INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Personalized, Cross-lingual TTS Using Phonetic Posteriorgrams

Lifa Sun, Hao Wang, Shiyin Kang, Kun Li and Helen Meng

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR, China
{lfsun,hwang,sykang,kli,hmmeng}@se.cuhk.edu.hk

Abstract

We present a novel approach that enables a target speaker (e.g. monolingual Chinese speaker) to speak a new language (e.g. English) based on arbitrary textual input. Our system includes a trained English speaker-independent automatic speech recognition (SI-ASR) engine using TIMIT. Given the target speaker's speech in a non-target language, we generate Phonetic PosteriorGrams (PPGs) with the SI-ASR and then train a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks (DBLSTM) to model the relationships between the PPGs and the acoustic signal. Synthesis involves input of arbitrary text to a general TTS engine (trained on any non-target speaker), the output of which is indexed by SI-ASR as PPGs. These are used by the DBLSTM to synthesize the target language in the target speaker's voice. A main advantage of this approach has very low training data requirement of the target speaker which can be in any language, as compared with a reference approach of training a special TTS engine using many recordings from the target speaker only in the target language. For a given target speaker, our proposed approach trained on 100 Mandarin (i.e. non-target language) utterances achieves comparable performance (in MOS and ABX test) of English synthetic speech as an HTS system trained on 1,000 English utterances.

Index Terms: phonetic posteriorgrams, personalized, cross-lingual, TTS, DBLSTM

1. Introduction

The advancements of speech technology in recent years bring growing demands for personalized speech service and applications. Cross-lingual TTS aims at synthesizing speech in a specific language not spoken by the target speaker. The technology can benefit various fields such as computer-aided language learning for personalized perceptual feedback and assistive technologies for the speech impaired.

Previous approaches can be divided roughly into two categories, i.e. GMM-HMM-based approaches and unit selection-based approaches. GMM-HMM-based approaches [1–4] require GMM-HMM training for corpora in two different languages (i.e. a non-target language and a target language). In [1], speech recordings from a reference bilingual (English and Mandarin) speaker were used to build language-specific decision trees separately. The Kullback-Leibler divergence (KLD) [5] was used to measure the distance between any pair of leaf nodes. Every leaf node (tied GMM states) in the Mandarin tree can be mapped to its nearest counterpart in the English tree in a minimum KLD sense. This process is called state mapping and the obtained mapping information may be applied to any monolingual English speaker for synthesizing his/her Mandarin speech. A similar approach was presented in [2], which established a state mapping between two Average Voice models in different languages. However, neither a corpus from a bilingual speaker nor speech data for training Average Voices is easily available. The spectral space warping [3] and the KLD-DNN [4] approaches used only speech recordings from a monolingual reference speaker in the target language and a monolingual target speaker in a non-target language. These proposed approaches can be regarded as modified versions of state mapping mechanism. The spectral space warping approach [3] equalized the numbers of the leaf nodes in two trained language-specific decision trees and found an optimal one-to-one leaf node mapping in a minimum total sum of KLDs sense. The KLD-DNN approach [4] employed a Deep Neural Network (DNN)-based speaker-independent automatic speech recognition (SI-ASR) to index TTS senones (leaf node) in both languages as their average posterior distributions. Then a senone mapping was established with a criterion of minimum KLD between the average posterior distributions.

Unit selection-based approaches [6–8] aim at rendering reference speech trajectories in a target language using the best matched speech segments from a target speaker's recordings in another language. In [6], a phoneme mapping algorithm was devised based on a proposed function for computing similarity scores between phonemes. However, phoneme-level unit selection is not robust when the two languages belong to different linguistic families. In [7, 8], frame-level units were applied and worked for very phonetically different languages, e.g. English and Mandarin. A reference speaker's recordings in the target language were modified with a piecewise-linear transform-based [7] or a bilinear transform-based [8] spectral frequency warping techniques for approximating the target speaker's voice. Then these modified speech data were used as a guide to select the most appropriate frames from the target speaker's recordings for generating the target speaker's speech in a new language.

This paper proposes a novel approach to personalized cross-lingual TTS based on our previous work [9, 10]. This approach uses a DNN-based SI-ASR (similar to [4]) for generating Phonetic PosteriorGrams (PPGs) of the target speaker's speech in a non-target language and a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM) for modeling the relationship between the PPGs and the corresponding speech signal. Cross-lingual TTS can be achieved by combining this framework with a general TTS engine in the target language. The TTS engine takes arbitrary text input and synthesizes speech in the target language. Then the synthesized speech is fed into the same SI-ASR to obtain the corresponding PPGs which are used to drive the trained DBLSTM model. Consequently, the target speaker's speech in the target language is generated. The advantages of this approach are: 1) it has a very low training data requirement of the target speaker's

http://dx.doi.org/10.21437/Interspeech.2016-1043

speech recordings in any language (e.g. 100 sentences [9]), thus enhancing practicability; 2) it can very easily be applied to synthesizing a fixed target speaker's speech in any language simply by inserting an arbitrary TTS engine in that language, thus offering flexibility.

The rest of the paper is organized as follows: Section 2 introduces PPGs. Section 3 describes our proposed cross-lingual TTS system using PPGs. Section 4 presents the experiments and the evaluation results. Section 5 concludes this paper.

2. Phonetic PosteriorGrams

A PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class for a specific time frame of one utterance [11, 12]. A phonetic class may refer to a word, a phone or a senone. In this paper, senones are treated as the phonetic class to represent the whole speaker-independent phonetic space. Fig. 1 shows an example of PPG representation for the spoken phrase "particular case". The horizontal axis represents time in seconds and the vertical one contains indices of phonetic classes.

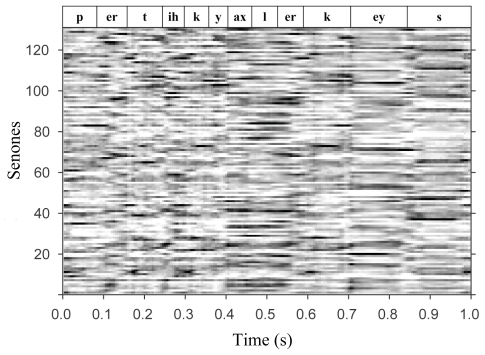

Figure 1: PPG representation of the spoken phrase "particular case". In this example, the number of senones is 131. Darker shade implies a higher posterior probability.

PPGs from an SI-ASR are used to represent articulation of speech sounds in a speaker-normalized space and correspond to speech content speaker-independently. As indicated in [1–4, 6, 7], small speech segments like tied-states and frames can be shared in different languages. Hence PPGs, as frame-level units, may be deemed language-independent. In view of the speaker-independent and language-independent properties of PPGs, we can use them as a representation of speech that bridge across speakers and language boundaries.

3. Cross-lingual TTS Using PPGs

Given a target speaker's speech in a non-target language (e.g. Mandarin), our objective is to build a TTS system which can synthesize this target speaker's speech for arbitrary textual input in the target language (e.g. English). In this system, a DBLSTM framework is used to represent the acoustic-phonetic space of target speaker which can map PPGs to acoustic features.

3.1. Overview

As illustrated in Fig. 2, the proposed approach is divided into three stages: training stage 1, training stage 2 and the synthesis stage. In training stage 1, an SI-ASR model is introduced to obtain a PPG representation of any input speech. Training stage 2 models the relationships between the PPGs and acoustic features of the target speaker's speech recordings (in a non-target language) for speech parameter generation. The 3rd stage of synthesis drives the trained DBLSTM model with PPGs of the waveform (obtained from an arbitrary TTS engine). These three stages will be elaborated in the following subsections.

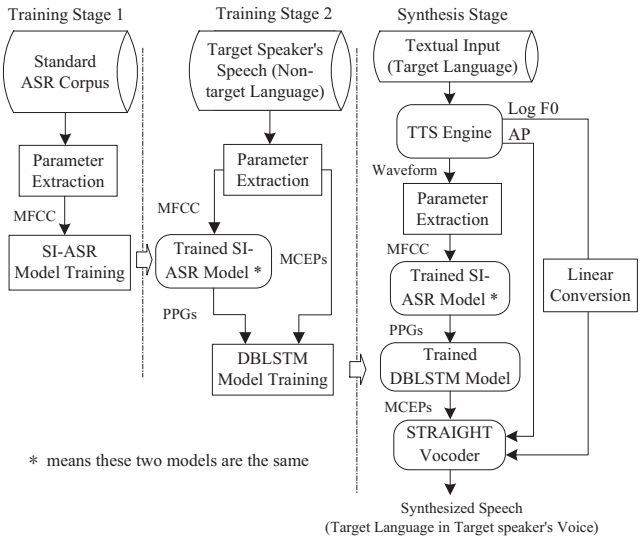

Figure 2: Schematic diagram of cross-lingual TTS using PPGs. SI stands for speaker-independent.

3.2. Training Stages 1 and 2

In training stage 1, an SI-ASR system is trained for PPGs. The input $\mathbf{x}_{n,t}$ is the acoustic MFCC feature vector at the t^{th} frame of the n^{th} utterance. The output $p(\mathbf{s}_{n,t} \mid \mathbf{x}_{n,t})$ is the posterior probability of the phonetic class vector $\mathbf{s}_{n,t}$ given $\mathbf{x}_{n,t}$. For simplicity, $p(\mathbf{s}_{n,t} \mid \mathbf{x}_{n,t})$ is also denoted as $\mathbf{p}_{n,t}$. To train the SI-ASR model, a multi-speaker ASR corpus is used.

Training stage 2 trains a DBLSTM model to get the mapping relationships between the PPGs, denoted as $\mathbf{p}_{n,t}$, and the acoustic features, i.e. Mel-cepstral coefficients (MCEPs), donated as $\mathbf{Y}_{n,t}$. The cost function can be represented as:

$$min \sum_{n=1}^{N} ||\mathbf{Y}_n^R - \mathbf{Y}_n^T||^2 \tag{1}$$

where $\mathbf{Y}_n^T = (\mathbf{Y}_{n,1}^T, \cdots, \mathbf{Y}_{n,t}^T, \cdots, \mathbf{Y}_{n,T_n}^T)$ is the target MCEP feature vector of the n^{th} utterance, whereas $\mathbf{Y}_n^R = (\mathbf{Y}_{n,1}^R, \cdots, \mathbf{Y}_{n,t}^R, \cdots, \mathbf{Y}_{n,T_n}^R)$ is the generated MCEP feature vector, i.e. the actual value of the output layer.

The model is trained to minimize the cost function through back-propagation through time (BPTT) technique [13]. Note that the DBLSTM model is trained using only the target speaker's MCEPs features and the PPGs without using any other linguistic information.

3.3. Synthesis Stage

In this stage, a TTS engine (trained on any non-target speaker's speech in the target language) is used to generate the pitch features ($F0$), aperiodic components (AP) and waveforms.

Table 1: Details of corpora in our experiments.

Corpus	TIMIT	CUBIL		CMU ARCTIC	
Speaker	Count: 462	WH		BDL	SLT
Data	English 3,696 Utterances	Mandarin 100 Utterances	English 1,000 Utterances	English 1,000 Utterances	English 1,000 Utterances
Model	SI-ASR	DBLSTM	TTS-WH Engine	TTS-BDL Engine	TTS-SLT Engine

The trained SI-ASR takes the MFCC features of a generated waveform as input and outputs corresponding PPGs. Then, the PPGs are fed into the trained DBLSTM model for generating MCEPs with the target speaker's voice characteristics. AP is directly copied from the TTS and $logF0$ is converted by equalizing the mean and the standard deviation of the $logF0$ generated by the TTS engine and that extracted from the target speaker's speech recordings. Finally, the target speaker's speech in the target language is synthesized using the STRAIGHT vocoder [14].

4. Experiments

4.1. Corpora

Three corpora including TIMIT [15], CUBIL and CMU ARCTIC [16] are involved in the training and synthesis stage of our proposed approach, as shown in table 1. 462 speakers' speech of TIMIT is used to train the SI-ASR model. CUBIL is a bilingual (Mandarin and English) corpus recorded at 16kHZ with mono channel, which is collected for these experiments in our research group at the Chinese University of Hong Kong. Speaker WH is treated as the target speaker in the following experiments. Mandarin and English are treated as the non-target language and the target language respectively. A DBLSTM model is trained using WH's 100 Mandarin utterances. The CMU ARCTIC BDL and the CMU ARCTIC SLT corpora are used for training two general speaker-dependent TTS engines – TTS-BDL and TTS-SLT, respectively.

4.2. Experimental Setup

The signals are sampled at 16kHZ with mono channel, windowed with 25 ms and shifted every 5 ms. Acoustic features, including spectral envelope, F0 (1 dimension) and AP (513 dimensions) are extracted by STRAIGHT analysis [14]. The 39^{th} order MCEPs plus log energy are extracted to represent spectral envelop.The HMM-based Speech Synthesis System (HTS) [17] is used to implement general TTS training, in which acoustic features together with their delta and delta-delta are modeled by multi-stream HMMs, and each phone HMM has a five-state topology with single Gaussian, diagonal covariance distributions.

We implement our proposed and reference approaches using different corpora, thus bringing the following five systems:

- **TW**: TTS-WH system, as a benchmark, is trained on 1,000 English utterances from WH.
- **TBWM**: TTS-BDL-WH-Mandarin is a cross-lingual system, which combines TTS-BDL and a DBLSTM model trained on 100 Mandarin utterances from WH.
- **TBWE**: TTS-BDL-WH-English is a intra-lingual system, as a reference with upper bound performance for comparison with TBWM, which combines TTS-BDL and a DBLSTM model trained on 100 English utterances from WH.
- **TSWM**: TTS-SLT-WH-Mandarin is a cross-lingual system, which combines TTS-SLT and a DBLSTM model trained on 100 Mandarin utterances from WH.
- **TSWE**: TTS-SLT-WH-English is a intra-lingual system, as a reference with upper bound performance for comparison with TSWM, which combines TTS-SLT and a DBLSTM model trained on 100 English utterances from WH.

For the four cross-lingual TTS approaches (TBWM, TBWE, TSWM and TSWE), the SI-ASR system is implemented using the Kaldi speech recognition toolkit [18] with TIMIT corpus [15]. The system has a DNN architecture with 4 hidden layers each of which contains 1024 units. The dimension of MFCC features we use is 13. Senones are treated as the phonetic class of PPGs. The total number of senones is 131 after clustering in training stage 1. The training time is about 11 hours under the hardware configuration of dual Intel Xeon E5-2640, 8 cores, 2.6GHZ.

After obtaining PPGs from SI-ASR model and MCEPs from STRAIGHT analysis, a DBLSTM framework is adopted to map the relationships between them. The machine learning library CURRENNT [19] is used for the implementation. The number of units in each layer is [131 64 64 64 64 39] respectively, where each hidden layer contains one forward LSTM layer and one backward LSTM layer. BPTT is used to train this model with a learning rate of 1.0×10^{-6} and a momentum of 0.9. It takes about 4 hours for 100 sentences training set in the support of a NVIDIA Tesla K40 GPU.

4.3. Evaluations

We conduct both objective and subjective evaluations on all the five systems. Natural human recording (denoted as RW) represents the upper bound in performance. The benchmark (TW) is a general TTS system that is trained directly on 1,000 English utterances from the target speaker while the other four systems (TBWM, TBWE, TSWM, TSWE) follow our proposed framework and each of them is trained using only 100 Mandarin utterances from the target speaker, i.e. only one-tenth the size of the training data in the benchmark. We expect that our proposed framework may achieve similar performance to the benchmark.

4.3.1. Objective Measure

Mel-cepstral distortion (MCD) is conducted to measure how close the synthesized speech is to RW. MCD is the Euclidean distance between the MCEPs of the synthesized speech and those of RW, denoted as

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{N}(c_d - c_d^g)^2} \qquad (2)$$

where N is the dimension of MCEPs (excluding the energy feature). c_d^g and c_d are the d^{th} coefficient of the generated MCEPs and the RW respectively.

Thirty sentences are randomly selected for testing. Dynamic time warping is used to align the MCEPs of RW and those of the five systems. As shown in table 2, the average MCD values for our proposed systems (TBWM and TSWM) are larger than that for the benchmark system (TW). It is reasonable because TW is directly trained using English utterances from the target speaker while TBWM and TSWM are cross-lingual TTS systems and they are trained using even fewer utterances from the target speaker. Comparison TBWM and TBWE shows they have the similar performance, which indicates that cross-lingual tasks can be well handled by our proposed approach.

Table 2: Average MCD of the five systems. TW is the benchmark. TBWM and TSWM are our proposed systems. TBWE and TSWE are the reference systems.

System	TW	TBWM	TBWE	TSWM	TSWE
MCD	5.09	6.59	6.35	6.54	6.37

4.3.2. *Subjective Tests*

The Mean Opinion Score (MOS) test and ABX preference test are conducted for subjective evaluation of the naturalness and similarity of the synthesized speech with the target speaker's speech. Ten sentences from each of the five systems are randomly selected for testing. We invite 20 listeners to participate in the subjective tests[1].

In the MOS test, the listeners are asked to rate the naturalness of the synthesized speech on a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Results are shown in Fig. 3, which indicates our proposed approach (TBWM and TSWM) achieves satisfactory performance which is comparable to the benchmark of TW.

Figure 3: MOS test results with the 95% confidence intervals. Natural human recording is shown as RW. TW is the benchmark trained on much more data from the target speaker. TBWM and TSWM are our proposed systems based on TTS-BDL and TTS-SLT engines respectively. TBWE and TSWE are the reference systems.

For the ABX preference test, the listeners are asked to choose which of A and B (generated by two different systems) sounds more like the target speaker's recording X. No preference is also recorded. Each pair of A and B are presented in randomized order to avoid preferential bias. As shown in Fig. 4, these four systems (TW, TBWM, TSWM and TBWE) are nearly equally preferred.

[1]The synthesized speech samples can be found at `http://www.se.cuhk.edu.hk/~lfsun/IS2016`

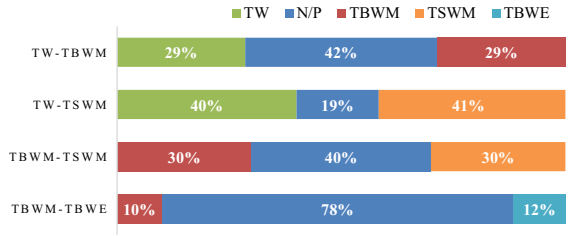

Figure 4: ABX preference test results. N/P stands for no preference. TW is the benchmark. TBWM and TSWM are our proposed systems. TBWE is the reference system.

Results from both the MOS test and the ABX test show that our proposed cross-lingual TTS systems (TBWM and TSWM) can achieve comparable performance to the benchmark (TW) in both speech quality and speaker similarity. The similar performance of TBWM and TSWM systems using different TTS engines suggests that our proposed cross-lingual technique can well be applicable to arbitrary general TTS engine in the target language. Comparison TBWM and TBWE also shows our proposed approach can well handle cross-lingual tasks. We can see there is a big difference in the N/P percentage between these four pairs. High N/P percentage (e.g. TBWM-TBME) means all the listeners consider there is no perference for the paired samples. Low N/P percentage (e.g. TW-TSWM) means some listeners consider A (e.g. TW) is better than B (e.g. TSWM) while some listeners consider B is better than A, but there is no preference in the statistical view. Compare to objective results, subjective results are more reliable because unavoidable alignment errors exist in objective measures.

5. Conclusions

This paper presents a novel approach to personalized, cross-lingual TTS. PPGs obtained from an SI-ASR are regarded as a bridge across speakers and language boundaries. Therefore, a DBLSTM model is trained using PPGs of the target speaker's data in a non-target language and the corresponding acoustic features, and this trained model can be driven to generate the target speaker's speech in the target language by feeding the PPGs of synthesized speech output from any general TTS in the target language for arbitrary input text. This approach has a very low training data requirement of the target speaker's speech recordings. In addition, it can very easily be applied to synthesizing a fixed target speaker's speech in any language by just inserting an arbitrary TTS in that language.

Experiments take a GMM-HMM-based TTS trained on a bilingual speaker's speech in the target language as a benchmark. Evaluation results show that the proposed system trained on this bilingual speaker's speech in the non-target language (e.g. 100 Mandarin utterances) is comparable to the benchmark (trained on 1,000 English utterances) in both speech quality and speaker similarity.

6. Acknowledgements

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14205814).

7. References

[1] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin–English) TTS," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1231–1239, 2009.

[2] Y. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Interspeech*, 2009, pp. 528–531.

[3] H. Wang, F. K. Soong, and H. Meng, "A spectral space warping approach to cross-lingual voice transformation in HMM-based TTS," in *ICASSP*, 2015, pp. 4874–4878.

[4] F. Xie, F. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *ICASSP*, 2016.

[5] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[6] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign TTS," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[7] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *ICASSP*, 2011.

[8] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language tts," in *Interspeech*, 2012.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*, 2016.

[10] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional Long Short-Term Memory based Recurrent Neural Networks," in *ICASSP*, 2015.

[11] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009.

[12] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Interspeech*, 2011.

[13] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.

[16] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[17] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on Hidden Markov Models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," Dec. 2011.

[19] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: the Munich open-source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.