different" to 5 as "very close". Mean opinion score (MOS) indicates the summarized measurement. In each subjective test, we invite 10 native English subjects to participate and each subject evaluates 40 groups with headsets.

4.2. Evaluation Results and Analysis

4.2.1. Intra-lingual Adaptation

To evaluate unsupervised adaptation, we first tested the proposed method on two speakers in WSJ1, including the context information, sample weight and iterative training. Due to identical scripts across all the speakers in WSJ1, we chose 1,100 sentences to train the multi-speaker DNN and reserved 100 sentences for adaptation.

The objective results are shown in Table 1, context information including the long-span features contributes to the reduction of LSD. Sample weight avoids some inaccurate labels and helps F0 modelling. Iterative training leads a more precise search with adapted model and improve adaptation effectively. Compared with supervised adaptation, unsupervised adaptation still has a gap of performance to fill, but not significant.

In subjective test, we take the voice synthesized with extracted features, denoted as "vocoder", as the upper bound of synthesized voice quality. Figures 2 and 3 suggest that unsupervised adaptation results are comparable to supervised one both in naturalness and similarity with a slightly worse MOS scores.

Table 1. Objective measures of intra-lingual adaptation on WSJ1.

Speaker	Female			Male		
Measures	LSD (dB)	V/U Err Rate (%)	F0 RMSE (Hz)	LSD (dB)	V/U Err Rate (%)	F0 RMSE (Hz)
1 Frame	4.29	4.11	26.5	4.63	4.10	18.1
3 Frames ($\pm$ 1 context)	4.25	4.04	26.4	4.59	4.09	18.8
5 Frames ($\pm$ 2 context)	4.22	4.06	26.5	4.57	4.10	19.0
+ Sample weight	4.23	4.06	25.6	4.58	4.08	17.7
+ Second iteration	4.18	4.05	25.2	4.47	4.05	17.0
Supervised	4.01	3.88	25.6	4.25	3.19	15.2

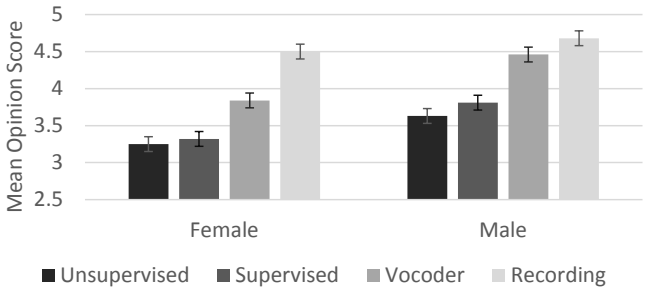

Fig. 2. Naturalness MOS results for intra-lingual adaptation.

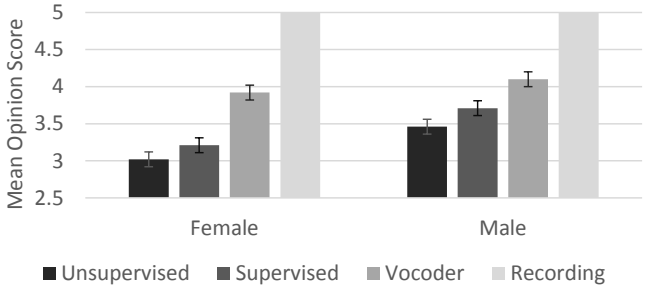

Fig. 3. Similarity MOS results for intra-lingual adaptation.

4.2.2. Cross-lingual Adaptation

In cross-lingual adaptation, data is from a native-Mandarin speaker who also can also speak English. Both Mandarin and English utterances are collected from the person. 100 Mandarin utterances are used for adaptation based upon English multi-speaker DNN, referred as cross-lingual adaption, while 100 English utterances for adaptation based on the same English multi-speaker DNN, called intra-lingual adaptation and employed as a comparison to that of cross-lingual adaptation.

Table 2. Subjective measures of intra-lingual and cross-lingual adaptation for a bi-lingual speaker.

	Naturalness MOS	Similarity MOS
Intra-lingual	3.27	2.50
Cross-lingual	3.06	2.65
Supervised	2.84	3.78
Recording	2.93	5.00

The results of objective measures in Table 2 show that the performance of unsupervised adaptation in cross-lingual scenario is comparable to the intra-lingual one with a slightly worse naturalness and better speaker similarity. In naturalness test, the unsupervised adaptation is better than the supervised one, which is due to that the non-native English speakers have some wrong pronunciation or accent affected by their own native language. The unsupervised adaptation without limitation of linguistic information can assign a closer label to the mispronunciation and reduce the accent in some cases. We conjecture also why the unsupervised adaptation gets worse results in similarity scores, comparing with the supervised adaptation. In addition, cross-lingual adaptation can keep more accent from the speaker's native language to achieve better similarity than the intra-lingual one.

5. CONCLUSIONS

In this paper, we propose a new approach to unsupervised speaker adaptation, based upon a multi-speaker DNN for TTS synthesis. The multi-speaker DNN, trained with many speakers' voices, serves as a big database for "label search" and robust linguistic transformation for adaptation. The experimental results on both intra-lingual and cross-lingual speaker adaptation show that the proposed approach is promising.

6. REFERENCES

[1] Yuchen Fan, Yao Qian, Frank K. Soong, and Lei He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[2] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[3] Heiga Zen and Andrew Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, 2014, pp. 3844–3848.

[4] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, 2015, pp. 4470–4474.

[5] Keiichi Tokuda and Heiga Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4215–4219.

[6] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015, pp. 4460–4464.

[7] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Proc. ICASSP*, 2014, pp. 3829–3833.

[8] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[9] Yuchen Fan, Yao Qian, Frank K. Soong, and Lei He, "Sequence generation error (SGE) minimization based deep neural networks training for text-to-speech synthesis," in *Proc. Interspeech*, 2015.

[10] Simon King, Keiichi Tokuda, Heiga Zen, and Junichi Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. Interspeech*, 2015, pp. 1869–1872.

[11] Matthew Gibson, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," pp. 4642–4645, 2010.

[12] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep. MSR-TR-2014-112, August 2014.