PHONETIC POSTERIORGRAMS FOR MANY-TO-ONE VOICE CONVERSION WITHOUT PARALLEL DATA TRAINING

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

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

ABSTRACT

This paper proposes a novel approach to voice conversion with non-parallel training data. The idea is to bridge between speakers by means of Phonetic PosteriorGrams (PPGs) obtained from a speaker-independent automatic speech recognition (SI-ASR) system. It is assumed that these PPGs can represent articulation of speech sounds in a speaker-normalized space and correspond to spoken content speaker-independently. The proposed approach first obtains PPGs of target speech. Then, a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM) structure is used to model the relationships between the PPGs and acoustic features of the target speech. To convert arbitrary source speech, we obtain its PPGs from the same SI-ASR and feed them into the trained DBLSTM for generating converted speech. Our approach has two main advantages: 1) no parallel training data is required; 2) a trained model can be applied to any other source speaker for a fixed target speaker (i.e., many-to-one conversion). Experiments show that our approach performs equally well or better than state-of-the-art systems in both speech quality and speaker similarity.

Index Terms— voice conversion, phonetic posteriorgrams, non-parallel, many-to-one, SI-ASR, DBLSTM

1. INTRODUCTION

Voice conversion (VC) aims to modify the speech of one speaker to make it sound as if it were spoken by another specific speaker. VC can be widely applied to many fields including customized feedback of computer-aided pronunciation trimming systems, development of personalized speaking aids for speech-impaired subjects, movie dubbing with various persons' voices, etc.

Typical VC training works as follows: speech segments (e.g., frames) with the same spoken content are aligned first. Then, the mapping from source acoustic features to target acoustic features is found. Many previous efforts on VC rely on parallel training data in which speech recordings come in pairs by the source speaker and the target speaker uttering the same sentences. Stylianou et al. [1] proposed a continuous probabilistic transformation approach based on Gaussian Mixture Models (GMMs). Toda et al. [2] improved the performance of GMM-based method by using global variance to alleviate the over-smoothing effect. Wu et al. [3] proposed a non-negative matrix factorization-based method to use speech exemplars to synthesize converted speech directly. Nakashika et al. [4] used a Deep Neural Network (DNN) to map the source and target in high order space. Sun et al. [5] proposed a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM)-based approach to model the relationships between source and target speeches by using spectral features and their context information.

All the above approaches provide reasonably good results. However, in practice, parallel data is not easily available. Hence, some researchers proposed approaches to VC with non-parallel data, which is a more challenging problem. Most of these approaches focused on finding proper frame alignments that is not so straightforward. Erro et al. [6] proposed an iterative alignment method to pair phonetically equivalent acoustic vectors from non-parallel utterances. Tao et al. [7] proposed a supervisory data alignment method, where phonetic information was used as the restriction during alignment. Silén et al. [8] extended a dynamic kernel partial least squares regression-based approach for non-parallel data by combining it with an iterative alignment algorithm. Benisty et al. [9] used temporal context information to improve the iterative alignment accuracy of non-parallel data.

Unfortunately, the experimental results [6–9] show that the performance of VC with non-parallel data is not as good as that of VC with parallel data. This outcome is reasonable because it is difficult to make non-parallel alignment as accurate as parallel alignment. Aryal et al. [10] proposed a very different approach that made use of articulatory behavior estimated by electromagnetic articulography (EMA). With the belief that different speakers have the same articulatory behavior (if their articulatory areas are normalized) when they speak the same spoken content, the authors took normalized EMA features as a bridge between the source and target speakers. After modeling the mapping between EMA features and acoustic features of the target speaker, VC can be

achieved by driving the trained model with EMA features of the source speaker.

Our approach is inspired by [10]. However, instead of EMA features which are expensive to obtain, we use easily accessible Phonetic PosteriorGrams (PPGs) to bridge between speakers. A PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class for each specific time frame of one utterance [11, 12]. Our proposed approach generates PPGs by employing a speaker-independent automatic speech recognition (SI-ASR) system for equalizing speaker differences. Then, we use a DBLSTM structure to model the mapping between the obtained PPGs and the corresponding acoustic features of the target speaker for speech parameter generation. Finally, we perform VC by driving the trained DBLSTM model with the source speaker's PPGs (obtained from the same SI-ASR). Note that we are not using any underlying linguistic information behind PPGs from SI-ASR in VC. Our proposed approach has the following advantages: 1) no parallel training data is required; 2) no alignment process (e.g., DTW) is required, which avoids the influence of possible alignment errors; 3) a trained model can be applied to any other source speakers as long as the target speaker is fixed (as in many-to-one conversion). But for the state-of-the-art approach with parallel training data, a trained model is only applicable to a specific source speaker (as in one-to-one conversion).

The rest of the paper is organized as follows: Section 2 introduces a state-of-the-art VC system that relies on parallel training data as our baseline. Section 3 describes our proposed VC approach with PPGs. Section 4 presents the experiments and the comparison of our proposed approach against the baseline in terms of both objective and subjective measures. Section 5 concludes this paper.

2. BASELINE: DBLSTM-BASED APPROACH WITH PARALLEL TRAINING DATA

The baseline approach is based on a DBLSTM framework which is trained with parallel data [5].

2.1. Basic Framework of DBLSTM

As shown in Fig. 1, DBLSTM is a sequence to sequence mapping model. The middle section, the left section and the right section (marked with "t", "t-1" and "t+1" respectively) stand for the current frame, the previous frame and the following frame respectively. Each square in the Fig. 1 represents one memory block, which contains self-connected memory cells and three gate units (i.e., input, output and forget gates) that can respectively provide write, read and reset operations. Furthermore, bidirectional connections of each layer can make full use of the context information in both forward and backward directions.

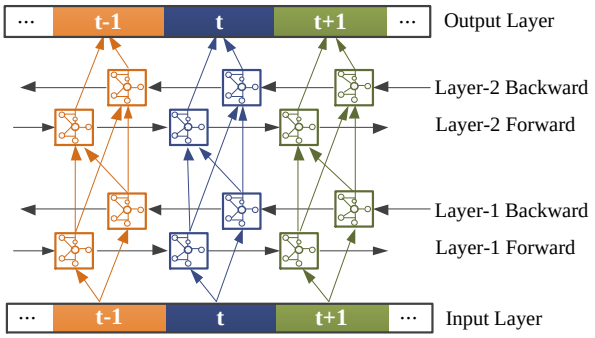

Fig. 1. Architecture of DBLSTM.

The DBLSTM network architecture including memory blocks and recurrent connections makes it possible to store information over a longer period of time and to learn the optimal amount of context information [5, 13].

2.2. Training Stage and Conversion Stage

The baseline approach is divided into training stage and conversion stage as illustrated in Fig. 2.

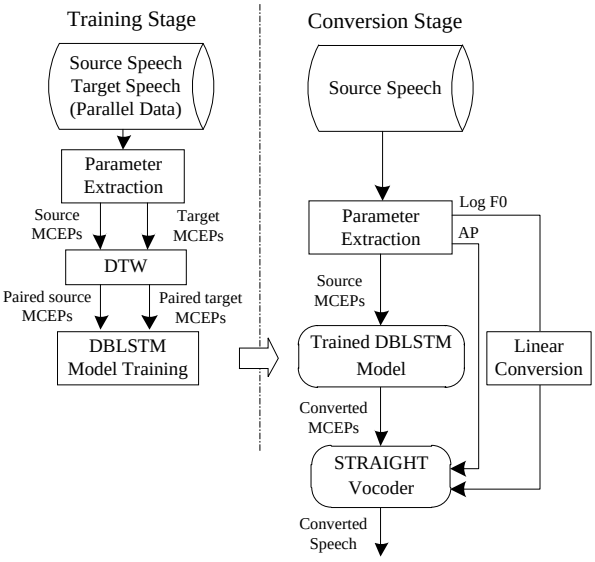

Fig. 2. Schematic diagram of the DBLSTM-based approach for VC with parallel training data.

In the training stage, the spectral envelope is extracted by STRAIGHT analysis [14]. Mel-cepstral coefficients (MCEPs) [15] are extracted to represent the spectral envelope and then MCEPs features from the same sentences of the source speech and the target speech are aligned by dynamic time warping (DTW). Then, paired MCEPs features of the source and target speeches are treated as the training data. Back-propagation through time (BPTT) is used to train DBLSTM model.

In the conversion stage, fundamental frequency (F0), MCEPs and an aperiodic component (AP) are extracted for one source utterance first. Then, parameters of the converted speech are generated as follows: MCEPs are mapped by the trained DBLSTM model. Log F0 is converted by equalizing the mean and the standard deviation of the source and target speeches. AP is directly copied. Finally, the STRAIGHT vocoder is used to synthesize the speech waveform.

2.3. Limitations

Despite its good performance, the DBLSTM-based approach has the following limitations: 1) it relies on parallel training data which is expensive to collect; 2) the influence of DTW errors on VC output quality is unavoidable.

3. PROPOSED APPROACH: VC WITH PHONETIC POSTERIORGRAMS (PPGS)

To solve the limitations of the baseline approach, we propose a PPGs-based approach with the belief that PPGs obtained from an SI-ASR system can bridge across speakers.

3.1. Overview

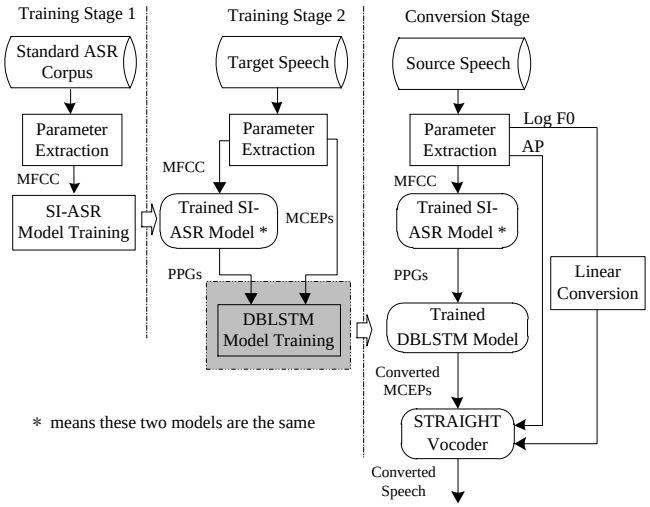

Fig. 3. Schematic diagram of VC with PPGs. SI stands for speaker-independent. Target speech and source speech do not have any overlapped portion. The shaded part will be presented in Fig. 5.

As illustrated in Fig. 3, the proposed approach is divided into three stages: training stage 1, training stage 2 and the conversion stage. The role of the SI-ASR model is to obtain a PPGs representation of the input speech. Training stage 2 models the relationships between the PPGs and MCEPs features of the target speaker for speech parameter generation. The conversion stage drives the trained DBLSTM model with PPGs of the source speech (obtained from the same SI-ASR) for VC. The computation of PPGs and the three stages will be presented in the following subsections.

3.2. Phonetic PosteriorGrams (PPGs)

A PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class for each specific time frame of one utterance [11, 12]. A phonetic class may refer to a word, a phone or a senone. In this paper, we treat senones as the phonetic class. Fig. 4 shows an example of PPG representation for the spoken phrase "particular case".

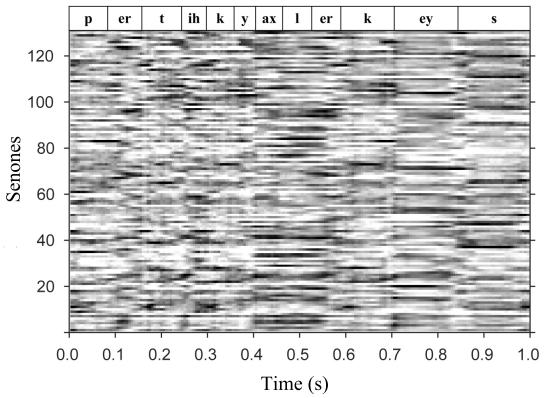

Fig. 4. PPG representation of the spoken phrase "particular case". The horizontal axis represents time in seconds and the vertical one contain indices of phonetic classes. The number of senones is 131. Darker shade implies a higher posterior probability.

We believe that PPGs obtained from an SI-ASR can represent articulation of speech sounds in a speaker-normalized space and correspond to speech content speaker-independently. Therefore, we regard these PPGs as a bridge between the source and the target speakers.

3.3. Training Stages 1 and 2

In training stage 1, an SI-ASR system is trained for PPGs generation using a multi-speaker ASR corpus. The equations are illustrated by the example of one utterance. The input is the MFCC feature vector of t^{th} frame, denoted as $\mathbf{X}_t$. The output is the vector of posterior probabilities $\mathrm{P}_t = (p(s|\mathbf{X}_t)|s = 1, 2, \cdots, C)$, where $p(s|\mathbf{X}_t)$ is the posterior probability of each phonetic class s.

As shown in Fig. 5, training stage 2 trains the DBLSTM model (speech parameter generation model) to get the mapping relationships between the PPG and the MCEPs sequence. For a given utterance from the target speaker, t denotes the frame index of this sequence. The input is the PPG $(\mathrm{P}_1, \cdots, \mathrm{P}_t, \cdots, \mathrm{P}_N)$, computed by the trained SI-ASR model. The ideal value of the output layer is the

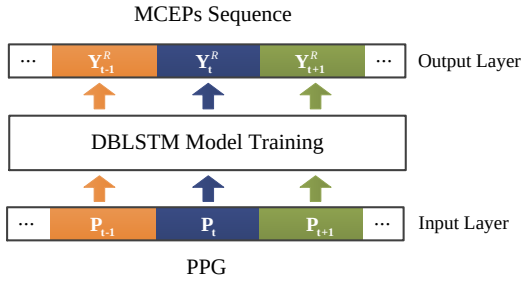

Fig. 5. Schematic diagram of DBLSTM model training.

MCEPs sequence $(\mathbf{Y}_1^T, \cdots, \mathbf{Y}_t^T, \cdots, \mathbf{Y}_N^T)$, extracted from the target speech. The actual value of the output layer is $(\mathbf{Y}_1^R, \cdots, \mathbf{Y}_t^R, \cdots, \mathbf{Y}_N^R)$. The cost function of training stage 2 is

$$min \sum_{t=1}^{N} ||\mathbf{Y}_t^R - \mathbf{Y}_t^T||^2 \quad (1)$$

The model is trained to minimize the cost function through the BPTT technique mentioned in Section 2. Note that the DBLSTM model is trained using only the target speaker's MCEPs features and the speaker-independent PPGs without using any other linguistic information.

3.4. Conversion Stage

In the conversion stage, the conversion of log F0 and AP is the same as that of the baseline approach. First, to get the converted MCEPs, MFCC features of the source speech are extracted. Second, PPGs are obtained from the trained SI-ASR model where the input is MFCC features. Third, PPGs are converted to MCEPs by the trained DBLSTM model. Finally, the converted MCEPs together with the converted log F0 and AP are used by the vocoder to synthesize the output speech.

4. EXPERIMENTS

4.1. Experimental Setup

The data we use for VC is the CMU ARCTIC corpus [16]. The within-gender conversion experiment (male-to-male: BDL to RMS) and the cross-gender conversion experiment (male-to-female: BDL to SLT) are conducted. The baseline approach uses parallel speech of the source and target speakers while our proposed approach uses only the target speaker's speech for model training.

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 39th order MCEPs plus log energy are extracted to represent the spectral envelope.

Two systems are implemented for comparison:

- **Baseline system**: DBLSTM-based approach with parallel training data. Two tasks: male-to-male (M2M) conversion and male-to-female (M2F) conversion.
- **Proposed PPGs system**: Our proposed approach uses PPGs to augment the DBLSTM. Two tasks: male-to-male (M2M) conversion and male-to-female (M2F) conversion.

In the PPGs-based approach, the SI-ASR system is implemented using the Kaldi speech recognition toolkit [17] with the TIMIT corpus [18]. The system has a DNN architecture with 4 hidden layers each of which contains 1024 units. Senones are treated as the phonetic class of PPGs. The number of senones is 131, which is obtained by clustering in training stage 1. Hardware configuration of the SI-ASR model training is dual Intel Xeon E5-2640, 8 cores, 2.6GHZ. The training time is about 11 hours.

Then, the DBLSTM model is adopted to map the relationships of PPGs sequence and MCEPs sequence for speech parameter generation. The implementation is based on the machine learning library, CURRENNT [19]. 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. The training process of DBLSTM model is accelerated by a NVIDIA Tesla K40 GPU and it takes about 4 hours for 100 sentences training set.

The baseline DBLSTM-based approach has the same model configuration except that its input has only 39 dimensions (instead of 131). It takes about 3 hours for 100 sentences training set.

4.2. Objective Evaluation

Mel-cepstral distortion (MCD) is used to measure how close the converted is to the target speech. MCD is the Euclidean distance between the MCEPs of the converted speech and the target speech, denoted as

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

where N is the dimension of MCEPs (excluding the energy feature). c_d and $c_d^{converted}$ are the d-th coefficient of the target and converted MCEPs respectively.

To explore the effect of the training data size, all the systems are trained using different amounts of training data – 5, 20, 60, 100 and 200 sentences. For the baseline approach, the training data consists of parallel pairs of sentences from the source and target speakers. For the proposed approach, the training data consists only of the sentences from the target speaker. The test data set has 80 sentences from the source speaker.

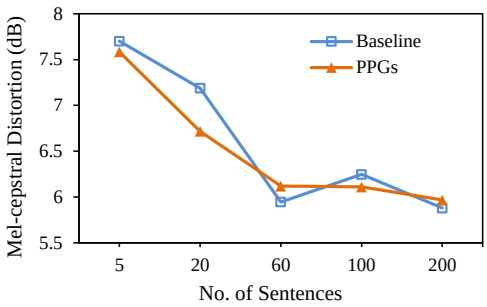

Fig. 6. Average MCD of baseline and proposed PPGs approaches. Male-to-male conversion experiment.

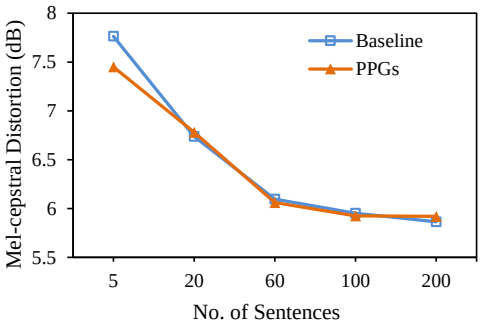

Fig. 7. Average MCD of baseline and proposed PPGs approaches. Male-to-female conversion experiment.

Fig. 6 and Fig. 7 show the results of male-to-male and male-to-female experiments respectively. As shown, when the training size is at 5, 20 and 60 sentences, the MCD value becomes smaller with the increase of the data size. The MCD value tends to converge when the training size is larger than 60 sentences. The results indicate that the baseline approach and the proposed approach have similar performance in terms of objective measure.

4.3. Subjective Evaluations

We conducted a Mean Opinion Score (MOS) test and an ABX preference test as subjective evaluations for measuring the naturalness and speaker similarity of converted speech. 100 sentences are used for training each system and 10 sentences (not in the training set) are randomly selected for testing. 21 participants are asked to do MOS test and ABX test. The questionnaires of these two tests and some samples are presented in `https://sites.google.com/site/2016icme/`.

In the MOS test, listeners are asked to rate the naturalness and clearness of the converted speech on a 5-point scale. The results of the MOS test are shown in Fig. 8. The average scores of the baseline and proposed PPGs-based approaches are 3.20 and 3.87 respectively.

For the ABX preference test, listeners are asked to choose which of the converted utterances A and B (generated by the two approaches) sounds more like the target speaker's recording X or no preference. Each pair of A and B are shuffled to avoid preferential bias. As shown in Fig. 9, PPGs-based approach is frequently preferred over the baseline approach.

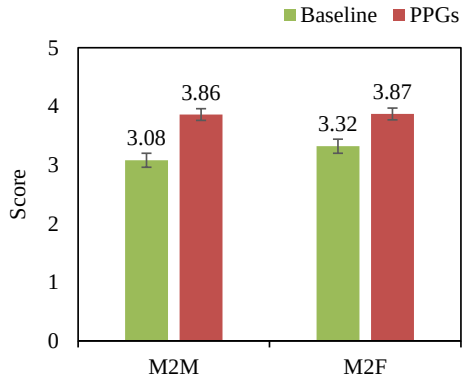

Fig. 8. MOS test results with the 95% confidence intervals. M2M: male-to-male experiment. M2F: male-to-female experiment. 5-point scale: 5: excellent, 4: good, 3: fair, 2: poor, 1: bad.

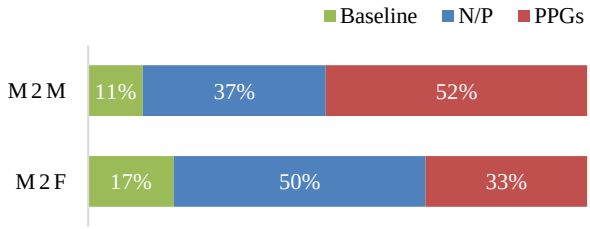

Fig. 9. ABX preference test results. N/P stands for no preference. M2M: male-to-male experiment. M2F: male-to-female experiment. The p-values of the two experiments are 2.94×10^{-16} and 4.94×10^{-3} respectively.

Results from both MOS test and ABX test show that our proposed PPGs-based approach perform better than the baseline approach in both speech quality and speaker similarity. Possible reasons include: 1) Proposed PPGs-based approach does not require alignment (e.g., DTW), which avoids the influence caused by possible alignment errors; 2) the DBLSTM model of the proposed approach is trained using only the speaker-normalized PPGs and the target speaker's acoustic features. This minimizes the interference from the source speaker's signal.

5. CONCLUSIONS

In this paper, we propose a PPGs-based voice conversion approach for non-parallel data. PPGs, obtained by an SI-ASR model, are used to bridge between the source and target speakers. The relationships between PPGs and acoustic

features are modeled by a DBLSTM structure. The proposed approach does not require parallel training data and is very flexible for many-to-one conversion, which are the two main advantages over the approach for voice conversion (VC) using parallel data. Experiments suggest that the proposed approach improves the naturalness of the converted speech and its similarity with target speech.

We have also tried applying our proposed model into cross-lingual VC and have obtained some good preliminary results. More investigation on the cross-lingual applications will be conducted in the future.

6. ACKNOWLEDGEMENTS

The work is partially supported by a grant from the HKSAR Government's General Research Fund (Project Number: 14205814)

7. REFERENCES

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[4] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using Deep Belief Nets," in *Proc. Interspeech*, 2013.

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

[6] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[7] J. Tao, M. Zhang, J. Nurminen, J. Tian, and X. Wang, "Supervisory data alignment for text-independent voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 932–943, 2010.

[8] H. Silén, J. Nurminen, E. Helander, and M. Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression," *Convergence*, vol. 1, p. 2, 2013.

[9] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *Proc. ICASSP*, 2014.

[10] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with Deep Neural Networks," in *Proc. Interspeech*, 2015.

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

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

[13] M. Wollmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. ICASSP*, 2013.

[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] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, 1983.

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

[17] 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.

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

[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.