SPEAKER AND LANGUAGE FACTORIZATION IN 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

We have successfully proposed to use multi-speaker modelling in DNN-based TTS synthesis for improved voice quality with limited available data from a speaker. In this paper, we propose a new speaker and language factorized DNN, where speaker-specific layers are used for multi-speaker modelling, and shared layers and language-specific layers are employed for multi-language, linguistic feature transformation. Experimental results on a speech corpus of multiple speakers in both Mandarin and English show that the proposed factorized DNN can not only achieve a similar voice quality as that of a multi-speaker DNN, but also perform polyglot synthesis with a monolingual speaker's voice.

Index Terms— statistical parametric speech synthesis, deep neural networks, speaker and language factorization, polyglot speech synthesis

1. INTRODUCTION

When data is collected from different speakers, languages and speaking styles, how to train Text-to-Speech (TTS) system effectively and efficiently becomes an interesting research topic.

The voice characteristics of speakers and languages are two dominant factors in Text-to-Speech (TTS) synthesis. Factorizing and integrating the speaker and languages dependent parts may make TTS more versatile to synthesize any speaker's voice in any language. Zen et al. [1] proposed the speaker and language factorization (SLF) framework to factorize the speaker and language characteristics for HMM-based TTS. First, non-polyglot speaker's voice becomes polyglot in multiple languages. Second, one speaker's voice with limited data can be pooled with multiple speakers in different languages. Third, new languages can be adapted with only limited data.

Deep Neural Networks (DNNs) have advanced parametric Text-to-Speech (TTS) synthesis to a new frontier [2, 3, 4, 5, 6, 7, 8, 9, 10]. Zen et al. [2] investigated DNN-based TTS and comprehensively pointed out some intrinsic limitations of the conventional HMM-based speech synthesis, e.g. decision-tree based contextual state clustering. They showed that, on a rather large training corpus ($\sim$ 35,000 sentences), DNN can yield better TTS performance than its GMM-HMM counterpart with a 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 upgrades the acoustic model from frame level to sequence (sentence) level.

In DNN-based TTS, DNN is used as regression model to map input linguistic features to output acoustic features. DNN can be viewed as layer-structured model, which jointly learns a complicated linguistic feature transformation in multiple hidden layers to a speaker-specific acoustic space. For DNN-based TTS, the concept of speaker factorization has been introduced in the multi-speaker DNN [9], in which all speakers share the same hidden layers and each speaker has a speaker-specific output layer. In multi-speaker DNN, network is decomposed into two parts: shared hidden layers which is used for linguistic transformation and speaker-specific layers which factorize the speaker characteristics in the data. The shared hidden layers across all speakers, which are populated with more linguistic diversities, are expected to yield an enriched linguistic to acoustic transformation to improve synthesized voice quality. Meanwhile, speaker adaptation with limited speech can also benefit by freezing the speaker-independent hidden layers and re-training the output layer only.

In this paper, we propose a speaker and language factorization framework for DNN-based TTS. The framework takes the advantages of the layer-wise structure and the flexible topology in DNN, and decomposes the network into three independent functional layers: language-specific layers; shared layers and speaker-specific layers. The modularized DNN gets the structural flexibility for modelling and synthesizing voice with any speaker and language characteristics. Language-specific layers exploiting with multiple speakers' data in specific languages and speaker-specific layers populated with multiple languages' data from specific speakers become more robust than separated and independent modelling. Shared layers serve as a bridge to connect the language and

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

speaker-specific layers. The barrier between languages and speakers is unblocked and monolingual voice can become polyglot, i.e., synthesizing speech in different language.

2. DNN-BASED SPEECH SYNTHESIS

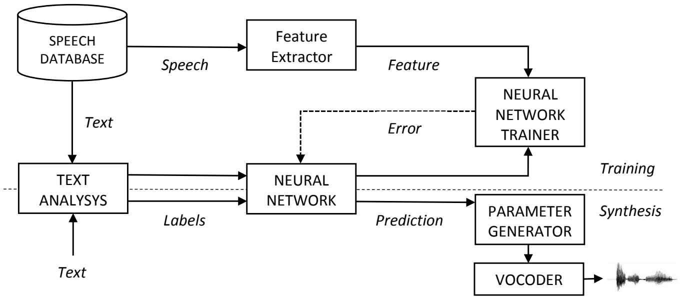

Fig. 1. Framework of DNN-based TTS synthesis.

Figure 1 shows a block diagram of a DNN-based speech synthesis system, which consists of both training and synthesis. In training, the acoustic features for DNN input are first extracted from the speech signal with the feature extraction module and the linguistic features for DNN output are converted from the contextual labels generated through text analysis. The parameters of DNN are trained by using pairs of input and output features with a mini-batched, back-propagation algorithm. The cost function is defined as the errors between the original acoustic features and the predicted outputs of each frame in the training data. In synthesis, input text is first analyzed into labels, then mapped onto the acoustic features by the trained DNN. In order to generate smooth parameter trajectories, dynamic features are used as constraints in speech parameter generation, where predicted features are used as mean vectors and global variances of the training data are adopted for generating speech parameters by maximizing the probability. Finally, the speech waveform is synthesized from the generated parameters with a vocoder.

3. SPEAKER AND LANGUAGE FACTORIZATION

DNN is a layer-structured model equipped with stacked multiple layers of linear transformations and non-linear activations, where linear transformation can also be built to connect between any two activations. So the speaker and language factorization in DNN-based TTS can be achieved by specifying transformations for any specific speaker in any specific languages.

3.1. Model Structure

Figure 2 shows the topology of DNN factorized in speaker and language factorization. In this framework, DNN is structured in three major layers: language-specific layers, shared layers and speaker-specific layers. Each layer can have multiple hidden layers. Language-specific layers are built to have a specific transformation for each language. Similarly, speaker-specific layers have a specific transformation for each speaker. Shared layers are both speaker and language independent. The linguistic features will first go through the language-specific layers which the features belong to, then pass through the shared layers, finally predict the speaker specific acoustic features with the corresponding speaker-specific layers.

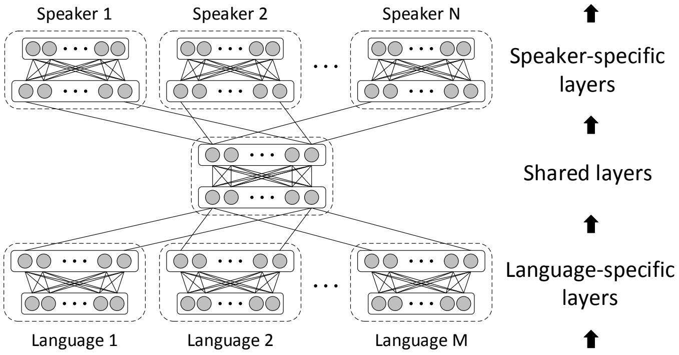

Fig. 2. DNN topology for speaker and language factorization.

Formally, the factorized DNN for speaker s in language l, denoted as $\mathcal{F}_{l,s}(\cdot)$, can be decomposed to

$$y = \mathcal{F}_{l,s}(\boldsymbol{x}) = \mathcal{S}_s(\mathcal{H}(\mathcal{L}_l(\boldsymbol{x})))$$

where $\boldsymbol{x}$ is the linguistic input feature vector; $\boldsymbol{y}$ is the acoustic output feature vector; and $\mathcal{L}_l(\cdot)$, $\mathcal{H}(\cdot)$, $\mathcal{S}_s(\cdot)$ are the language-specific layers for language l, shared layers and speaker-specific layers for speaker s, respectively.

In this framework, the language-specific layers are trained by multiple speakers' voices as multi-speaker DNN [9], which will benefit the speakers whose data are limited. Similarly, the speaker-specific layers are trained by the specific speaker's data in different languages, which will also benefit the languages which are under-resourced.

3.2. Model Training

Training of the speaker and language factorized DNN is still based on the mini-batched stochastic gradient decent (SGD) algorithm as the conventional DNN. For each linguistic and acoustic feature pair of speaker s in language l, only the related parts' gradient in the network will be computed for update, while gradient of the rest will be set to zero. Mini-batched parallelization is used to speed up the neural network training. In our proposed DNN, data from different speakers and languages needs different parts of network for calculation, so that parallelization becomes very hard.

To achieve efficient parallelization, model structure must be identical to data from all the speakers and languages.

Hence, we replace selective operations in model updates by multiplication and summation. Formally, the factorized DNN becomes

$$y = \mathcal{F}_{\boldsymbol{l},\boldsymbol{s}}(\boldsymbol{x}) = \sum_{i=1}^{N} \boldsymbol{s}_i \mathcal{S}_i(\mathcal{H}(\sum_{j=1}^{M} \boldsymbol{l}_j \mathcal{L}_j(\boldsymbol{x})))$$

where $\boldsymbol{s}$ and $\boldsymbol{l}$ are the one-hot vectors to indicate speaker and language identity, respectively. The uniform structure for all the speakers and languages can be directly applied in mini-batched parallelization. Although it introduces some redundant computations for irrelevant speakers and languages, the efficiency of parallelization makes the redundancy worthwhile.

3.3. Polyglot

In speaker and language factorized DNN, modules from different parts can be arbitrarily combined to synthesized voice of any speaker in any language, that is, all speakers' voices can perform polyglot speech synthesis in the languages covered in training. The speaker and language independent shared layers play a crucial role in the proposed polyglot synthesis. In the layer-wise structure of DNN, there are no connections among the nodes in the same hidden layer, so that the nodes in the same layer behave uncorrelated. Without the shared layers, the intermediate nodes between speaker-specific and language-specific layers may be speaker or language dependent, which makes the modules in factorized DNN unable to transfer to unseen speaker or language. The shared hidden layers rebuild the relations among the intermediate nodes and make the polyglot speech synthesis possible.

4. EXPERIMENTS

4.1. Experimental Setup

A corpus of 3 native Mandarin speakers, including 2 females and 1 male, who can also speak English, is used in our experiments. The corpus, in which the text is got from one year's newspaper by brute-forth search algorithm, is both phonetically and prosodically rich. Each speaker has 900 utterances in Mandarin and 900 utterances in English for training, and 40 utterances in Mandarin and 40 utterances in English for testing. The average length of the utterances is 3s. The sentences are uttered in the style of reading. 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 positions of a phone, syllable and word in phrases and sentences, the length of a word and a phrase, stress of a syllable, POS of a word.

In training DNN, the Mandarin input feature vectors contain 611 dimensions, among them 576 are binary features for categorical linguistic contexts and the rest are numerical linguistic contexts, while the English input feature vectors contain 331 dimensions, where 304 are binary features. The output feature vector contains a voiced/unvoiced flag, log F0, LSP, gain and their dynamic counterparts, in a total of 127 dimensions. Voiced/unvoiced flag is a binary feature to indicate the voicing status of the current frame. 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 silence frames are removed from the training data to balance the training data and to reduce the computational cost. Removing silence frames in DNN training was found useful for avoiding DNN over-learning the silence label in speech recognition task. 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) [11].

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

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) errors and normalized spectrum distance in log spectral distance (LSD). The subjective measures are used to measure the naturalness and speaker similarity. In the naturalness subjective test, each subject is to compare natural speech with synthesized speech and give a 5-point score, from 1 ("bad") to 5 ("excellent"). The speaker similarity is measured similarly, from 1 ("very different") to 5 ("very close"). Mean opinion score (MOS) indicates the summarized measurements.

4.2. Evaluation Results and Analysis

4.2.1. Network topology

The number of layers in the speaker, language and shared layers of the factorized DNN determines the network topology and corresponding performance. In this section, we take all three speakers' bilingual voices for training. Considering the size of training corpus, we evaluate different combinations of topologies for the proposed factorized DNN based on a structure of 3 hidden layers and 1 output layer.

Table 1 shows the average objective test results of the three speakers and two languages for the factorized DNN in different topologies. In Table 1, L, H, S denote the number of

Table 1. Objective Measures of speaker and language factorized DNN in different topologies

Topologies			LSD	V/U Err	F0 RMSE
L	H	S	(dB)	(%)	(Hz)
1	2	1	4.49	2.39	26.8
1	1	2	4.60	2.46	27.2
2	1	1	4.49	2.39	26.4
Multi-speaker			4.44	2.36	26.3

linguistic, shared and speaker layers, respectively. From the objective results, factorized DNN with 2 linguistic-specific layers, 1 shared layer and 1 speaker-specific layer gets the best performance and is similar to the multi-speaker DNN [9], which builds multi-speaker network for Mandarin and English separately.

In the subjective test, we compare the factorized DNN with the optimal topology, multi-speaker DNN on monolanguage and recording by naturalness MOS. For different systems, speakers and languages, 120 judgements by 10 subjects of native speakers are performed.

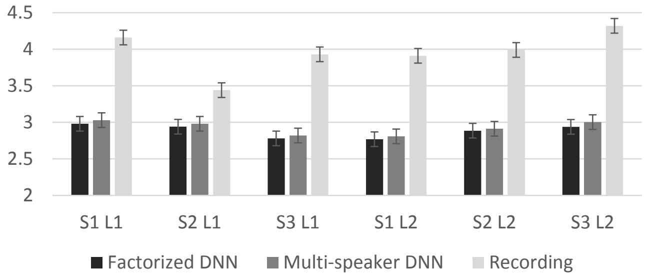

Fig. 3. Naturalness MOS test results of speaker and language factorized DNN.

Figure 3 shows the naturalness MOS results for all three training speakers: S1 (female), S2 (male) and S3 (female), and two languages: L1 (English) and L2 (Mandarin). The proposed factorized DNN behaves almost the same as the multi-speaker DNN in naturalness test, which indicates the factorized DNN can benefit the synthesis quality with multiple speakers' voice as multi-speaker DNN and build transformations for linguistic features in all the languages.

4.2.2. Polyglot Synthesis

To evaluate the capability of factorized DNN for polyglot synthesis without the training data from multi-lingual speakers, we remove the English data of S3 (female) from training and conduct a naturalness MOS test and a similarity MOS test in English for polyglot synthesis. DNN-based monolingual synthesis is built upon the female speaker's English recordings and used as baseline. In both naturalness and similarity test, we invited 10 native English subjects and each subject was to evaluates 40 groups by using headsets.

Table 2. Subjective measures of polyglot synthesis.

	Naturalness MOS	Similarity MOS
Polyglot	2.44	2.13
Monolingual	2.69	2.71
Recording	3.71	5.00

Subjective results in Table 2 show that factorized DNN can achieve polyglot synthesis without using speech data from a multi-lingual speaker. In contrast to recording, the naturalness of the polyglot synthesis is comparable with DNN-based monolingual system, i.e., 2.44 vs. 2.69, while the speaker similarity is lower but still acceptable, i.e., 2.13 vs. 2.71.

5. CONCLUSIONS

In this paper, we propose a speaker and language factorized DNN. The factorized DNN can model voices of multiple speakers in multiple languages simultaneously. The shared layers in the middle of factorized DNN can exploit the commonalities among different languages and speakers so as to transfer learned knowledge to a new speaker and language combination. Experimental results on a corpus of multiple speakers in both Mandarin and English show that the proposed factorized DNN can achieve polyglot synthesis for a monolingual speaker. Our future research will use more speakers in more languages to evaluate the performance of factorized DNN for TTS synthesis in a scale-up manner.

6. REFERENCES

[1] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark JF Gales, Kate Knill, Sacha Krstulović, and Javier Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[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, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[10] 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.

[11] 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.