ACCENT CONVERSION USING PHONETIC POSTERIORGRAMS

Guanlong Zhao[1], Sinem Sonsaat[2], John Levis[2], Evgeny Chukharev-Hudilainen[2], Ricardo Gutierrez-Osuna[1]
[1]Department of Computer Science and Engineering, Texas A&M University, USA
{gzhao, rgutier}@tamu.edu
[2]Department of English, Iowa State University, USA
{sonsaat, jlevis, evgeny}@iastate.edu

ABSTRACT

Accent conversion (AC) aims to transform non-native speech to sound as if the speaker had a native accent. This can be achieved by mapping source spectra from a native speaker into the acoustic space of the non-native speaker. In prior work, we proposed an AC approach that matches frames between the two speakers based on their acoustic similarity after compensating for differences in vocal tract length. In this paper, we propose an approach that matches frames between the two speakers based on their phonetic (rather than acoustic) similarity. Namely, we map frames from the two speakers into a phonetic posteriorgram using speaker-independent acoustic models trained on native speech. We evaluate the proposed algorithm on a corpus containing multiple native and non-native speakers. Compared to the previous AC algorithm, the proposed algorithm improves the ratings of acoustic quality (20% increase in mean opinion score) and native accent (69% preference) while retaining the voice identity of the non-native speaker.

***Index Terms*—**speech synthesis, accent conversion, frame pairing, posteriorgram, acoustic model

1. INTRODUCTION

Learners who acquire a second language (L2) after a “critical age” [1] usually speak with a non-native accent. This may result in lower intelligibility [2] and speakers can be subjected to discriminatory attitudes [3]. Therefore, L2 learners interacting with native speakers have much to gain by improving their pronunciation. Several studies [4, 5] have suggested that having a suitable native speaker to imitate – a so-called “golden speaker,” can be beneficial in pronunciation training. Felps et al. [6] suggested that such a “golden speaker” could be created by resynthesizing the L2 learner’s own voice but with a native accent.

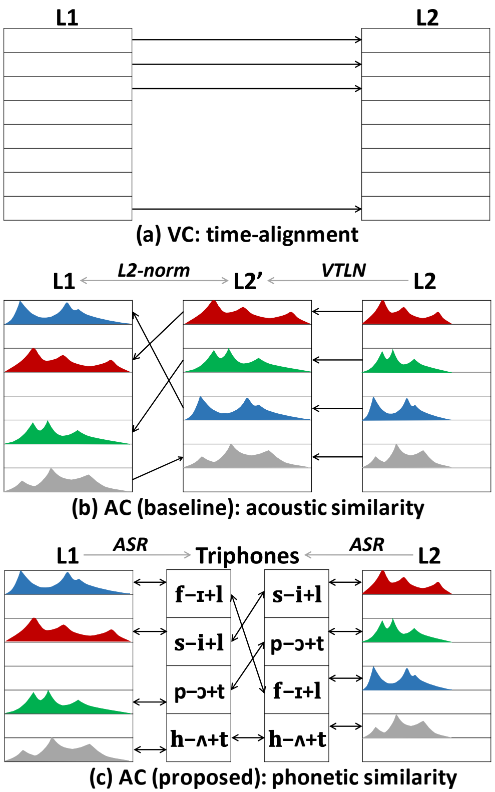
Traditional voice-conversion (VC) methods [7-10] cannot be used for this purpose since they cannot decouple the speaker’s voice quality from their pronunciation, i.e., they assume that pronunciation is part of the speaker’s identity. To address this issue, Aryal and Gutierrez-Osuna [11] proposed a modified VC method, where source frames (i.e., from the native speaker) and target frames (i.e., from the L2 learner) were paired based on their *acoustic* similarity. In a first step, the authors apply vocal-tract length normalization (VTLN) to the source speech, so it matches the target’s vocal-tract length. Then, each frame in the source corpus is paired with the closest frame in the target corpus, and vice versa. Though VTLN did improve frame pairing (e.g., compared to forced alignment), vocal-tract length is just one of potentially many differences between two speakers, and it is too coarse to account for differences in pronunciation.

To address this issue, we present an approach that matches source and target frames based on their phonetic content. Leveraging advances in acoustic modeling [12], we extract phonetic information from the posteriorgram [13]. In a first step, we compute the posteriorgram for each source and target speech frame through a speaker-independent acoustic model trained on native speech. Then, we use the symmetric Kullback-Leibler (KL) divergence in the posteriorgram space to match source and target frames. The result is a set of source-target frames that are aligned based on their *phonetic* similarity. In a final step, we use the frame pairs to train a GMM that models the joint distribution of source and target Mel-Cepstral Coefficients (MCEPs), then map source MCEPs into target MCEPs using maximum likelihood estimation of spectral parameter trajectories considering the global variance of the target speaker.

Relation to prior work. Previous AC methods directly modify speech features that carry accent information, including prosody, formants, spectral envelopes, or articulatory gestures [6, 14-16]. In contrast, our approach uses VC techniques to capture the voice identity of the L2 learner while preserving the native speaker’s pronunciation characteristics – both segmental and prosodic. Unlike VC methods, however, we avoid the issue of time aligning source and target utterances (which is problematic when the target speaker is non-native). Our approach is related to Xie et al. [17], who used speaker-adaptive acoustic models to generate posteriorgram for VC. In contrast with their work, however, we use the posteriorgram to correct mispronunciations and reduce non-native accents, and we focus on speaker-independent acoustic models so we can measure phonetic similarity without retraining the acoustic models.

2. LITERATURE REVIEW

Existing methods for AC can be broadly categorized into acoustic-based and articulatory-based. Among acoustic methods, Yan et al. [18] used a VoiceMorph software to change the trajectories of the formants, pitch, and duration of speech to convert between three different English accents. Huckvale and Yanagisawa [15] blended the spectral envelope



(a) VC: time-alignment

(b) AC (baseline): acoustic similarity

(c) AC (proposed): phonetic similarity

Fig. 1: (a) conventional VC approach; the L1 and L2 frames are time-aligned following their ordering in the data; (b) the AC baseline that uses acoustic similarity through VTLN to pair frames; (c) the proposed AC algorithm that uses phonetic similarity to pair frames

of non-native speech with its native counterpart through voice morphing to reduce the accent. Other acoustic methods for AC have used spectral-envelope vocoders [6] and voice morphing [16].

An alternative to acoustic methods is to consider articulatory gestures. Along these lines, Felps et al. [14] used an articulatory synthesizer based on unit-selection to replace mispronounced L2 diphones with those from the L2 corpus that matched the articulatory configuration of a reference utterance from a native speaker. Later, Aryal and Gutierrez [19] used GMMs and DNNs [20] to map an L2 speaker's articulatory gestures into acoustics, then drove the GMM/DNN with articulatory gestures from a reference utterance from a native speaker.

3. METHODS

Conventional *voice* conversion methods use time alignment to pair frames from source and target utterances, see Fig. 1 (a). As such, a VC model trained from time-aligned frame pairs will retain the L2 speaker's accent. Instead, to perform *accent* conversion, the pairing must be based on the phonetic similarity between source and target frames.

3.1. Frame pairing based on phonetic similarity

Accordingly, our proposed approach uses the phonetic posteriorgram to pair acoustic frames from the source and target speakers. Our rationale is simple: if a speech recognizer trained on native speech data determines that an L2 speech segment $\boldsymbol{y}$ is close to the native speech production of a particular phoneme, then it is reasonable to pair up $\boldsymbol{y}$ with an native speech segment $\boldsymbol{x}$ with the same phonetic label. See Fig. 1 (c).

Our approach works as follows. In a first step, we compute a feature vector of phonetic posteriors for speech frames from the two speakers:

$$\mathcal{L}_{x_i} = [P(l_1|\boldsymbol{x_i}), P(l_2|\boldsymbol{x_i}), \dots, P(l_V|\boldsymbol{x_i})] \tag{1}$$

where $\boldsymbol{x_i}$ is the acoustic feature vector of the ith speech frame; $V = \{l_1, l_2, \dots, l_V\}$ is the predefined senone set; $P(l_j|\boldsymbol{x_i})$ is the conditional probability that the speech frame belongs to senone l_j given $\boldsymbol{x_i}$; $\sum_j P(l_j|\boldsymbol{x_i}) = 1$.

We compute phonetic posteriors using DNN acoustic models, the state-of-the-art for large vocabulary continuous speech recognition tasks. Namely, we adopt a p-norm DNN [21] with multiple p-norm and normalization layers between inputs and outputs. Each p-norm layer uses the p-norm non-linearity followed by a normalization layer that scales down all dimensions of its input in order to stop the average squared output from exceeding one. Inputs to the DNN consist of concatenated MFCC frames $\boldsymbol{X}$, whereas target outputs $\boldsymbol{Y}$ are senones obtained from forced-alignment using a pre-trained acoustic model. After the DNN is fine-tuned using Stochastic Gradient Descent, we compute the posterior probability of a senone using the softmax non-linearity:

$$p(l_j|\boldsymbol{X}_i) = \frac{\exp(x_j')}{\sum_k \exp(x_k')} \tag{2}$$

where x_k' is the output of the hidden layer that precedes the softmax layer. Additional details on DNN acoustic modeling may be found in [12].

Given phonetic posterior feature vectors $\mathcal{L}_{x_i}$ and $\mathcal{L}_{x_j}$, we calculate their distance using the symmetric KL divergence,

$$D\left(\mathcal{L}_{x_i}, \mathcal{L}_{x_j}\right) = \left(\mathcal{L}_{x_i} - \mathcal{L}_{x_j}\right) \cdot \left(\log \mathcal{L}_{x_i} - \log \mathcal{L}_{x_j}\right) \tag{3}$$

For each source (i.e., native) frame $\boldsymbol{x_i}$ we find its closest target (i.e., L2) frame $\boldsymbol{y_j^*}$ as,

$$\boldsymbol{y_j^*} = \underset{\forall \boldsymbol{y}}{\operatorname{argmin}} D\left(\mathcal{L}_{x_i}, \mathcal{L}_{y}\right) \tag{4}$$

Likewise, for each L2 frame $\boldsymbol{y_j}$ we find its closest native frame $\boldsymbol{x_i^*}$,

$$\boldsymbol{x_i^*} = \underset{\forall \boldsymbol{x}}{\operatorname{argmin}} D\left(\mathcal{L}_{x}, \mathcal{L}_{y_i}\right) \tag{5}$$

The resulting frame pairs are used to train a GMM.

3.2. Baseline methods for frame pairing

We compared the proposed posteriorgram method against two baseline techniques for frame pairing: the acoustic similarity method of Aryal and Gutierrez-Osuna [11], and dynamic time warping (DTW).

Baseline 1. Following [11], we measured acoustic similarity as the inverse of the L2-norm between the source and target speaker, after normalizing the source speaker to match the vocal tract length of the L2 speaker; see Fig. 1 (b).

In a first step, we learn a VTLN transform to reduce physiological differences in vocal tract between the two speakers. For this purpose, we time-align parallel training utterances of the two speakers, each utterance represented as a sequence of MFCCs. Following Panchapagesan and Alwan [22], we then learn a linear transform between the MFCCs of both speakers using ridge regression:

$$T^* = \underset{T}{\operatorname{argmin}} \|\boldsymbol{x} - T\boldsymbol{y}\|^2 \tag{6}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of MFCCs from the native and L2 speakers, respectively, and T^* is the VTLN transform. Next, for each native vector $\boldsymbol{x_i}$ we find its closest L2 vector $\boldsymbol{y_j^*}$ as:

$$\boldsymbol{y_j^*} = \underset{\forall \boldsymbol{y}}{\operatorname{argmin}} \|\boldsymbol{x_i} - T^*\boldsymbol{y}\|^2 \tag{7}$$

We repeat the process for each L2 vector $\boldsymbol{y_i}$ to find its closest match $\boldsymbol{x_i^*}$:

$$\boldsymbol{x_j^*} = \underset{\forall \boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x} - T^*\boldsymbol{y_i}\|^2 \tag{8}$$

The above process results in a lookup table where each native and L2 frame in the database is paired with the closest one from the other speaker.

Baseline 2. As our second baseline method, we use DTW to time-align source and target frames.

3.3. Spectral conversion

To ensure a fair comparison among the three frame-pairing methods, we use a common spectral conversion technique to map a source speaker's spectral features to match a target speaker. Namely, we use a GMM to model the joint distribution of source and target frame pairs and then use maximum likelihood parameter generation (MLPG) that also considers global variance of the target speaker to generate the converted speech given a testing speech signal from the source speaker. Additional details may be found in [7].

3.4. Pitch scaling

Previous studies [6, 15, 18] have shown that prosody modification is an essential part of accent conversion. Following Toda et al. [7], we use the pitch trajectory from the source (native) speaker, which captures native intonation patterns, then normalize it to match the pitch range of the target (L2) speaker using mean and variance normalization in log F_0 space.

4. EXPERIMENT SETUP

4.1. DNN acoustic model

We obtained a pre-trained DNN acoustic model via Kaldi's [23] online archive[1]. The model is a p-norm DNN with 18 hidden layers. The input features are computed from a 13-dim MFCC vectors with a 9-frame context; the concatenated 117-dim (13 × 9) MFCCs are passed through a Linear Discriminant Analysis to generate a 40-dim input feature vector. The output layer has 5,816 nodes that correspond to the senones. The DNN acoustic model was trained on Librispeech's [24] training set, which contains 960 hours of native English speech. For more details about the DNN acoustic model we used, please refer to [25].

4.2. Speech corpus

For the native speech corpus, we used two speakers from the CMU ARCTIC dataset [26]: BDL (male) and CLB (female). For the non-native (L2) English speech corpus, we collected recordings from five speakers: two native Hindi speakers (RRBI, male; TNI, female), two native Korean speakers (HKK and YKWK, both male); and one native Arabic speaker (ABA, male). Each L2 speaker produced the full ARCTIC dataset. For each AC direction, we used 100 parallel utterances for training and 50 utterances for testing; there was no overlap between the two sets.

4.3. System configuration

We used STRAIGHT [27] to decompose speech into aperiodicity (AP), F_0, and a 513-dim spectral envelope. Then, we computed 25 MFCCs from the spectral envelopes to learn the VTLN transform and pair up frames using acoustic similarity. We also computed 25 MCEPs from the spectral envelopes as the acoustic feature (excluding MCEP_0) to train the GMMs and convert speech from the native speaker to the L2 speaker. Following our prior work [11], all GMMs had 128 mixture components with diagonal covariance matrices. Once we converted native MCEPs to the L2's space, we reconstructed the spectrogram from the converted MCEPs and combined it with the native's AP and normalized F_0 to synthesize speech.

We considered five speaker pairings for accent conversion: BDL to RRBI, BDL to HKK, BDL to YKWK, BDL to ABA, and CLB to TNI. For each pairing, we performed accent conversion on all 50 testing utterances.

5. RESULTS

To evaluate the three systems (posteriorgram, two baselines), we conducted listening studies on Mechanical Turk to rate the acoustic quality, speaker identity and native accent of the resynthesized speech. All human subjects passed a screening test that consisted of identifying various American English accents. All test samples were randomly ordered. Audio samples are available at: `http://people.tamu.edu/~guanlong.zhao/icassp18_demo.html`

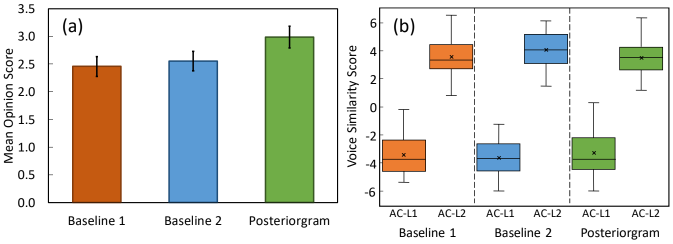


Fig. 2: (a) Speech quality results with 95% confidence interval (b) Speaker identity results (AC-L2: VSS b/w AC and L2 speaker)

[1] http://kaldi-asr.org/models.html

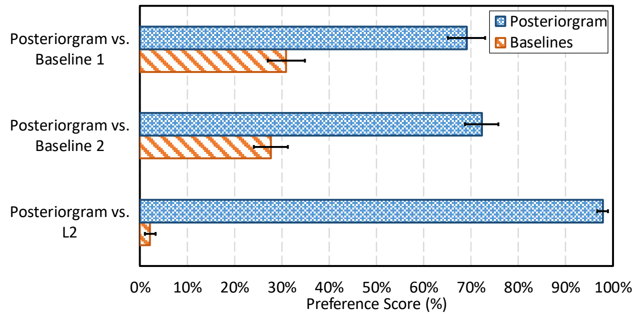


Fig. 3: Accent preference score with 95% confidence interval

Acoustic quality. We used a standard five-point (1-Bad, 5-Excellent) Mean Opinion Score (MOS) to rate the acoustic quality of the synthesized speech. Thirty listeners rated 150 test samples: 50 per system, ten per conversion direction. Results are shown in Fig. 2a. We found no statistical differences between the two baseline systems (2.6 vs. 2.5; $p = 0.43$; two-tailed t-test; null hypothesis: the two averages are the same). The proposed method (posteriorgram) received a 3.0 MOS, which was statistically higher than baseline 1 (20% improvement; $p \ll 0.001$; single-tailed t-test) and baseline 2 (16% improvement; $p \ll 0.001$; single-tailed t-test). These results suggest that the proposed algorithm can boost the acoustic quality of the converted speech significantly using exactly the same training data without even having to modify the GMM training and spectral conversion methods.

Speaker identity. Following [28], we used a voice similarity score (VSS) ranging from -7 (definitely different speakers) to +7 (definitely same speaker) to assess the speaker's identity. Twenty-six participants rated 150 utterance pairs: 50 pairs per system (25 AC-native (L1) and 25 AC-L2 pairs), ten pairs per conversion direction (randomly drawn from the 50 testing utterances). Presentation order within a single utterance pair was counterbalanced. Native (L1) and L2 utterances were resynthesized from their MCEPs to match the acoustic quality of the accent conversions. In addition, and following [6], we played utterances in reverse to prevent the accent from interfering with the perception of voice identity. Results are summarized in Fig. 2b. Overall, the three systems have similar voice similarity scores, and AC-native received a VSS around -3.5, indicating that listeners were "*confident*" that the AC utterances had a different voice identity from those of the native speaker. Likewise, AC-L2 pairs received a VSS around 3.5, indicating that listeners were "*confident*" that the same speaker produced the AC and L2 utterances. We found no statistically-significant differences in VSS between the posteriorgram and baseline methods (AC-native VSS, $p \gg 0.05$; AC-L2 VSS, $p \gg 0.05$; two-tailed t-test; null hypothesis: the averages of the two comparison groups are the same), which shows that the posteriorgram method does not sacrifice the converted speech's voice identity.

Accentedness. In a final experiment, we used a preference test to determine if the posteriogram method does indeed make L2 speech sound more native-like. Thirty native English speakers rated 150 utterance pairs: 50 pairs for each comparison: Posteriorgram vs. Baseline 1, Posteriorgram vs. Baseline 2, and Posteriorgram vs. L2 (i.e., original utterances from the L2 speaker), ten pairs per conversion direction randomly drawn from the 50 testing utterances. The order of the systems within a single comparison pair was counterbalanced; each utterances pair was from the same sentence. Listeners were asked to choose the most native-like (least foreign) utterance from each pair. Aryal and Gutierrez [11] had previously established that Baseline 1 outperforms Baseline 2 and L2 in this task; therefore, we omitted those comparisons in this study. Results are summarized in Fig. 3. On average, listeners were very confident (mean: 98%, STD: 3%) that the Posteriorgram conversions were more native-like than the original L2 utterances. More importantly, listeners were positive that the Posteriorgram method outperformed both Baseline 1 (mean: 69%, STD: 11%) and Baseline 2 (mean: 72%, STD: 10%). All the above preference scores are statistically significant ($p \ll 0.001$; single-tailed t-test) compared with chance levels (50%).

6. CONCLUSION

We have proposed a new frame-pairing method based on the phonetic similarity between acoustic frames. To measure phonetic similarity, we map source and target frames into a phonetic posteriorgram space using speaker-independent acoustic models trained on a native English corpus. Through a series of perceptual studies, we have shown that merely changing the frame pairing method can lead to significant improvement in acoustic quality and "nativeness" while keeping the voice quality of the L2 learner. Our results also show that the approach works well across multiple L2 speakers with different native tongues. Our approach only requires 5-10 minutes of speech data from the L2 learner, making it practical for pronunciation training in realistic settings [29].

A few future directions are worth exploring. At present, pairing speech frames requires computing pairwise symmetric KL divergence for all possible frame combinations in a high-dimensional (posteriorgram) space. Though our implementation was carefully optimized, it is still computationally expensive (it takes about 10 minutes on a high-end desktop to process 100 parallel utterances). Further reductions in compute time may be achieved via dimensionality reduction and clustering. Another future direction is to directly modify the speech waveform [8], which has been shown to reduce over-smoothness in the synthesis. Our ultimate goal is to apply this technique to pronunciation training in classroom settings.

7. ACKNOWLEDGEMENTS

This work was supported by NSF awards 1619212 and 1623750. We would like to thank Dr. Genna Angello for her assistance.

REFERENCES

[1] E. H. Lenneberg, "The biological foundations of language," *Hospital Practice*, vol. 2, no. 12, pp. 59-67, 1967.

[2] M. Munro and T. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73-97, 1995.

[3] D. L. Rubin and K. A. Smith, "Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants," *International Journal of Intercultural Relations*, vol. 14, no. 3, pp. 337-353, 1990.

[4] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors–in search of the golden speaker," *Speech Communication*, vol. 37, no. 3, pp. 161-173, 2002.

[5] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in *Australian International Conference on Speech Science & Technology*, 2006, pp. 24-29.

[6] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920-932, 2009.

[7] 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.

[8] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Interspeech*, 2014.

[9] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 3, pp. 580-587, 2015.

[10] C. Liberatore, S. Aryal, Z. Wang, S. Polsley, and R. Gutierrez-Osuna, "SABR: Sparse, Anchor-Based Representation of the Speech Signal," in *Interspeech*, 2015, pp. 608-612.

[11] S. Aryal and R. Gutierrez-Osuna, "Can Voice Conversion Be Used to Reduce Non-Native Accents?," in *ICASSP*, 2014, pp. 7879-7883.

[12] G. Hinton *et al*., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[13] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 421-426.

[14] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301-2312, 2012.

[15] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," presented at the ISCA Speech Synthesis Workshop, 2007.

[16] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Interspeech*, 2013, pp. 3077-3081.

[17] F.-L. Xie, F. K. Soong, and H. Li, "A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences," in *Interspeech*, 2016, pp. 287-291.

[18] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 676-689, 2007.

[19] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 433-446, 2015.

[20] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with Deep Neural Networks," in *Interspeech*, 2015, pp. 3385-3389.

[21] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014, pp. 215-219.

[22] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer Speech and Language*, vol. 23, no. 1, pp. 42-64, Jan 2009.

[23] D. Povey *et al*., "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition & Understanding*, 2011.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206-5210: IEEE.

[25] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[26] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223-224.

[27] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *ICASSP*, 2008, pp. 3933-3936: IEEE.

[28] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *ICASSP*, 2017, pp. 5525-5529: IEEE.

[29] S. Ding *et al*., "Golden Speaker Builder: an interactive online tool for L2 learners to build pronunciation models," in *PSSLT*, 2017.