Articulatory-based conversion of foreign accents with deep neural networks

Sandesh Aryal, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University

{sandesh, rgutier}@cse.tamu.edu

Abstract

We present an articulatory-based method for real-time accent conversion using deep neural networks (DNN). The approach consists of two steps. First, we train a DNN articulatory synthesizer for the non-native speaker that estimates acoustics from contextualized articulatory gestures. Then we drive the DNN with articulatory gestures from a reference native speaker –mapped to the nonnative articulatory space via a Procrustes transform. We evaluate the accent-conversion performance of the DNN through a series of listening tests of intelligibility, voice identity and nonnative accentedness. Compared to a baseline method based on Gaussian mixture models, the DNN accent conversions were found to be 31% more intelligible, and were perceived more native-like in 68% of the cases. The DNN also succeeded in preserving the voice identity of the nonnative speaker.

Index Terms: articulatory synthesis, deep neural networks, electromagnetic articulography, voice conversion

1. Introduction

Foreign accent conversion [1] seeks to transform utterances from a second language (L2) learner to sound native-like while preserving the voice quality of the learner. This transformation is achieved by transposing accent cues and voice-identity cues between the L2 utterance and that from a native (L1) speaker. Due to the difficulty in decoupling accent and voice-identity cues in the audio signal [2], however, acoustic-based methods for accent conversion often lead to utterances that appear to have been produced by a third speaker, i.e., a morph between the L1 and L2 speakers [1, 3]. To address this issue, in prior work [4, 5] we have shown that articulatory information (e.g., from electromagnetic articulography) may be used to decouple both sources of information to produce accent conversions.

Shown in Figure 1, a typical articulatory-based method for accent conversion consists of building an articulatory synthesizer for the L2 speaker and driving it with normalized articulatory gestures from an L1 speaker. Several techniques may be used to build the articulatory synthesizer in a data-driven fashion, including unit-selection synthesis [4] and statistical parametric synthesis [5]. Statistical techniques tend to be more effective since, unlike unit selection, they can operate with a small L2 corpus and can also interpolate L1 phones that may not exist in the L2 corpus. Accordingly, in recent work [5] we have used the statistical parametric synthesizer of Toda et al. [6]. The approach consists of modeling the joint acoustic-articulatory distribution with a Gaussian mixture, then applying optimization to find the maximum-likelihood trajectory of acoustics features for a given articulatory sequence. This trajectory-optimization stage can substantially improve acoustic quality by reducing spectral discontinuities across adjacent frames, but requires that the entire utterance be processed at once, making it impractical for real-time conversion.

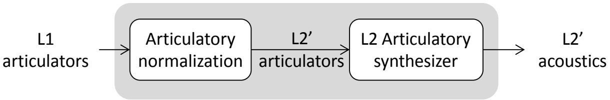


Figure 1: *Articulatory foreign accent conversion*

Here we propose using a deep neural network (DNN) as an articulatory synthesizer to perform accent conversion in real-time. The DNN uses a tapped-delay line to contextualize the input articulatory features in the time domain [7], in this way avoiding the costly trajectory optimization of the conventional GMM synthesizer. We compare the performance of the DNN articulatory synthesizer against a baseline GMM synthesizer [5] through a series of perceptual studies of acoustic quality, voice identity and native accentedness.

The remainder of this paper is structured as follows. Section 2 reviews previous work on accent conversion. Section 3 describes the proposed DNN accent conversion technique and the GMM-based baseline method. Section 4 discusses the experimental setup used to evaluate the DNN accent-conversion method. Results from the perceptual tests are presented in section 5. Finally, section 6 discusses our findings and proposes directions for future work.

2. Related work

Studies have shown that segmental cues are as important for accent perception as prosodic cues in the speech signal [1, 8]. As a step towards modifying both types of cues (segmental and prosodic), in early work we used a vocoding technique to transpose linguistic (e.g., accent) and organic (i.e., voice identity) information from the vocal tract spectra of L1 and L2 utterances [1]. Due to the complex interaction of linguistic and organic information in the acoustic domain, the results often led to the perception of a third speaker, one who shared voice quality characteristics from the L1 and L2 speaker. In later work [4] we suggested performing the accent conversion in the articulatory domain, where a voice-independent representation of linguistic gestures may be readily available. For this purpose, we used a unit-selection framework to replace the most accented portions of the L2 utterance with alternative segments from the L2 corpus based on their articulatory similarity to those from a reference L1 utterance. Although the approach avoided the third-speaker problem, the small corpus size and the lack of native-like units in the L2 corpus led to unreliable synthesis quality.

Unlike unit-selection, statistical parametric synthesizers have low-data requirements and the flexibility to interpolate sounds for previously unseen articulatory gestures [6]. In recent work [5], we performed accent conversion by first building a GMM articulatory synthesizer for an L2 speaker, and then driving the synthesizer with articulatory trajectories from an L1 speaker. In our study, the articulatory data consisted of trajectories for a few critical articulators (e.g., tongue tip, lips) recorded via electromagnetic articulography (EMA). Through a series of subjective listening tests, we showed that driving the L2 synthesizer with L1 articulators led to more intelligible and native-like utterances than driving the L2 synthesizer with the original L2 articulators. As noted in the introduction, however, the method requires an expensive trajectory optimization stage to incorporate the dynamics of acoustic features, making it unsuitable for real-time conversion. Though low-delay implementation of this trajectory optimization step have been proposed [9, 10], this comes at a cost of lower-quality speech synthesis.

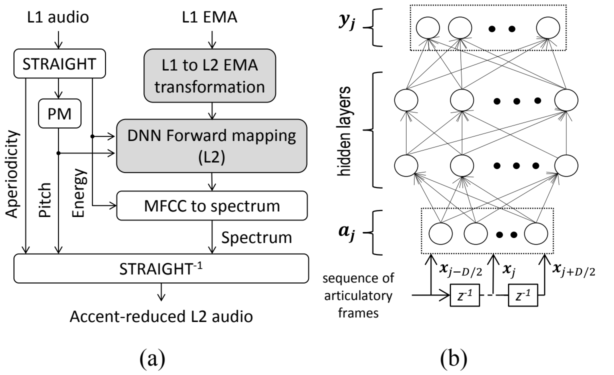
To address this issue, recently we have also proposed a DNN-based articulatory synthesis technique for real-time synthesis that uses a tapped-delay line to contextualize the articulatory trajectory [7]. When compared to a baseline GMM articulatory synthesizer, the DNN reduced the Mel Cepstral distortion by 9.8% *within speaker*. In addition, perceptual evaluation through listening tests rated the DNN synthesis as more natural in 73% of the cases. Here, we examine whether the DNN articulatory synthesizer can also outperform the GMM articulatory synthesizer *across speakers*, as needed for accent conversion.

3. Method

Following our prior work [5], our overall approach for foreign accent conversion consists of four main stages –see Figure 2a: (1) articulatory normalization to map L1 EMA positions into L2 articulatory space, (2) DNN forward mapping to estimate L2 acoustic parameters from normalized L1 EMA positions, (3) scaling of the L1 pitch contour to match the pitch range of the L2 speaker, and (4) reconstructing the speech waveform via STRAIGHT synthesis. In what follows, we provide a brief overview of articulatory normalization, the DNN forward mappings and the baseline GMM forward mappings. For details on the pitch scaling and waveform generation, please refer to [5]

3.1. Cross-speaker articulatory normalization

A set of cross-speaker articulatory mappings are used to transform the EMA articulatory coordinates for the L1 speaker into the equivalent position in the L2 articulatory space. For this purpose, we build a set of Procrustes transforms for each flesh-point using pairs of corresponding articulatory landmarks from both the speakers. Following [11], we use phone-centroids of the EMA positions as the articulatory landmarks. Please refer to [5] for details.



(a) (b)

Figure 2: *(a) DNN-based foreign accent conversion (PM: pitch modification) (b) Forward mapping using a DNN with a tapped-delay line input*

3.2. DNN-based forward mapping

Given a trajectory of articulatory features $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \dots \boldsymbol{x_T}]$ for an utterance, the DNN estimates the corresponding sequence of acoustic feature vectors $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_T]$. As illustrated in Figure 2b, the DNN consists of an input layer, an output layer, and multiple layers of hidden units between them. In this particular topology, units in a layer are fully connected to units in the immediate layer above it, but there is no connection among units within a layer. The network contains a tapped-delay line to contextualize the input with features from past and future frames, resulting in the input vector $\boldsymbol{a}_j = \{\boldsymbol{x}_{j-D/2} \dots \boldsymbol{x}_j \dots \boldsymbol{x}_{j+D/2}\}$, where $\boldsymbol{x}_j$ is the articulatory configuration at frame j, and D is the number of delay units. The DNN consists of Gaussian input units and binary hidden units, all units with sigmoid activation functions since the mapping is likely to be nonlinear.

Training the DNN is a two stage process. First, a Gaussian-Bernoulli Boltzmann machine [12] is trained in an unsupervised fashion. Finally, a layer of output nodes (one node for each acoustic parameter) is added on top of the trained GDBM to form a DNN, which is then fine-tuned via back-propagation [13].

3.3. Global variance adjustment

Statistical mappings are known to over-smooth the acoustic trajectories, resulting in muffled sounds [14]. For this reason, GMM synthesizers generally incorporate the global variance (GV) of the acoustic feature vectors to reduce over-smoothing effects. To ensure a fair comparison with the baseline, we adjust the DNN estimated acoustic features as follows. Let the acoustic feature vector estimated by the DNN at frame j of the test utterance be $\boldsymbol{y}_j$, then, the GV-adjusted feature vector $\hat{\boldsymbol{y}}_j$ is given by:

$$\hat{\boldsymbol{y}}_j = (\boldsymbol{y}_j - \boldsymbol{\mu})\mathbf{A} + \boldsymbol{\mu} \qquad (1)$$

where $\boldsymbol{\mu}$ is the mean of the estimated acoustic feature vectors, and $\boldsymbol{A}$ is a diagonal matrix whose elements are the square roots of the ratios between the GVs for the natural and estimated trajectories. Calculating the exact values for $\boldsymbol{\mu}$ and $\mathbf{A}$ requires the estimated acoustic features for the entire utterance, which is not possible in real-time conversion. Therefore, we calculate these parameters $(\boldsymbol{\mu}, \mathbf{A})$ for all the training sentences and use their average value as an approximation during run-time.

3.4. GMM-based forward mapping

The baseline method [5] uses a GMM to estimate the maximum-likelihood trajectory of acoustic features $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_T]$ for a sequence of articulatory feature vectors $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \dots \boldsymbol{x_T}]$ in a test utterance. The mapping considers the dynamics and the global variance of the acoustic features to estimate the trajectories of acoustics features $\hat{\boldsymbol{y}}$ as:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\text{argmax}}\ P(\boldsymbol{Y}|\boldsymbol{x})^{1/2T}.P(\boldsymbol{v}(\boldsymbol{y})) \qquad (2)$$

where $\boldsymbol{Y} = [\boldsymbol{y_1}, \Delta\boldsymbol{y_1}, \boldsymbol{y_2}, \Delta\boldsymbol{y_2}, \dots \boldsymbol{y_T}, \Delta\boldsymbol{y_T}]$ is the time sequence of acoustic vectors (both static and dynamic) and $\boldsymbol{v}(\boldsymbol{y})$ is the

GV of static acoustic feature vectors. The probability distribution of global variance $P(\boldsymbol{v}(\boldsymbol{y}))$ is modeled using a Gaussian distribution whereas the conditional probability $P(\boldsymbol{Y}|\boldsymbol{x})$ is inferred from the joint probability distribution function $P(\boldsymbol{x}_i, \boldsymbol{y}_i)$ modeled using Gaussian mixtures. For more details, please refer to [5, 6].

4. Experimental

We evaluated the DNN and GMM accent conversion models on an experimental corpus of parallel recordings of articulatory and audio signal from a native and a nonnative speaker of American English [4, 5] collected via Electromagnetic articulography (EMA). Both speakers recorded the same set of 344 sentences, out of which 294 sentences were used for training the model and the remaining 50 sentences were used only for testing. Six standard EMA pellets positions (tongue tip, tongue body, tongue dorsum, upper lip, lower lip, and lower jaw) were recorded at 200Hz. For each acoustic recording, we also extracted aperiodicity, fundamental frequency and the spectral envelop using STRAIGHT analysis [15]. STRAIGHT spectra were sampled at 200Hz to match the EMA recording and then converted into Mel frequency cepstral coefficients (MFCCs). MFCCs were extracted from the STRAIGHT spectrum by passing it through a Mel frequency filter bank (25 filters, 8 KHz cutoff) and then calculating discrete cosine transformation of these filter-bank energies. Following our prior work [5], the articulatory input feature vector consisted of the $x - y$ coordinates for the six EMA pellets, fundamental frequency (log scale), frame energy ($MFCC_0$) and *nasality* (binary feature extracted from the text transcript), while the acoustic feature vector consisted of $MFCC_{1-24}$. The baseline GMMs were trained with 128 mixture components (full covariance), whereas the DNNs contained 2 layers of 512 hidden nodes, and a 60ms tapped-delay input (seven 10-ms frames: 3 previous, 1 current, 3 future). These GMM and DNN structures were found to be suitable for forward mapping in our earlier studies [5, 7].

In order to evaluate the DNN-based accent conversion method, we synthesized test sentences in five experimental conditions –see Table 1: a) the proposed accent conversion method (AC_{DNN}), b) articulatory resynthesis by driving the DNN with L2 articulators ($L2_{EMA}$), c) accent conversion using the baseline GMM-based method (AC_{GMM}), d) MFCC compression of L2 speech ($L2_{MFCC}$), and e) L1 utterances modified to match the vocal tract length [16] and pitch range of L2 ($L1_{GUISE}$). We evaluated these conditions through a series of subjective listening tests on Mturk, Amazon's crowd sourcing tool. To qualify for the study, participants were required to reside in the United States and pass a screening test that consisted of identifying various American English accents, including Northeast, Southern, and General American.

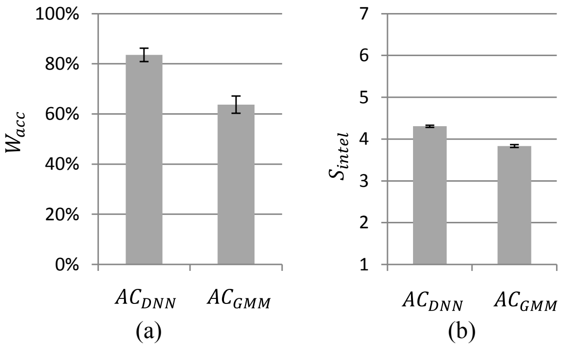


Figure 3: *(a) Word accuracy and (b) subjective intelligibility ratings for AC_{DNN} and AC_{GMM}*

5. Results

5.1. Intelligibility assessment

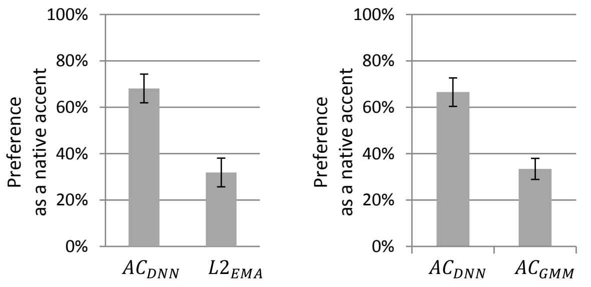
In a first listening test we assessed the intelligibility of the proposed method (AC_{DNN}). We asked a group of participants (N=15) to transcribe 46 test utterances from AC_{DNN}, and also rate the (subjective) intelligibility (S_{intel}) of those utterances using a seven-point Likert scale (1: not intelligible at all, 7: extremely intelligible). From the transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly-identified words to the total number of words in the utterance. To compare the intelligibility of the proposed method against the baseline method, we used the same set of test sentences in our prior study [5]. Figure 3 shows the word accuracy and the subjective intelligibility ratings for the two accent-conversion models (AC_{DNN} and AC_{GMM}). The DNN model had higher scores ($W_{acc} =$ 84%, $S_{intel} =$ 4.3) than the baseline GMM model ($W_{acc} =$ 64%, $S_{intel} =$ 3.84), and the differences were statistically significant (W_{acc}: $t(45) = 7.4, p < 0.001; S_{intel}$: $t(45) = 3.66, p < 0.001$).

5.2. Assessment of non-native accentedness

In a second set of listening tests, we examined the ability of the DNN to reduce the perceived non-native accent of L2 utterances. Following our previous work [5, 17], participants were asked to listen to pairs of utterances –one from the accent conversion (AC_{DNN}) method, the other an articulatory resynthesis of the L2 utterance ($L2_{EMA}$) for the same sentence, and select the most native-like. The articulatory resynthesis ($L2_{EMA}$) was used instead of the original L2 recording to account for losses in acoustic quality due to the articulatory-synthesis step in the accent conversion process, which are known to affect accent perception [1]. As before, we tested on the same subset of 15 test sentences in our prior study [5] so that the results could be compared. Participants listened to 30

Table 1: *Experimental conditions for the listening tests*

Experimental conditions	Aperiodicity and energy	Pitch	Articulators	Spectrum	Forward-mapping model
AC_{DNN}	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping	DNN
$L2_{EMA}$	L2	L2	L2	L2 forward mapping	DNN
AC_{GMM}	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping	GMM
$L2_{MFCC}$	L2	L2	N/A	L2 MFCC	N/A
$L1_{GUISE}$	L1	L1 scaled to L2	N/A	L1 warped to L2	N/A



(a) (b)

Figure 4: *Subjective evaluation of accentedness. Participants selected the most native-like utterances (a) between AC_{DNN} vs. L2 articulatory resynthesis, and (b) between AC_{DNN} vs. AC_{GMM}*

pairs of utterances (15 $AC_{DNN} - L2_{EMA}$ pairs and 15 $L2_{EMA} - AC_{DNN}$ pairs) presented in random order to account for ordering effects. As shown in Figure 4(a), participants rated AC_{DNN} more native-like than L2 articulatory resynthesis in 68% $(s.e = 6\%)$ of the sentences, which is significantly higher $(t(14) = 3.03, p < 0.01)$ than the 50% chance level. *This result shows that the proposed DNN-based method is effective in reducing perceived nonnative accents.*

Next, we compared the DNN accent conversion method against the baseline GMM method. For this purpose, a different group of participants listened to the 30 pairs of utterances (15 $AC_{DNN} - AC_{GMM}$ pairs and 15 $AC_{GMM} - AC_{DNN}$ pairs) presented in random order. Shown in Figure 4(b), AC_{DNN} utterances were rated as more native-like than AC_{GMM} utterances in 67% $(s.e. = 5\%)$ of the sentences, which is also significantly higher than the 50% chance level $(t(14) = 3.6674, p < 0.01)$.

5.3. Voice identity assessment

In a third and final listening experiment we evaluated if the DNN accent-conversion method was able to preserve the voice identity of the L2 speaker. For this purpose, participants were asked to compare the voice similarity between pairs of utterances, one from AC_{DNN}, the other from $L2_{MFCC}$ (MFCC compression of the original L2 recordings). As a sanity check [5], we also included pairs of utterances from $L2_{MFCC}$ and $L1_{GUISE}$, the latter a simple guise of L1 utterances to match the pitch range and vocal tract length of the L2 speaker. Following [1, 5], utterances in each pair were linguistically different, and presentation order was randomized for conditions within each pair and for pairs of conditions. Participants $(N = 15)$ rated 40 pairs, 20 from each group $(L2_{MFCC} - AC_{DNN}$, $L2_{MFCC} - L1_{GUISE})$ randomly interleaved, and were asked to (1) determine if the utterances were from the same or a different speaker and (2) rate how confident they were in their assessment using a seven-point Likert scale (1: not confident at all, 3: somewhat confident, 5: quite a bit confident, and 7: extremely confident). The responses and their confidence ratings were then combined to form a voice similarity score (VSS) ranging from -7 (extremely confident they are different speaker) to $+7$ (extremely confident they are from the same speaker).

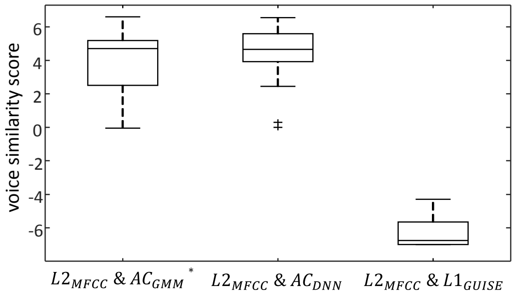


Figure 5: *Average pairwise voice similarity scores (* $L2_{MFCC}$ & AC_{GMM} scores are from* [5])

Figure 5 shows the boxplot of average VSS between the pairs of experimental conditions. Participants were 'quite' confident $(VSS = 4.3, s.e. = 0.5)$ that the $L2_{MFCC}$ and AC_{DNN} were from the same speaker, suggesting that the method successfully preserved the voice-identity of L2 speaker. The VSS was also comparable $(t(14) = -0.37, p = 0.71)$ to the VSS between AC_{GMM} and $L2_{MFCC}$ $(VSS = 4.0, s.e. = 0.5)$ reported for the baseline method in our prior study [5]. The participants were also 'quite' confident that $(VSS = -6.3, s.e. = 0.2)$ the $L2_{MFCC}$ and $L1_{GUISE}$ were from different speakers, corroborating the finding in our prior study [5] that a simple guise of L1 utterances is not sufficient to match the voice of the L2 speaker. These findings suggest that the run-time capabilities of the DNN did not compromise its ability to preserve the voice identity.

6. Discussion

We have presented an articulatory method for real-time modification of non-native accents. The approach uses a DNN with a 60ms tapped-delay input to map L2 articulatory trajectories into L2 acoustic observations (MFCCs). Driving the DNN with articulatory trajectories recorded via EMA from an L1 speaker –normalized to the L2 articulatory space— results in speech that captures the linguistic gestures of the L1 speaker and the voice quality of the L2 speaker.

We evaluated the DNN accent-conversion method against the baseline GMM method in [5]. Accent conversions with the DNN were more intelligible and were perceived as more native-like than those using the GMM. A possible explanation for the difference in perceived accentedness between both methods is that acoustic quality affects the perception of nonnative accents (i.e., the lower the quality, the higher the non-native rating) [1]; although both methods use articulatory synthesis, a recent study [7] shows that the DNN tends to synthesize speech of higher acoustic quality than the GMM.

Additional work is required to validate the approach beyond the specific L1-L2 speaker pair in our study, including nonnative speakers with different levels of proficiency. An interesting new resource in this regard is the Marquette University EMA Mandarin Accented English (EMA-MAE), which contains a large EMA corpus from multiple Mandarin L2 speakers of American English [18]. Future work may also extend this study using the richer articulatory representation provided by real-time magnetic resonance imaging (rt-MRI) [19]. In comparison to EMA, which only captures a few fleshpoints in the frontal oral cavity, rt-MRI provides information about the entire vocal tract, from lips to glottis, which may result in more intelligible and native-like accent conversions. Considering the cost of recording articulatory features, future studies may also evaluate the feasibility of using speaker-independent inverted articulatory features [20] as opposed to the measured EMA positions used in this study.

7. References

[1] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech commun.,* vol. 51, pp. 920-932, 2009.

[2] H. Hermansky and D. J. Broad, "The effective second formant F2' and the vocal tract front-cavity," in *Proceedings of ICASSP*, 1989, pp. 480-483.

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

[4] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Trans. Audio Speech Lang. Process.,* vol. 20, pp. 2301-2312, 2012.

[5] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *J. Acoust. Soc. Am.,* vol. 137, pp. 433-446, 2015.

[6] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.,* vol. 50, pp. 215-227, 2008.

[7] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language,* 2015 (in press).

[8] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *IEEE Trans. Audio, Speech, and Lang. Process.,* vol. 15, pp. 676-689, 2007.

[9] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proceedings of INTERSPEECH*, 2008, pp. 1076-1079.

[10] N. Xingyu, X. Xiang, and K. Jingming, "Low latency parameter generation for real-time speech synthesis system," in *Proceedings of ICME*, 2014, pp. 1-6.

[11] C. Geng and C. Mooshammer, "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," *J. Acoust. Soc. Am.,* vol. 125, pp. 3278-3288, May 2009.

[12] K. H. Cho, T. Raiko, and A. Ilin, "Gaussian-Bernoulli deep Boltzmann machine," in *Proceedings of IJCNN*, 2013, pp. 1-7.

[13] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature,* vol. 323, pp. 533-536, 1986.

[14] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Trans. Audio Speech Lang. Process.,* vol. 15, pp. 2222-2235, 2007.

[15] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proceedings of ICASSP*, 1997, pp. 1303-1306.

[16] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* St. Thomas, U.S. Virgin Islands, 2003, pp. 676-681.

[17] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *Proceedings of ICASSP*, 2014, pp. 7744-7748.

[18] A. Ji, J. Berry, and M. T. Johnson, "The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) Corpus of Acoustic and 3D Articulatory Kinematic Data," in *Proceedings of ICASSP*, 2014, pp. 7769-7773.

[19] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim*, et al.*, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in *Proceedings of INTERSPEECH*, 2011, pp. 837-840.

[20] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *ICASSP*, 2011, pp. 4624-4627.