Prosodic and segmental factors in foreign-accent conversion

Daniel Felps [1], Heather Bortfeld [2] and Ricardo Gutierrez-Osuna [1]

[1] Dept. of Computer Science, [2] Dept. of Psychology, Texas A&M University, College Station, USA
dlfelps@cs.tamu.edu, bortfeld@psyc.tamu.edu, rgutier@cs.tamu.edu

Abstract

We propose a signal processing method that transforms foreign-accented speech to resemble its native-accented counterpart. The problem is closely related to voice conversion, except that our method seeks to preserve the organic properties of the foreign speaker's voice; i.e., only those features which cue foreign-accentedness are to be transformed. Our method operates at two levels: prosodic and segmental. Prosodic transformation is performed by means of time and pitch scaling. Segmental transformation is performed by convolving the foreign speaker's excitation with the warped spectral envelope of the native speaker. Perceptual results indicate that our model is able to provide a 63% reduction in foreign-accentedness. Multidimensional scaling also shows that the segmental transformation causes the perception of a new speaker to emerge, though the identity of this new speaker is three times closer to the foreign speaker than to the native speaker.

Index Terms: voice conversion, foreign accent, speaker identity

1. Introduction

Voice conversion (VC) seeks to transform utterances from a speaker so they sound as if another speaker had produced them. First proposed by Abe et al. in 1988 [1], numerous techniques have since been spawned for converting a *source* speaker to sound like a *target* speaker. Applications for VC include personalizing text-to-speech synthesizers [2], normalizing features in speech recognition [3], and film dubbing/looping [4]. VC has a well-defined goal: making the source sound like the target. This article addresses a related but orthogonal problem: transforming only those features that contribute to foreign/regional accent while maintaining those that contribute to identity. According to Traunmüller's modulation theory [5], the speech signal can be viewed as the convolution/modulation of a voice-quality carrier with linguistic gestures. From this perspective, accent conversion seeks to extract the voice-quality carrier of the source and convolve it with the linguistic gestures of the target.

Accent conversion has various applications in automatic recognition of foreign-accented speech and second language learning. In particular, accent conversion may enable a new generation of *personalized* computer-assisted pronunciation training tools. To this end, Probst et al. [6] investigated the choice of which native speaker to imitate in pronunciation training. Results from this study showed that users who imitate a well-matched speaker improved their pronunciation more than those who imitated a poor match, suggesting the existence of a user-dependent "golden speaker." Thus, one can argue that accent conversion would provide learners with the optimal "golden speaker": their native-accented selves.

2. FD-PSOLA framework

Our accent-conversion transformation is based on the general framework of Pitch-Synchronous Overlap and Add (PSOLA) [7]. Several versions of PSOLA have been proposed in the literature, including Fourier-domain FD-PSOLA, linear-prediction LP-PSOLA, and time-domain TD-PSOLA [7, 8]. These algorithms perform comparably under modest modification factors, but FD-PSOLA is the most robust to spectral distortion during the pitch modification step. For this reason, and despite its higher computational requirements, FD-PSOLA was adopted for this work. FD-PSOLA operates in three stages: analysis, modification, and synthesis, as described next.

2.1. Analysis

The speech signal is first decomposed into a series of pitch-synchronous short-time analysis windows. Instants of glottal closure are estimated using a pitch-marking algorithm [9]. This stage can greatly affect the quality of the output signal, because pitch marks determine the center and the width of each analysis window. Each analysis window is then framed with a Hanning window, and transformed into the frequency domain. As a result, all pitch-synchronous short-time spectra are represented with the same length (e.g., 2,048 frequencies in our implementation).

2.2. Modification

In this stage, the short-time spectra and their locations are modified to meet the desired pitch and timing (i.e., those of the native speaker, in our case). This modification consists of three steps. First, a new set of synthesis pitch marks are defined according to the target pitch and timing. Second, the short-time spectra are copied (i.e., duplicated or deleted) onto the synthesis pitch marks. Finally, the short-time spectra are transformed to match the new pitch period; since we operate in the frequency domain, this is equivalent to resampling, i.e., spectral compression lowers the pitch and expansion raises it. However, naïve compression of the spectrum also shifts speech formants. For this reason, we first flatten the spectrum with a spectral envelope vocoder (SEEVOC) [10]. We also use a spectral folding technique [11] to regenerate high frequency components that are lost when performing spectral compression (Figure 1(b)).

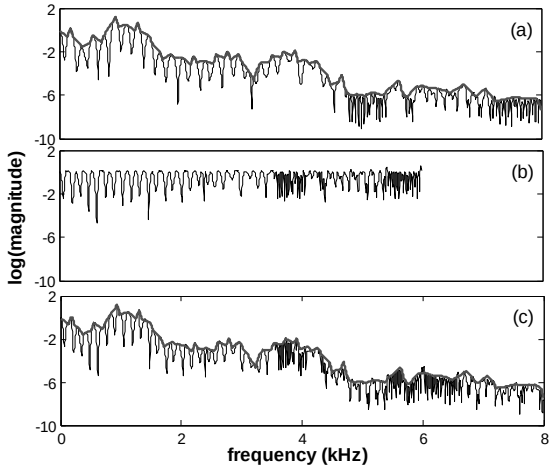


Figure 1 Pitch lowering in the frequency domain. (a) Spectrum of a female vowel /a/ with f_0=188 Hz. (b) The spectrum is flattened and compressed to f_0=141 Hz; notice the spectral hole that occurs at 6-8 kHz. (c) The flattened spectrum in (b) is folded at 6 kHz to fill the hole, and then multiplied by the spectral envelope in (a).

Finally, we multiply the flattened spectrum by the SEEVOC spectral envelope estimate, thus restoring its original resonances.

2.3. Synthesis

The modified short-time spectra are finally transformed back to the time domain, and combined by means of a least-squared-error signal estimation criterion:

$$\hat{x}(n) = \frac{\sum_{m=-\infty}^{\infty} w(m-n) F_m^{-1}(n)}{\sum_{m=-\infty}^{\infty} w^2(m-n)}$$

where $F_m^{-1}(n)$ is the inverse Fourier transform of the short-time spectra at time m and $w(m-n)$ is the windowing function (e.g. Hanning) [12].

3. Accent conversion

For convenience we will use the conventional terminology in voice conversion of *source* vs. *target speaker*; in our case, source will refer to a second-language (foreign) speaker of American English, and target will refer to a native speaker of American English. We will also assume that parallel English utterances are available for both speakers, as well as a common phonetic transcript.

Our accent transformation method proceeds in two distinct steps. First, prosodic conversion is performed by modifying the phoneme durations and pitch contour of the source to imitate those of the target. Second, formants from the source utterance are replaced with those from the target. These two steps are performed simultaneously in our implementation.

3.1. Prosodic conversion

To perform *time-scale* conversion, we first segment the utterances with a forced alignment tool [13]. From these phonetic segments, the ratio of target/source durations is then used to specify a time-scale modification factor α (confined to [0.25, 4]) for the source on a phoneme-by-phoneme basis.

Table 1. Stimulus conditions for the perceptual studies

#	**Stimulus**
1	Foreign utterance
2	Foreign w/ prosodic conversion
3	Foreign w/ segmental conversion
4	Foreign w/ prosodic & segmental conversion
5	Native utterance

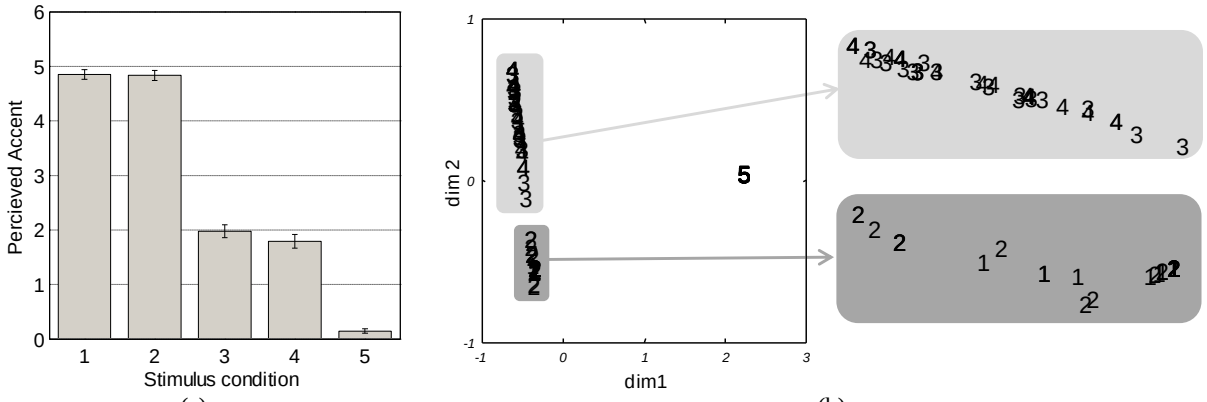
Our *pitch-scale* modification combines the pitch dynamics of the target with the pitch baseline of the source. This is achieved by replacing the pitch contour of the source utterance with a transformed (i.e., shifted and scaled) version of the pitch contour of the target utterance. For this purpose, we first estimate average pitch values for the source ($\overline{f_0^S}$) and target ($\overline{f_0^T}$) from a corpus of utterances. Next, we define a piecewise-linear time-warping (Ψ_{ST}) to align source and target utterances at phoneme boundaries. Finally, given pitch contours (f_0^S and f_0^T) for the specific source and target utterances to be converted, we define a pitch scale as:

$$\beta = \frac{\Psi_{ST}(f_0^T) + \overline{f_0^S} - \overline{f_0^T}}{f_0^S};\ 0.5, < \beta < 2$$

This process allows us to preserve speaker identity by maintaining a reasonable pitch baseline and range [14, 15], while acquiring the pitch dynamics of the target, which provides an important cue to native accentedness [16-18]. Once the time-scale and pitch-scale modification parameters are calculated, standard FD-PSOLA is used to perform the prosodic conversion.

3.2. Segmental conversion

Our segmental accent conversion stage is motivated by the source-filter model [19]. Namely, we assume that the glottal excitation signal is largely responsible for voice quality, whereas the filter contributes to most of the linguistic information. Thus, our goal is to combine the target's spectral envelope (filter) with the source's excitation. FD-PSOLA allows us to perform this step in a straightforward fashion: in the final step of Section 2.2, we multiply the source's flat spectra by the target's spectral envelope rather than by the source's spectral envelope. In order to reduce speaker-dependent information in the target's spectral envelope, we also perform Vocal Tract Length Normalization (VTLN) using a piecewise linear function defined by the average formant pairs of the two speakers [20]. These formant locations are estimated with Praat [21] over the entire corpus. The result is a signal that consists of the source's excitation and the target's warped spectral envelope.



(a) (b)

Figure 2. Experimental results from the perceptual tests. (a) Accent ratings showing mean ± standard error for each stimulus category. Application of the prosodic and segmental transformations yielded a 63% reduction in accent. (b) All 100 stimuli are shown in the perceptual identity space; labels correspond to those in Table 1. Tight clusters may appear as a single point (even in the insets). A third speaker close the source emerges with the segmental transformation.

4. Perceptual experiments

The proposed accent conversion method was evaluated through a series of perceptual experiments. We were interested in determining (1) the degree of foreign accent reduction that could be achieved with the model and (2) the extent to which the transformation preserved the identity of the foreign speaker. To establish the relative contribution of segmental and prosodic information, these two factors were manipulated independently, resulting in three accent conversions: prosodic only, segmental only, and both. Original utterances from both foreign and native speakers were tested as well, resulting in a total of five stimulus conditions (see Table 1). Sample audio files for the five conditions are available as supplemental material (1-5.wav).

A total of 82 participants were recruited from the undergraduate pool maintained by the Department of Psychology at TAMU. All participants were native speakers of American English and had no hearing or language impairments. Audio stimuli were presented via headphones.

Two speakers were selected from the CMU_ARCTIC database [22]: *ksp_indianmale* and *rms_usmale2*. Given that our participants were native speakers of American English, utterances from *ksp_indianmale* were treated as the foreign-accented source, and utterances from *rms_usmale2* were treated as the native-accented target. The same twenty sentences were chosen for each of the five conditions, for a total of 100 unique utterances.

4.1. Foreign accentedness experiment

Thirty-nine students participated in a 25-minute scaled-rating test to establish the degree of accentedness of individual utterances. Following Munro and Derwing [23] participants responded on a 7-point Empirically Grounded, Well-Anchored (EGWA) scale (0=not at all accented; 2=slightly accented; 4=quite a bit accented; 6=extremely accented) [24]. Each participant rated all 100 utterances.

4.2. Identity experiment

Forty-three students participated in a 25-minute speaker identification test. Following Kreiman and Papcun [25], participants heard two linguistically different utterances presented consecutively, and were instructed to focus on the organic aspects of the voice. Participants were asked to determine if the two sentences were produced by the same speaker or by two different speakers, and also to rate their confidence on a 7-point EGWA scale (0=not at all confident; 2=slightly confident; 4=quite a bit confident; 6=extremely confident). These two responses were then converted into a 15-point perceptual distance metric from 0 (extremely confident the speakers were the same) to 14 (extremely confident the speakers were different). Each participant listened to 60 pairs of utterances, uniformly distributed across all possible pairings (5×5) in Table 1.

5. Results

Results from the foreign-accentedness experiment are summarized in Figure 2(a). Original recordings from the source received the highest average accent rating (4.85), while target recordings had the lowest average rating (0.15). The prosodic transform decreased the source's accent slightly (4.83), but this change was not statistically significant. On the other hand, the segmental transform lowered the rating to 1.97, or a 59% reduction in foreign accentedness. When used in concert, both transformations yield an average score of 1.79, or a 63% reduction. Both of these reductions were statistically significant.

To examine the results of the identity experiment, we analyzed the pair-wise perceptual distances with ISOMAP [26]. For this purpose, we first created a (100×100) distance matrix that contained the average perceptual distance between any two of the 100 utterances. This matrix is sparse due to the small number of participants relative to the number of combinations. To guard against outliers, we eliminated those pairs for which only one sample was

available. An ε-neighborhood with a radius of 7 units was used to define the local connectivity of the geodesic graph (scores of 0-7 indicates pairs of utterances that the subjects believed to have been produced by the same speaker). Results are shown in Figure 2(b). Samples from conditions 1 and 2 map close together in the manifold, which indicates that the prosodic transformation only had a small effect on the perceived identity of the speakers. On the other hand, participants were able to clearly distinguish between source utterances (1) and their segmental transformation (3 and 4), which indicates that they perceived the latter as a third speaker; this type of inference is not possible with the ABX tests commonly used in voice conversion. Nonetheless, by calculating the average Euclidean distance across conditions we find that this "third speaker" is perceived to be three times closer to the source than to the target. Interestingly, the ISOMAP embedding can be interpreted in terms of the source-filter theory; the first dimension separates samples in condition 5, which uses the target excitation, from samples in the remaining conditions, which use the source excitation; the second dimension separates samples in conditions 1-2, which employ the source filter, from samples in conditions 3-5, which employ the target filter.

6. Discussion

We have proposed a new method for accent conversion. The method is based on the assumption that accent is contained in the prosody and formant structure, whereas speaker identity is captured by vocal tract length and glottal shape. The method employs FD-PSOLA to adapt the speaking rate and pitch of the source towards those of the target, and a segmental transformation to replace the spectral envelope of the source with that of the target.

Our method has been validated using two perceptual tests designed to quantify foreign accent and speaker identity. The results show that the method provides a 63% reduction in foreign accentedness while preserving much of the voice quality of the original speaker. The emergence of a third speaker suggests a strong relationship between identity and accent; this issue deserves further investigation.

Huckvale and Yanagisawa have recently proposed a similar method for accent morphing [27]. The authors simulated foreign-accented speech by means of a text-to-speech (TTS) system. Namely, they used an English TTS to generate Japanese utterances by replacing Japanese phonemes with their closest English counterparts. The authors evaluated the intelligibility of these utterances against those from a Japanese TTS system, and several prosodic and segmental transformations of the former. Their results show that both segmental and prosodic transformations are required to significantly improve the intelligibility of the English-accented Japanese utterances. In contrast, our study uses natural speech and evaluates the effect of both transformations on the perceived accentedness and identity of the speaker. Thus, results from both studies can be considered to be complementary.

7. Acknowledgment

Hart Blanton is greatly acknowledged for his suggestions regarding the EGWA scale for the perceptual ratings. These experiments were performed in his laboratory, for which we are also "6: extremely grateful."

8. References

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

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

[3] L. Qiguang and C. Chiwei, "Normalizing the vocal tract length for speaker independent speech recognition," *IEEE Signal Processing Letters*, vol. 2, pp. 201-203, 1995.

[4] O. Turk and L. M. Arslan, "Subband Based Voice Conversion," in *Proc. 7th Intl. Conf. Spoken Language Processing*, 2002, pp. 289-292.

[5] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: a modulation theory," *Phonetica*, vol. 51, pp. 170-183, 1994.

[6] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors - In search of the golden speaker," *Speech Communication*, vol. 37, pp. 161-173, 2002.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453-467, 1990.

[8] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175-205, 1995.

[9] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. Intl. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. I-349-I-352.

[10] D. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, pp. 786-794, 1981.

[11] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. Intl. Conf. Acoustics, Speech, and Signal Processing*, 1979, pp. 428-431.

[12] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*: Prentice Hall PTR, 2001.

[13] Sphinx, "SphinxTrain: building acoustic models for CMU Sphinx," Carnegie Mellon University, 2001.

[14] A. J. Compton, "Effects of filtering and vocal duration upon identification of speakers, aurally," *Journal of the Acoustical Society of America*, vol. 35, pp. 1748-1755, 1963.

[15] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 23, pp. 176-182, 1975.

[16] B. Vieru-Dimulescu and P. B. d. Mareüil, "Contribution of prosody to the perception of a foreign accent:a study based on Spanish/Italian modified speech," in *Proc. ISCA Workshop on Plasticity in Speech Perception* London, UK, 2005, pp. 66-68.

[17] J. M. Munro, "Non-segmental factors in foreign accent: ratings of filtered speech," *Studies in Second Language Acquisition*, vol. 17, pp. 17-34, 1995.

[18] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, pp. 28-40, 1997.

[19] G. Fant, *Acoustic theory of speech production*. s'Gravenhage: Mouton, 1960.

[20] D. Sundermann, H. Ney, and H. Hoge, "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.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.5.15)," Universiteit van Amsterdam, Institute of Phonetics Sciences (IFA), 2007.

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

[23] M. J. Munro and T. M. Derwing, "Evaluations of foreign accent in extemporaneous and read material," *Language Testing*, vol. 11, pp. 253-266, 1994.

[24] B. Pelham and H. Blanton, *Conducting Research in Psychology, Measuring the Weight of Smoke*, 3rd ed. Belmont, CA: Thomson Higher Education, 2007.

[25] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, pp. 265-275, 1991.

[26] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.

[27] M. Huckvale and K. Yanagisawa, "Spoken Language Conversion with Accent Morphing," in *Proc. ISCA Speech Synthesis Workshop* Bonn, Germany, 2007, pp. 64-70.