Foreign Accent Conversion through Voice Morphing

Sandesh Aryal, Daniel Felps, and Ricardo Gutierrez-Osuna

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

{sandesh, dlfelps, rgutier}@cse.tamu.edu

Abstract

We present a voice morphing strategy that can be used to generate a continuum of accent transformations between a foreign speaker and a native speaker. The approach performs a cepstral decomposition of speech into spectral slope and spectral detail. Accent conversions are then generated by combining the spectral slope of the foreign speaker with a morph of the spectral detail of the native speaker. Spectral morphing is achieved by representing the spectral detail through pulse density modulation and averaging pulses in a pair-wise fashion. The technique is validated on parallel recordings from two ARCTIC speakers using both objective and subjective measures of acoustic quality, speaker identity and foreign accent.

Index Terms— voice morphing, accent conversion.

1. Introduction

During the last two decades, a few studies have suggested that it would be beneficial for second language (L2) learners to be able to listen to their own voices producing native-accented speech [1] The rationale behind this proposal is that removing information that is only related to the teacher's voice quality makes it easier for learners to perceive differences between their accented utterances and their ideal accent-free counterparts. As a step towards this goal, we have developed voice-transformation techniques that can be used to synthesize native-accented utterances from their foreign-accented counterpart while preserving the speaker's voice quality [1-3].

In this paper, we propose a morphing technique that generates a continuum of accent-conversions between the learner's productions and those of the teacher. The technique works as follows. First, we decompose the speech spectra into two components: one that captures broad spectral features (i.e., spectral slope), and a second one containing the spectral details (i.e., formant positions). Then, we generate accent morphs by combining the learner's broad spectra with a morph of the spectral detail of both speakers. Generating the morph requires that we establish correspondence between the two detailed spectra. This is achieved by encoding both spectra as a pulse density, and then averaging the position of corresponding pulses.

Morphing accent conversions may serve as a behavioral shaping strategy in computer assisted pronunciation training. In behavioral shaping, the teacher asks the student to compare their utterances against their "best" previous efforts rather than against a separate standard [4]. Using a normative reference can be detrimental early in training, when the student's utterances are very distant from the ideal pronunciation. Instead, by using a "floating" reference (i.e., one that adapts to the performance of the learner), the teacher can provide carefully graded evaluations of the learners' performance and guide them towards the ultimate goal. Likewise, morphing accent conversions during the early stages of learning may be used to produce utterances that have less ambitious prosodic and segmental goals, slowly improving the reference by incorporating the best pronunciation of the learner and higher degrees of morphing towards the teacher's productions.

2. Related work

Morphing techniques have been extensively used for face perception, but are challenging when applied on speech. Whereas facial landmarks are well defined and relatively easy to detect (eyes, mouth, jaw lines, etc.), spectral features in speech (i.e., formant frequencies) are difficult to measure and ill-defined in the case of unvoiced phones. Rather than use formant-tracking techniques, which are notoriously unreliable, a number of methods have been proposed to generate morphs directly from the spectra of two speakers. Slaney et al. [5] generate separate spectrograms for the pitch and broad spectral shape of a sound, and interpolate each channel separately by means of dynamic programming and harmonic alignment, respectively. Pfitzinger [6] also uses dynamic programming to find a frequency warp between two spectra, but in this case the warping is performed on the first-order derivative of the two LP spectral envelopes. Ezzat et al. [7] also use the derivative of the two (log magnitude) spectra but instead employ a technique similar to optical flow to find a correspondence between the two spectra. More recently, Shiga [8] has proposed a method where spectral envelopes are encoded as a distribution of pulses (see Figure 2). In this case, morphing can be performed by pairing individual pulses from the two spectra (according to their order) and then computing the weighted average of each pair. This results in significant time savings as compared to previous methods based on dynamic programming or optical flow. An added advantage of pulse density coding is that, unlike all-pole models such as LPC, it can model spectral zeros accurately.

Our work is related to the problem of voice conversion [9-14]. However, voice conversion seeks to transform utterances from a speaker so they sound as if another speaker had produced them, whereas accent conversion seeks to transform only those features of an utterance that contribute to accent while maintaining those that carry the identity of the speaker. Only a handful of studies have been published on the subject of accent conversion. Yan et al. [15] proposed an accent-synthesis method based on formant warping. First, the authors developed a formant tracker based on HMMs and LPC, and applied it to a corpus containing several regional English accents (British, Australian, and American). Analysis of the formant trajectories revealed systematic differences in the vowel formant space for the three regional accents. Second, the authors re-synthesized utterances by warping formants from a foreign accent onto the formants of a native accent; pitch-scale and time-scale modifications were also applied. An ABX test showed that 75% of the re-synthesized utterances were perceived as having the native accent. Huckvale and

Yanagisawa [16] used an English TTS system to simulate English-accented Japanese utterances; this was achieved by transcribing Japanese phonemes with their closest English counterparts. The authors then evaluated the intelligibility of a Japanese TTS against the English TTS, and against several prosodic and segmental transformations of the English TTS. Their results showed that both segmental and prosodic transformations are required to improve significantly the intelligibility of English-accented Japanese utterances.

Our work differs from Yan et al. [15] in two respects. First, our method uses pulse coding to represent speech spectra, which makes it more robust than formant tracking, particularly for unvoiced segments. Second, we evaluate not only the accentedness but also the perceived speaker identity of the re-synthesized speech. The latter is critical because a successful accent-conversion model should preserve the identity of the foreign-accented speaker. In contrast with Huckvale and Yanagisawa [16], our study is performed on natural speech, and focuses on accentedness and identity rather than on intelligibility; as noted by Munro and Derwing [17], a strong foreign accent does not necessarily limit the intelligibility of the speaker. Finally, unlike these previous methods and our own prior work [1-3], the work presented here allows us to achieve different degrees of accent conversion by virtue of a morphing coefficient, as described next.

3. Methods

3.1. Morphing through pulse density modulation

Our method for morphing accent conversion is based on the pulse density modulation (PDM) technique of Shiga [8]. The PDM technique employs a delta-sigma modulator to convert a log spectral envelope $x(n)$, where n denotes frequency, into a pulse sequence $y(n) = PDM[x(n)]$ as follows:

$$e(n) = x(n) - v_c\, y(n-1) \tag{1}$$

$$r(n) = e(n) - r(n-1) \tag{2}$$

$$y(n) = sign\big(r(n)\big) \tag{3}$$

with initial conditions $r(1) = e(1) = x(1)$ and $y(n) = 0$; the term v_c is the feedback gain of the delta-sigma modulator: $v_c = max(x)$. In turn, the pulse sequence $y(n)$ can be decoded back into a log spectral envelope $\hat{x}(n) = PDM^{-1}[y(n)]$ through the discrete cosine transform (DCT):

$$c(n) = DCT[y(n)] \tag{4}$$

$$c(n) = 0 \quad \forall\ n > k \tag{5}$$

$$\hat{x}(n) = DCT^{-1}[c(n)] \times v_c \tag{6}$$

which acts as a low-pass filter by truncating the DCT expansion with an appropriate cutoff ($k = 100$ in our case.) Thus, given a pair of spectral envelopes $x_1(n)$ and $x_2(n)$, a morphed spectral envelope can be computed by averaging the position of corresponding pulses in the two spectra:

$$x_m(n) = PDM^{-1}\big[(1-\alpha)PDM[x_1(n)] + \alpha PDM[x_2(n)]\big] \tag{7}$$

where the morphing coefficient α $(0 \leq \alpha \leq 1)$ can be used to generate a continuum of morphs between the two spectral envelopes $x_1(n)$ and $x_2(n)$.

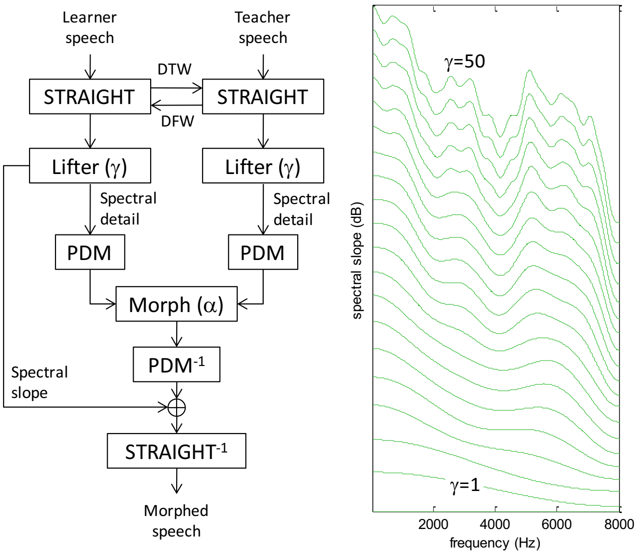

Figure 1*: (a) Morphing accent conversion strategy. (DTW/DFW: dynamic time/frequency warping). (b) Spectral slope $x^L(n)$ as a function of liftering cutoff* $\gamma \in \{1,2,3 \ldots 9,10,12 \ldots 20,25 \ldots 50\}$*. Individual spectra have been shifted vertically for visualization purposes.*

3.2. Accent conversion through voice morphing

Given parallel recordings from the learner $x_1(n)$ and the teacher $x_2(n)$, equation (7) produces a morph of *both* the identity and the accent of the two speakers. In accent conversion, however, we seek to morph only the accent while preserving the learner's identity. For this purpose, prior to the PDM encoding in equations (1-3), each spectra $x_i(n)$ is separated into two components, $x_i^L(n)$ carrying the broad spectral features (i.e., spectral slope) and $x_i^H(n)$ carrying the spectral detail (i.e., formant positions). This, again, is performed by liftering in the DCT domain as:

$$x_i^H(n) = DCT^{-1}\big[DCT\big(x(n)\big) \times l(n)\big] \tag{8}$$

$$x_i^L(n) = DCT^{-1}\big[DCT\big(x(n)\big) \times \big(1 - l(n)\big)\big] \tag{9}$$

where $l(n)$ are the liftering coefficients, defined by:

$$l(n) = \begin{cases} n/\gamma & 1 \leq n \leq \gamma \\ 1 & n > \gamma \end{cases} \tag{10}$$

An accent morph $x_m(n)$ is then produced by combining the learner's broad spectra $x_1^L(n)$ with a morph of the spectral detail of both speakers $x_m^H(n)$:

$$x_m(n) = x_1^L(n) + x_m^H(n) \tag{11}$$

$$x_m^H(n) = PDM^{-1}\Big[\alpha PDM[x_1^H(n)] + (1-\alpha)PDM[x_2^H(n)]\Big] \tag{12}$$

Larger values of the liftering coefficient γ in (10) ensure that increasing spectral detail is preserved in the learner's broad envelope $x_1^L(n)$ and that, likewise, equivalent spectral detail is discarded from the teacher's spectral detail $x_2^H(n)$. The overall accent-conversion process and liftering results for different values of γ are illustrated in Figure 1.

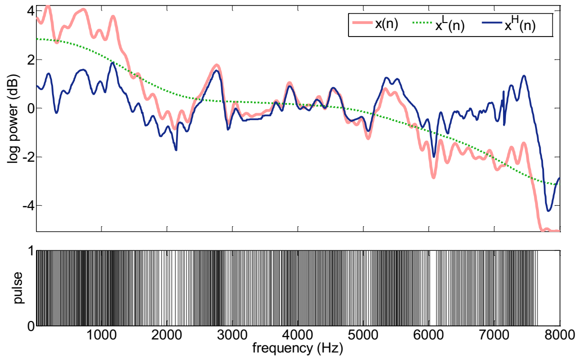

Figure 2: *(a) Decomposition of the spectral envelope $x(n)$ into global shape $x^L(n)$ and spectral detail $x^H(n)$. (b) Encoding of spectral detail $x^H(n)$ through pulse density modulation.*

4. Experimental validation

The proposed method was evaluated on two speakers from the ARCTIC corpus [18]: *ksp_indianmale*, who was treated as the foreign-accented learner, and *rms_usmale2*, who was treated as the native-accented teacher. The STRAIGHT vocoder [19] was used to generate smooth spectrograms and resynthesize the resulting voice morphs. Prior to performing the morphing accent conversions, learner utterances were time-aligned at the frame level (80ms windows, 1ms shift; default frame shift in STRAIGHT) to those of the teacher using dynamic time warping (DTW) and a conventional 39-dimensional feature vector (13 MFCCs, delta and delta-delta features) computed from the STRAIGHT spectrum. To account for differences in vocal tract length, teacher utterances were then frequency warped to those of the target; a global warping function was obtained by applying DTW in the frequency domain [20] to 100 sentences in ARCTIC's "B" set. Finally, utterances were resynthesized using the teacher's pitch contour shifted to the baseline of the learner. As a result of these steps, all subsequent morphs conformed to the timing and pitch dynamics of the teacher, but had the global frequency warp and pitch baseline of the learner.

Morphing accent conversions were generated for parameter values $\gamma \in \{1,2,3 \dots 9,10,12,14 \dots 20,25,30 \dots 50\}$ and $\alpha \in \{0, 0.1, 0.2 \dots 1\}$. One hundred sentences from ARCTIC's "A" set were synthesized for each of these 11×21 combinations, and analyzed in terms of their acoustic quality, speaker identity and foreign accentedness. Three objective measures shown in our earlier work [2] to correlate with listening tests were used for this purpose. Namely, acoustic quality was estimated through the ITU-T recommendation P.563, speaker identity was estimated from a linear discriminant analysis (LDA) of natural utterances from the learner and the teacher, and foreign accent was assessed by the forced-alignment score (log-likelihood) of acoustic models trained on North American speakers using HTK; see [2]. These objective ratings were also verified through subjective listening tests on a subset of the 11×21 combinations.

5. Results

5.1. Objective measures

Figure 3 shows the average performance of the morphing accent conversion in terms of the three objective measures. Acoustic quality improves for higher values of the liftering cutoff γ and low values of the morphing parameter α. This result can be explained as follows. As the value of γ increases, additional spectral structure is retained for the learner's broad envelope $x_1^L(n)$. As a result, the spectral detail $x_i^H(n)$ becomes flatter for large γ, which improves the PDM encoding (i.e., for a spectrum with a significant spectral slope most of the pulses will be placed at the lower frequencies). Overall, however, the result in Figure 3(a) shows that the acoustic quality of the morphed accent conversions remains at an estimated mean-opinion-score (MOS) above 4.7, which in our earlier study [2] corresponds to a perceived MOS of 4.1.

Results from the speaker identity scores are shown in Figure 3(b) in terms of the ratio:

$$ID = \frac{\sum_u \sum_i [d(y_{u,i}, \mu_L)/\sigma_L - d(y_{u,i}, \mu_T)/\sigma_T]}{d(\mu_L, \mu_T)/(\sigma_L + \sigma_T)/2} \tag{13}$$

where $d(\cdot)$ is the Euclidean metric, $y_{u,i}$ is the projection of acoustic frame i in utterance u onto the LDA solution for the two speakers, μ_L, μ_T are the average LDA projection for learner and teacher utterances, respectively, and σ_L, σ_T are their standard deviations. Thus, ID values greater than 0 indicate that the morph is closer to the learner than to the teacher, and vice versa. As shown in Figure 3(b), the morphed accent conversions remain closer to the learner except for a small number of parameter combinations (large α and small γ); the dashed line indicates the maximum-likelihood decision boundary between both speakers. These results are to be expected since for large α the morph is dominated by the target speaker (the teacher) and for small γ only the overall spectral slope of the source speaker (the learner) is preserved.

Results from the accented measure are shown in Figure 3(c) in terms of the HTK forced-alignment score:

$$ACC = \frac{\sum_u \sum_p (s_{u,p} - s_{u,sil})}{N_u N_p} \tag{14}$$

where $s_{u,p}$ is the score (log-likelihood) of phone p on utterance u, N_u is the number of test utterances and N_p is the size of the phone set $(N_u = 100;\ N_p = 39 + sil)$. Subtraction of the silence score $s_{u,sil}$ compensates for misalignment errors. As may be expected, large values of the morphing parameter α reduce the foreign accentedness. In addition, the more information about the learner that is preserved in the spectral slope (i.e., by increasing the liftering cutoff γ), the larger the morphing value will have to be in order to achieve a given accent score. Comparison of Figure 3(b) and Figure 3(c) shows that the accent measure improves (i.e., morphs become more native) faster than the identity measure degrades (i.e., morphs become more like the teacher), which suggests that there is a "sweet spot" where foreign accent reduction can be achieved while preserving the identity of the learner.

5.2. Subjective measures

To verify these objective measures, we ran additional subjective studies on the five selected (α, γ) pairs shown in Table 1 and Figure 3, which represent intermediate degrees of morphing: V1 and V5 being nearest to the learner and the teacher, respectively. For each condition, we transformed the same 10 sentences. Participants performed the following tests through Amazon's Mechanical Turk:

- *Accent* – 10 subjects rated the accentedness of 50 utterances (5 conditions, 10 sentences) using a 7-point scale (0=not at all accented; 2=slightly; 4=quite a bit; 6=extremely). Subjects had to qualify for this test by passing a dialect identification test, which only native speakers of American English are likely to complete.
- *Quality* – 10 subjects rated the quality of the same 50 utterances using a 5-point MOS (1=bad, 2=poor, 3=fair, 4=good, 5=excellent).
- *Identity* – 10 subjects participated in a forced choice test. They listened to two pairs of utterances; each pair consisted of V1 or V5, a separating beep, and one of the intermediate voices (V2-V4). Subjects were asked to select the pair whose voices were more different from each other. The order of presentation was random, and utterances were time-reversed so subjects focused on the physiological components of the voices [2]. Identity scores were calculated as the fraction of times that a given voice was perceived to be closer to V1.

Subjective ratings were consistent with the corresponding objective measures. Accent ratings decrease monotonically through the sequence (V1-V5) and the perceived accent drops suddenly at V3 near a similar drop-off in Figure 3. Quality tended to decrease with γ, as predicted by the objective measure. Due to the design of the test, no identity score is available for V1 or V5. The intermediate voices were considered closer to the learner than to the teacher, and the order of similarity agrees with the objective measures. Original and morphed utterances for ARTIC sentence "*We have plenty of capital ourselves, and yet we want more*" for the five conditions in the listening tests are available at http://research.cs.tamu.edu/prism/publications/is2013.zip.

6. Discussion

We have presented a method for foreign accent conversion that combines a cepstral decomposition of the spectral envelope and a morphing technique through pulse density modulation. Given parallel recordings from a native speaker and a foreign speaker, we decompose the spectral envelope into its overall shape, which captures speaker-dependent cues (i.e. spectral slope), and its spectral detail, which captures linguistic content. The critical step in the morphing process is matching peaks across two spectra. We address this issue by representing the spectral detail as a distribution of a large number of pulses. In this manner, morphing two spectra is equivalent to averaging the position of their pulses in a pair-wise fashion. The overall procedure contains two parameters: a liftering cutoff γ that determines the amount of information to be preserved in the foreign speaker's spectral slope, and a morphing coefficient α that determines the degree of morphing between the spectral detail of both speakers.

The procedure was evaluated on two ARCTIC speakers using objective and subjective measures of acoustic quality, speaker identity and foreign accent. The results indicate that there is a trade-off between quality, identity and accent. Higher quality and identity scores are obtained by retaining as much of the learner's spectral information as possible (large γ and small α) at the expense of reducing accent scores. However, our results also show a region in parameter space where significant reductions in accent are obtained while preserving cues to the learner's identity.

Our approach preserves the pitch and overall vocal tract length of the learner, and assumes that speaker-dependent and linguistic cues in the spectral envelope can be separated through cepstral decomposition (i.e., spectral slope vs. spectral detail, respectively). While F0, vocal tract length and spectral slope are known to be good discriminator among speakers [21], additional acoustic cues from the learner's voice could be captured and preserved before the morphing stage, such as jitter and shimmer [22] and fine structure in the speech signal [23]. Other features from the speaker recognition literature (see [24] for a recent review) may also be investigated while considering that our goal is synthesis rather than recognition. Future work may also investigate filtering techniques (i.e., head-related transfer functions) to reduce differences between speakers' perception of self-produced speech and their speech recordings [25], which may become important in computer assisted pronunciation training.

7. Acknowledgements

This work was supported by NSF award 0713205 and a DoD SMART scholarship. We are grateful to Prof. Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during RGO's sabbatical stay at CSTR (Univ. Edinburgh). We are also grateful to Prof. Kawahara for permission to use the STRAIGHT analysis-synthesis method.

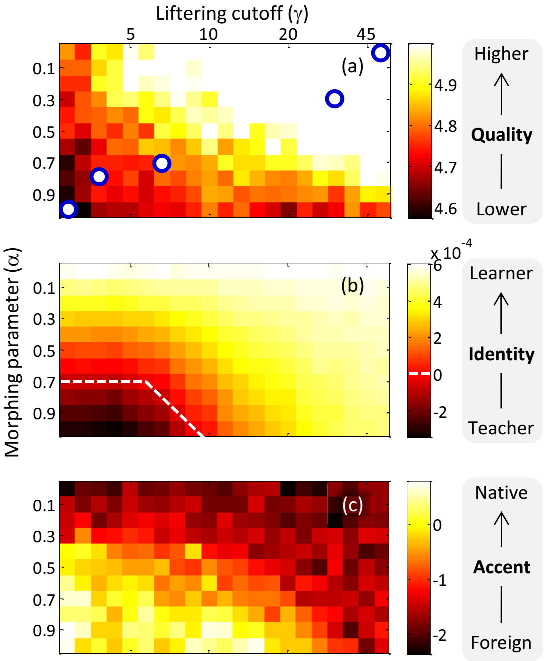

Figure 3: *(a) Quality, (b) identity, and (c) accentedness of the conversions as a function of the liftering cutoff γ and morphing coefficient α. Lighter color denotes desirable effects (e.g., high quality, learner identity, and native accent). Dashed line in (b) represents the maximum-likelihood boundary between both speakers, as measured by LDA. Circles in (a) indicate the five conditions used in the listening tests.*

Table 1. *Subjective ratings of accent, quality and identity.*

(α,γ)	Accent	Quality	Learner's ID
V1 (0,50)	2.94	3.38	n/a
V2 (0.3,30)	2.61	3.41	74%
V3 (0.7,7)	0.40	3.26	69%
V4 (0.8,3)	0.26	3.07	64%
V5 (1,1)	0.14	2.56	n/a

8. References

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

[2] Felps, D. and Gutierrez-Osuna, R., "Developing objective measures of foreign-accent conversion," IEEE Trans.Audio Speech Lang. Process., vol. 18, pp. 1030-1040, 2010.

[3] Felps, D., Geng, C., and Gutierrez-Osuna, R., "Foreign Accent Conversion Through Concatenative Synthesis in the Articulatory Domain," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 2301-2312, 2012.

[4] Watson, C. and Kewley-Port, D., "Advances in computer-based speech training: Aids for the profoundly hearing impaired," Volta-Review, vol. 91, pp. 29-45, 1989.

[5] Slaney, M., Covell, M., and Lassiter, B., "Automatic audio morphing," presented at the ICASSP, 1996.

[6] Pfitzinger, H. R., "Unsupervised speech morphing between utterances of any speakers," in Proc10th Australian Intl Conf Speech Science & Technology, Macquarie University, Sydney, 2004, pp. 545–550.

[7] Ezzat, T., Meyers, E., Glass, J., and Poggio, T., "Morphing Spectral Envelopes Using Audio Flow," in INTERSPEECH, 2005, pp. 2545-2548.

[8] Shiga, Y., "Pulse Density Representation of Spectrum for Statistical Speech Processing," in INTERSPEECH, Brighton, UK, 2009, pp. 1771-1774.

[9] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H., "Voice conversion through vector quantization," in Proc. Intl. Conf. Acoustics, Speech, and Signal Processing, New York, NY 1988, pp. 655-658.

[10] Arslan, L. M. and Talkin, D., "Voice Conversion By Codebook Mapping Of Line Spectral Frequencies And Excitation Spectrum," in Eurospeech '97, Rhodes, Greece, 1997, pp. 1347–1350.

[11] Childers, D. G., Wu, K., Hicks, D. M., and Yegnanarayana, B., "Voice conversion," Speech Communication, vol. 8, pp. 147-158, 1989.

[12] Kain, A. and Macon, M. W., "Spectral voice conversion for text-to-speech synthesis," in Proc. Intl. Conf. Acoustics, Speech, and Signal Processing, 1998, pp. 285-288.

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

[14] Turk, O. and Arslan, L. M., "Robust processing techniques for voice conversion," Computer Speech & Language, vol. 20, pp. 441-467, 2006.

[15] Yan, Q., Vaseghi, S., Rentzos, D., and Ho, C.-H., "Analysis by synthesis of acoustic correlates of British, Australian and American accents," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04), Montreal, Quebec, Canada, 2004, pp. I-637-40.

[16] Huckvale, M. and Yanagisawa, K., "Spoken Language Conversion with Accent Morphing," presented at the Proc. ISCA Speech Synthesis Workshop, Bonn, Germany, 2007.

[17] Munro, M. and Derwing, T., "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," Language Learning & Technology, vol. 45, pp. 73-97, 1995.

[18] Kominek, J. and Black, A., "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University Language Technologies Institute 2003.

[19] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., "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 Commun, vol. 27, pp. 187-207, 1999.

[20] Neuburg, E. P., "Frequency warping by dynamic programming," in ICASSP, 1988, pp. 573-575.

[21] Murthy, H. A., Beaufays, F., Heck, L. P., and Weintraub, M., "Robust text-independent speaker identification over telephone channels," IEEE Trans. Speech Audio Process., vol. 7, pp. 554-568, 1999.

[22] Farrus, M. and Hernando, J., "Using jitter and shimmer in speaker verification," IET Signal Processing, vol. 3, pp. 247-257, 2009.

[23] Jankowski, C. R., Jr., Quatieri, T. F., and Reynolds, D. A., "Measuring fine structure in speech: application to speaker identification," in ICASSP, 1995, pp. 325-328.

[24] Kinnunen, T. and Li, H., "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun, vol. 52, pp. 12-40, 2010.

[25] Shuster, L. I. and Durrant, J. D., "Toward a better understanding of the perception of self-produced speech," J Commun Disord, vol. 36, pp. 1-11, 2003.