Published in final edited form as:
Speech Commun. 2009 October ; 51(10): 920–932. doi:10.1016/j.specom.2008.11.004.

Foreign accent conversion in computer assisted pronunciation training

Daniel Felps[a], **Heather Bortfeld**[b], and **Ricardo Gutierrez-Osuna**[a,*]

Daniel Felps: dlfelps@cs.tamu.edu; Heather Bortfeld: bortfeld@psyc.tamu.edu; Ricardo Gutierrez-Osuna: rgutier@cs.tamu.edu

[a] Department of Computer Science, Texas A&M University, 3112 TAMU, College Station, TX 77843-3112, USA

[b] Department of Psychology, Texas A&M University, 3112 TAMU, College Station, TX 77843-3112, USA

Abstract

Learners of a second language practice their pronunciation by listening to and imitating utterances from native speakers. Recent research has shown that choosing a well-matched native speaker to imitate can have a positive impact on pronunciation training. Here we propose a voice-transformation technique that can be used to generate the (arguably) ideal voice to imitate: the own voice of the learner with a native accent. Our work extends previous research, which suggests that providing learners with prosodically corrected versions of their utterances can be a suitable form of feedback in computer assisted pronunciation training. Our technique provides a conversion of both prosodic and segmental characteristics by means of a pitch-synchronous decomposition of speech into glottal excitation and spectral envelope. We apply the technique to a corpus containing parallel recordings of foreign-accented and native-accented utterances, and validate the resulting accent conversions through a series of perceptual experiments. Our results indicate that the technique can reduce foreign accentedness without significantly altering the voice quality properties of the foreign speaker. Finally, we propose a pedagogical strategy for integrating accent conversion as a form of behavioral shaping in computer assisted pronunciation training.

Keywords

Voice conversion; Foreign accent; Speaker identity; Computer assisted pronunciation training; Implicit feedback

1. Introduction

Despite years or decades of immersion in a new culture, older learners of a second language (L2) typically speak with a so-called “foreign accent,” sometimes despite concerted efforts at improving pronunciation. Similar learning phenomena have been observed in the animal world: a critical period exists beyond which animals cannot learn certain behaviors, e.g. bird singing, nest building, courting. In analogy with this phenomenon, Penfield and Roberts (1959), and later Lenneberg (1967), proposed the concept of a critical period for language acquisition. Initially proposed for the acquisition of a first language, this critical period (roughly between the age of two and puberty) has also been studied in the context of L2

*Corresponding author. Tel.: +1 979 845 2942; fax: +1 979 847 8578.

acquisition (Major, 2001). Among the many aspects of proficiency in a second language (e.g. lexical, syntactic, semantic, phonological), native-like pronunciation is the most severely affected by a critical period because of the neuromusculatory basis of speech production (Scovel, 1988). Thus, according to this theory, foreign-accented production is unavoidable if a second language is learned beyond the critical period years.

To address this gloomy outlook on L2 pronunciation, many authors argue that what really matters is that the speech be intelligible, rather than accent-free (Neri et al., 2002; Pennington, 1999). However, although foreign accentedness does not necessarily affect a person's ability to be understood (Munro and Derwing, 1995), foreign-accented speakers can be subjected to discriminatory attitudes and negative stereotypes (Anisfeld et al., 1962; Arthur et al., 1974; Lippi-Green, 1997; Ryan and Carranza, 1975; Schairer, 1992). Thus, by achieving near-native pronunciation, L2 learners stand more to gain than just better intelligibility. A second, and more direct, counter-argument to the critical period hypothesis has been provided by a number of studies showing that native-like pronunciation can be achieved by adults learning the second language well beyond puberty (Bongaerts, 1999). Nonetheless, the proportion of such native-like L2 speakers is believed to be small: between 0.1% and 3% (Markham, 1997). Given the small probability of attaining native-like pronunciation, it is unrealistic to believe that an L2 learner should hold this as their ultimate goal. However, most L2 learners can make significant strides towards reducing their accent and possibly achieving near-native performance. According to Bongaerts (1999), several factors contribute to the success of such L2 speakers: (1) a high motivation to achieve accent-free pronunciation, (2) unlimited access to L2 speech, and (3) intensive training in L2 perception and L2 production. These characteristics suggest that computer assisted pronunciation training (CAPT) is an ideal medium for the attainment of near-native pronunciation.

Although not as effective as human instruction, CAPT offers several features that make it advantageous in classroom settings (Neri et al., 2002; Pennington, 1999). Most notably, CAPT allows users to follow personalized lessons, at their own pace, and practice as often as they like. One study (Murray, 1999) showed that users are more comfortable practicing pronunciation in a private setting, where they can avoid anxiety and embarrassment. Users are also more likely to practice when and where it is convenient. The most praised systems are those that incorporate Automatic Speech Recognition (ASR) (e.g. FLUENCY (Eskenazi and Hansma, 1998), ISLE (Menzel et al., 2000), and Talk to Me (Auralog, 2002)) because they are able to provide users with objective and consistent (though not always correct) feedback. CAPT systems also keep several speakers in their databases, which helps listeners improve their listening comprehension (McAllister, 1998).

Despite the potential advantages of CAPT, the technology remains controversial for reasons beyond the debate surrounding the critical period hypothesis (Pennington, 1999). As noted by Neri et al. (2002), part of the problem is that many commercial products have chosen technological novelty over pedagogical value. For instance, a product may provide a display of the learner's utterance (e.g. a speech waveform or a spectrogram) against that from a native speaker. These visualizations are not only difficult to interpret for non-specialists but are also misleading: two utterances can have different acoustic representations despite having been pronounced correctly.[1] A second criticism stems from the limitations of ASR technology when used for detecting pronunciation errors and evaluating pronunciation quality (Neri et al., 2003); CAPT is a challenging domain for ASR because of the inherent variability of foreign-accented speech. ASR errors not only frustrate and mislead the learner

[1]On the other hand, it has been shown that displaying pitch contours improves intonation training, and that audio-visual feedback also improves prosody and segmental accuracy (Hincks, 2003).

but also, and more importantly, undermine their trust in the CAPT tool (Levy, 1997; Wachowicz and Scott, 1999).

Feedback is a critical component in pronunciation training; unfortunately, research data on the effectiveness of various feedback strategies is scarce (Neri et al., 2002). Hansen (2006) prescribes four critical criteria for feedback in CAPT; feedback should be easy to understand (comprehensive), feedback should determine if the correct phoneme was used (qualitative) and if the phoneme was of the correct length (quantitative), and feedback should suggest actions for improvement (corrective). CAPT systems employing ASR technology usually satisfy the first three requirements, but often have difficulty providing meaningful corrective suggestions. One of the more ambitious CAPT systems to date (ISLE; (Menzel et al., 2000)) satisfied all four of the criteria in (Hansen, 2006), but poor ASR accuracy ultimately limited its adoption. ASR errors can be so disruptive to the learner that Wachowicz and Scott (1999) have suggested that CAPT systems should rely on implicit rather than explicit feedback. As an example, recasts—a rephrasing of the incorrectly pronounced utterance-have been shown to be superior to explicit correction of phonological errors (Lyster, 2001).

2. Motivation for our work

During the last two decades, a handful of studies have suggested that it would be beneficial for L2 students to be able to listen to their own voices producing native-accented utterances (Jilka and Möhler, 1998; Sundström, 1998; Tang et al., 2001; Watson and Kewley-Port, 1989). The rationale is that, by stripping away information that is only related to the teacher's voice quality, accent conversion makes it easier for students to perceive differences between their accented utterances and their ideal accent-free counterparts. In addition, it can be argued that accent-corrected utterances provide a form of feedback that is implicit (Wachowicz and Scott, 1999), corrective, and encouraging. A series of previous studies support this view.

Nagano and Ozawa (1990) evaluated a prosodic-conversion method for the purpose of teaching English pronunciation to Japanese learners. One group of students was trained to mimic utterances from a reference English speaker, whereas a second group was trained to mimic utterances of their own voices, previously modified to match the prosody of the reference English speaker. Pre- and post-training utterances from both groups of students were evaluated by native English listeners. Post-training utterances from the second group of students were rated as more native-like than those from the first group. More recently, Bissiri et al. (2006) investigated the use of prosodic modification to teach German prosody to Italian speakers. Their results were consistent with those of Nagano and Ozawa (1990), and indicate that the learner's own voice (with corrected prosody) is a more effective form of feedback than prerecorded utterances from a German native speaker. Peabody and Sene. (2006) proposed a similar strategy to teach pronunciation of a tonal language (Mandarin), a problem that is very challenging for students whose native language is non-tonal (English). For this purpose, the authors used three different datasets of Mandarin utterances: a corpus produced by native speakers, and two corpora produced by L2 speakers. Using a phase vocoder, the authors transformed the pitch contour of L2 utterances to match the tonal shapes of native utterances. The transformed L2 utterances were twice as likely to be classified correctly by a pattern classifier. This result is not surprising since the classifier was trained with tones, but it does highlight the importance of prosody and its ability to indicate accent. Anecdotal support for the use of accent conversion is also provided by studies of categorical speech perception and production. In particular, Repp and Williams (1987) compared the accuracy of speakers imitating isolated vowels in two continua: [u]-[i] and [i]-[æ]. Their results indicate that speakers were more accurate when imitating their own

(earlier) productions of those vowels than when imitating vowels produced by a speech synthesizer.

More recently, a few CAPT tools have begun to incorporate prosodic-conversion capabilities. These tools allow L2 learners to re-synthesize their own utterances with a native prosody, either through a manual editing procedure (Martin, 2004) or with automated algorithms (GenevaLogic, 2007). Proper intonation and stress are critical because they provide a temporal structure that helps the listener parse the continuous speech waveform (Celce-Murcia et al., 1996). Thus, a number of authors have suggested that prosody should be emphasized early on in teaching a second language (Chun, 1998; Eskenazi, 1999). However, speech intelligibility can also degrade as a result of segmental/ spectral errors (Rogers and Dalby, 1996), which indicates that both segmental and supra-segmental features should be considered in pronunciation training (Derwing et al., 1998b). This suggests that full accent conversion (i.e. prosodic and segmental) would be beneficial in teaching pronunciation of a foreign language.

Probst et al. (2002) investigated the relationship between the student/teacher voice similarity and pronunciation improvement. Results from this study showed that learners who imitated 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 full accent conversion would provide learners with the optimal "golden speaker": their native-accented selves. As a step towards this goal, this manuscript describes a speech processing method that can be used to transform foreign-accented utterances into their native counterparts, and provides a thorough validation of the method through a series of perceptual tests. In addition, we discuss implementation issues and propose a pedagogical strategy that would integrate accent conversion into computer assisted pronunciation training as a form of behavioral shaping (Kewley-Port and Watson, 1994; Watson and Kewley-Port, 1989).

3. Foreign accent conversion

What constitutes a foreign accent? A foreign accent can be defined as deviations from the expected acoustic (e.g. formants) and prosodic (e.g. intonation, duration, and rate) norms of a language. According to the modulation theory of speech (Traunmüller, 1994), a speaker's utterance results from the modulation of a voice quality carrier with linguistic gestures. In this context, Traunmüller identifies the carrier as the organic aspects of a voice that "*reflect the morphological between-speaker variations in the dimensions of speech*," such as those that are determined by physical factors (e.g. larynx size and vocal tract length). Thus, in analogy with the source/filter theory of speech production (Fant, 1960), which decomposes a speech signal into excitation and vocal tract resonances, modulation theory suggests that one could deconvolve an utterance into its voice quality carrier and its linguistic gestures. According to this view, then, a foreign accent may be removed from an utterance by extracting its voice quality carrier and convolving it with the linguistic gestures of a native-accented counterpart.

In contrast with voice conversion, which seeks to transform utterances from a speaker so they sound as if another speaker had produced them (Abe et al., 1988; Arslan and Talkin, 1997; Childers et al., 1989; Kain and Macon, 1998; Sundermann et al., 2003; Turk and Arslan, 2006), 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. Accent conversion is a relatively new concept; as a result, only a handful of studies have been published on the subject. Yan et al. (2004) proposed an accent-synthesis method based on formant warping. First, the authors developed a formant tracker based on hidden Markov

models and linear predictive coding, 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, which indicates that segmental accent conversion is feasible. More recently, Huckvale and Yanagisawa (2007) used an English text-to-speech (TTS) system to simulate English-accented Japanese utterances; foreign accentedness 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 the study of Yan et al. (2004) in two respects. First, our accent conversion method uses a spectral envelope vocoder, which makes it more suitable than formant tracking for unvoiced segments. Second, we evaluate not only the accentedness of the re-synthesized speech but also the perceived identity of the resulting speaker. 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 (2007), our study is performed on natural speech, and focuses on accentedness and identity rather than on intelligibility; as noted by Munro and Derwing (1995), a strong foreign accent does not necessarily limit the intelligibility of the speaker.

The remaining sections of this article are organized as follows. Section 4 provides an overview of the speech modification framework (FD-PSOLA) adopted for this work, and describes our method of accent conversion. Section 5 describes the perceptual protocol we have employed to evaluate our method along three dimensions: foreign accentedness, speaker identity, and signal quality; results from these perceptual experiments are analyzed in Section 6. The article concludes with a discussion of our findings and the implications of accent conversion in CAPT.

4. Accent conversion with FD-PSOLA

Our accent conversion transformation is based on the general framework of Pitch-Synchronous Overlap and Add (PSOLA) (Moulines and Charpentier, 1990). Several versions of PSOLA have been proposed in the literature, including Fourier-domain FD-PSOLA, linear-prediction LP-PSOLA, and time-domain TD-PSOLA (Moulines and Charpentier, 1990; Moulines and Laroche, 1995). 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. During the analysis stage, the speech signal is decomposed into a series of pitch-synchronous short-time analysis windows; our implementation uses a pitch-marking algorithm (Kounoudes et al., 2002) to estimate instants of glottal closure. Each analysis window is framed with a Hanning window, and transformed into the frequency domain.[2] As a result, all pitch-synchronous

[2]Our implementation follows the recommended window length of four times the local pitch period for voiced segments or a constant 10 ms for unvoiced segments (Moulines and Charpentier, 1990).

short-time spectra are represented with the same length (e.g. 2048 frequencies in our implementation).

In the modification 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 native 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 last step 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) (Paul, 1981). We also use a spectral folding technique (Makhoul and Berouti, 1979) to regenerate high frequency components that are lost when performing spectral compression (Fig. 1b). Finally, we multiply the flattened spectrum by the SEEVOC spectral envelope estimate, thus restoring its original resonances (Fig. 1c).

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:

$$\widehat{X}(n)=\frac{\sum_{m=-\infty}^{\infty} w(m-n)F_m^{-1}(n)}{\sum_{m=-\infty}^{\infty} w^2(m-n)} \quad (1)$$

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) (Griffin and Lim, 1984).

4.1. Accent conversion

For convenience, we will call the second-language (foreign) speaker of American English the *learner*, and the native speaker of American English the *teacher*. We also assume that parallel English utterances are available from both speakers.

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

4.1.1. Prosodic conversion—To perform *time-scale* conversion, we assume that the speech has been phonetically segmented by hand or with a forced-alignment tool (Sphinx, 2001; Young, 1993). From these phonetic segments, the ratio of teacher-to-learner durations is used to specify a time-scale modification factor α for the learner on a phoneme-by-phoneme basis; as prescribed by Moulines and Laroche (1995), we limit time-scale factors to the range of $\alpha = [0.25, 4]$.

Our *pitch-scale* modification combines the pitch dynamics of the teacher with the pitch baseline of the learner. This is achieved by replacing the pitch contour of the learner utterance with a transformed (i.e. shifted and scaled) version of the pitch contour of the teacher utterance. For this purpose, we first estimate average pitch values for the learner ($\overline{f_0^L}$) and teacher ($\overline{f_0^T}$) from a corpus of utterances. Next, we define a piecewise-linear time-warping, $\Psi_{LT}(f(t))$, to align learner and teacher utterances at phoneme boundaries. Finally,

given pitch contours $f_0^L(t)$ and $f_0^T(t)$ for the specific learner and teacher utterances to be converted, we define a pitch-scale factor β as

$$\beta(t)=\frac{\psi_{LT}(f_0^T(t))+\overline{f_0^L}-\overline{f_0^T}}{f_0^L(t)} \quad (2)$$

where we also limit pitch-scale factors to the range of $\beta = [0.5, 2]$. This process allows us to preserve speaker identity by maintaining a reasonable pitch baseline and range (Compton, 1963; Sambur, 1975), while acquiring the pitch dynamics of the teacher, which provides important cues to native accentedness (Arslan and Hansen, 1997; Munro, 1995; Vieru-Dimulescu and Mareüil, 2005). Once the time-scale and pitch-scale modification parameters (α,β) are calculated, standard FD-PSOLA is used to perform the prosodic conversion.

4.1.2. Segmental conversion—Our segmental accent conversion stage assumes that the glottal excitation signal is largely responsible for voice quality, whereas the filter contributes to most of the linguistic information. Thus, our strategy consists of combining the teacher's spectral envelope (filter) with the learner's glottal excitation. FD-PSOLA allows us to perform this step in a straightforward fashion: in the final step illustrated in Fig. 1c, we multiply the learner's flat spectra by the teacher's envelope rather than by the learner's envelope. In order to reduce speaker-dependent information in the teacher'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 (see Fig. 2) (Sundermann et al., 2003). These formant locations are estimated with Praat (Boersma and Weenink, 2007) over the entire corpus. The result is a signal that consists of the learner's excitation, and the teacher's spectral envelope normalized to the learner's vocal tract length.

5. Perceptual experiments

The proposed accent conversion method was evaluated through a series of perceptual experiments. We were interested in determining (1) the degree of reduction in foreign accent that could be achieved with the model, and (2) the extent to which the transformation preserved the identity of the original 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 five stimulus conditions (see Table 1). Sample video files for the five conditions are available as Supplemental material (1–5.mpg and rev1–5.mpg), and spectrograms of the three primary conditions (1, 4 and 5) are shown in Fig. 3.

One hundred and ninety one participants were recruited from the undergraduate pool maintained by the Department of Psychology at Texas A&M University. All participants were native speakers of American English and had no hearing or language impairments. Two speakers were selected from the CMU_ARCTIC database (Kominek and Black, 2003): *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 learner, and utterances from *rms_usmale2* were treated as the native-accented teacher.[3] The

[3]In voice conversion, the choice of speakers is known to have a significant impact on the quality of the output (Turk and Arslan, 2005); we suspect that accent conversion is no different, but have not yet determined those factors that predict the success of the transformation. This is one of our immediate priorities as it must be investigated before a robust, learner-independent training tool is created.

same twenty sentences were chosen for each of the five conditions, or 100 unique utterances. Audio stimuli were presented via headphones.

5.1. 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 (1994), 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) (Pelham and Blanton, 2007). Each participant rated all 100 utterances.

5.2. Acoustic quality experiment

Forty-three students participated in a 25-minute Mean Opinion Score test to obtain a numerical indication of the perceived quality (e.g. lack of distortions) of the recorded/ synthesized utterances. Following Kain and Macon (Kain and Macon, 1998), participants heard an utterance and were asked to indicate the acoustic quality of the stimulus on a standard MOS scale from 1 (bad) to 5 (excellent), where "excellent" was defined as a sound that had no distortions. Before the test began, students listened to examples of sounds with various accepted MOS values. This task included feedback, which allowed students to calibrate themselves to the reference scores. Each participant rated all 100 utterances.

5.3. Identity experiment

Forty-three students participated in a 25-minute speaker identification test. Following Kreiman and Papcun (1991), participants heard two linguistically different[4] utterances presented consecutively, and were instructed to "focus on those aspects of the voice that determine identity." Participants were asked to determine if the two sentences were produced by the same speaker or by two different speakers, and 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 score from 0 to 14 (Table 2). Each participant listened to 60 pairs of utterances.[5]

5.4. Identity experiment with reversed speech

Sixty-six students participated in a 25-minute speaker identification test similar to that in Section 5.3, except that utterances were played backwards. Although the instructions in the previous experiment clearly stated that participants should focus on the identity-related aspects of the speaker voices, we suspected that it would be difficult for participants to ignore linguistic cues in those utterances. Fortunately, reversed speech removes most of the linguistic cues (e.g. language, vocabulary, and accent) that may be used to identify a speaker, while retaining the pitch, pitch range, speaking rate, and vocal quality of the speaker, which can be used to identify familiar and unfamiliar voices (Sheffert et al., 2002; van Lancker et al., 1985). Thus, this reversed-speech identification test allowed us to determine the extent to which the accentedness of our speakers (or possibly distortions resulting from the re-synthesis) had been used as a cue in the previous identification study.

[4]The 20 distinct sentences were divided into two sets (1–10 and 11–20) to ensure that pairs were linguistically unique. Presentation was counterbalanced across sets (i.e. a sentence from the first set was not always played first).

[5]All possible pairings can be expressed as a 5 × 5 matrix. To ensure that all pairs were sampled with the same frequency, diagonal elements in this matrix (i.e. same–same pairings) were sampled twice as often as off-diagonal elements, thus leading to 60 pairs (=(25 + 5) × 2 repetitions).

6. Results

Results from the foreign accentedness experiment are summarized in Fig. 4a. Original recordings from the foreign speaker received the highest average accent rating (4.85), while native speaker recordings had the lowest average rating (0.15). The prosodic transformation decreased the perceived accent slightly (4.83), but this change was not statistically significant; $t(38) = 0.38$, *n.s*. On the other hand, the segmental transform lowered the rating to 1.97; $t(38) = 24.14$, $p \ll 0.01$. When used in concert, both transformations yield an average score of 1.79; $t(38) = 24.06$, $p \ll 0.01$. Both of these reductions were statistically significant.

Results from the acoustic quality experiment are summarized in Fig. 4b. Original recordings from the native speaker received the highest average quality rating (4.84), while the unmodified foreign speaker averaged a lower rating (4.0); this difference was statistically significant; $t(42) = 6.68$, $p \ll 0.01$. This lower rating may have been caused by differences in the recording conditions for both speakers, but it is also possible that subjects penalized the "quality" of non-native speech because it was less intelligible. All transformations lowered the quality ratings: starting from the original baseline (4.0), the prosodic and segmental transformations reduced quality ratings to 2.96 and 2.67, respectively; these differences were statistically significant ($t(42) = 12.49$, $p \ll 0.01$ and $t(42) = 9.19$, $p \ll 0.01$), with respect to the rating of foreign-accented utterances.

The identity experiments yield a collection of perceptual distances between pairs of utterances (0/14: the participant was extremely confident that the speakers were the same/different). Because only the relative distance between stimuli is available, we resort to multi-dimensional scaling (MDS) to find a low-dimensional visualization (e.g. 2D) of the data that preserves those pair-wise distances; see (Matsumoto et al., 1973) for a classical use of MDS in speech perception. Namely, we use ISOMAP (Tenenbaum et al., 2000), an MDS technique that attempts to preserve the geodesic distance[6] between stimuli. For clarity, technical details of ISOMAP are included in Appendix A.

ISOMAP visualizations of the identity tests are shown in Fig. 5. In the case of forward speech, samples from conditions 1 and 2 map closely together in the manifold. Thus, this result indicates that the prosodic transformation had only a small effect on the perceived identity of the speakers. On the other hand, samples of learner utterances (condition 1) and their segmental transformations (conditions 3 and 4) are clearly separated in the ISOMAP manifold. This result indicates that participants were able to distinguish between learner and segmentally-transformed utterances, which suggests that they perceived the latter as a "third" speaker; note that this type of inference is not possible with the ABX tests commonly used in voice conversion. However, all samples containing the learner's glottal excitation (conditions 1–4) appear to map on a linear subspace that is separate from the teacher utterances (condition 5), which indicates that the former are perceived as being closer to each other than to the teacher. In fact, by calculating the average Euclidean distance across conditions, we find that this "third" speaker (conditions 3–4) is perceived to be three times closer to the learner (condition 1) than to the teacher (condition 5).

Results from the reversed-speech experiment, shown in Fig. 5b, indicate that participants were unable to differentiate between conditions 1–2 and conditions 3–4. These results support our hypothesis, namely, that participants in the forward speech experiment had

[6]ISOMAP assumes that samples exist on an intrinsically low-dimensional surface – a manifold. The geodesic distance is defined as the Euclidean distance between samples measured *over* this manifold. In ISOMAP, the geodesic distance is estimated as the shortest path in a graph where nodes represent samples and edges indicate neighboring samples.

identified conditions 3–4 as a "third" speaker because of the association between accentedness and speaker identity. Since most linguistic cues (including accent) are not accessible with reversed speech, participants perceive conditions 1–4 as utterances from the same speaker.

Interestingly, the ISOMAP embedding in both cases (though more clearly with forward speech) can be interpreted in terms of the source-filter theory. As shown in Fig. 5, the first dimension separates samples in condition 5, which uses the teacher's glottal excitation, from samples in the remaining conditions, which use the learner's glottal excitation. In contrast, the second dimension separates samples in conditions 1–2, which employ the learner's filter, from samples in conditions 3–5, which employ the teacher's filter.

7. Discussion

The perceptual results presented in the previous section indicate that our accent conversion approach can reduce the perceived foreign accentedness of an utterance (by about two-thirds) while preserving information that is unique to the voice quality of the speaker. Thus, these results support our choice of a spectral envelope vocoder to decompose utterances into their voice quality and linguistic components. Although foreign-accented utterances (condition 1) are already perceived as being of lower quality, the technique itself introduces perceivable distortions, as indicated by the lower quality ratings for conditions 2–4. This result could be attributed to several factors, including segmentation/alignment errors, voicing differences between speakers, and phase distortions that result from combining glottal excitation with spectral envelope from different speakers. Our results of accentedness seem to underplay the importance of prosody when compared with other studies (Jilka and Möhler, 1998; Nagano and Ozawa, 1990). This could be a consequence of the elicitation procedure used in the ARCTIC database (Kominek and Black, 2003), since read speech is more prosodically flat than spontaneous or conversational speech (Kenny et al., 1998).

Identity tests with forward speech indicate that the segmental transformations (with or without prosodic transformation) are perceived as a third speaker. This third speaker disappears, however, when participants are asked to discriminate reversed speech.[7] One could argue that the emergence of a third speaker on forward speech is merely the result of distortions introduced by the segmental transformation; these distortions are imperceptible when utterances are played backward, which may explain why the third speaker "disappears" with reversed speech. In other words, accentedness and acoustic quality would be confounded in our experiments. This view, however, is inconsistent with the acoustic quality ratings obtained in the second experiment. As shown in Fig. 4b, quality ratings for condition 2 are similar to those of conditions 3–4, rather than to those of condition 1; if participants had used acoustic quality as a cue in the identification study, condition 2 would have been perceived also as belonging to the third speaker. Thus, our identification experiments with forward and reverse speech indicate that participants used not only organic cues (voice quality) but also linguistic cues (accentedness) to discriminate speakers. This suggests that something is inevitably lost in the identity of a speaker when accent conversion is performed. After all, would foreign-born public figures (e.g. Arnold Schwarzenegger, Javier Bardem) be recognized as themselves without their distinct accents?

[7]One could argue that the results in Fig. 5b contain three clusters, but that they are more spread due to the increased difficulty of the task. The fact remains that, even with reverse speech, subjects are able to discriminate teacher utterances from other utterances, whereas they can hardly discriminate learner utterances from their accent converted versions.

7.1. Relevance to pronunciation training

As discussed in Section 2, several studies have suggested the use of speech modification as a training tool for second-language pronunciation, and have shown promising results (Bissiri et al., 2006; Nagano and Ozawa, 1990; Peabody and Seneff, 2006). In addition, new CAPT tools have also begun to incorporate speech modification capabilities (GenevaLogic, 2007; Martin, 2004). This previous work has focused on time-scale and pitch-scale modifications, arguably because of the impact that prosody has on foreign accent and intelligibility. However, segmental pronunciation errors are also detrimental to intelligibility (Rogers and Dalby, 1996), and both aspects of pronunciation should be considered during training (Derwing et al., 1998a).

Our work has focused on developing a method for full (i.e. prosodic and segmental) accent conversion, and characterizing the model on three perceptual criteria: foreign accentedness, speaker identity, and acoustic quality. While our perceptual results are encouraging, the proposed accent conversion model has yet to be validated for the purposes of pronunciation training. To this end, our immediate issues deal with the implementation of the accent conversion method as a CAPT tool:

- Establishing a calibration procedure to each particular user; this will require extracting speaker-dependent variables such as F0 and average formant values from an initial collection of utterances from the learner.
- Improving the run-time of the model; at present, the accent conversion method alone (i.e. without forced alignment) requires 12.5 s of processing per second of utterance. These estimates are based on our current MATLAB® implementation running on a 2 GHz desktop; we expect that porting to a compiled language with optimizations will improve run time by a factor of 20–100.
- Integrating the accent conversion model with forced-alignment of utterances in real time. We have developed a working prototype of the forced-alignment stage (Young, 1993); our initial results indicate that the current level of performance of forced-alignment tools is sufficient for the purposes of accent conversion, but the effect of segmentation errors on pronunciation training will need to be investigated.
- Developing quantitative metrics of user pronunciation performance; these may be based on time-scale and pitch-scale modification factors (α,β), as well as on dissimilarity between spectral envelopes (following vocal tract length normalization).

In addition to these technical challenges, special attention will have to be paid to feedback and pedagogical issues. Fortunately, earlier work by Kewley-Port and Watson (Kewley-Port and Watson, 1994; Watson and Kewley-Port, 1989) provides a framework for integrating accent conversion in CAPT. Namely, the authors proposed a three-dimensional taxonomy of CAPT systems according to their feedback strategy. The first two dimensions characterize feedback in terms of the level of detail (e.g. a spectrogram of the learner's utterance versus a quality rating) and type of physical media (e.g. acoustic vs. visual). The third dimension is quite relevant to our work, because it characterizes feedback according to the standard against which the system evaluates productions of the learner. Two types of references are considered in the taxonomy: a normative standard (e.g. the teacher's speech) and actual samples of the learner's speech.[8] The authors argue that using the student's own voice as a standard can be considered as an attempt to incorporate "behavioral shaping" procedures

[8]The authors also advance that "*another approach to generate client-specific standards would be to generate a speech model or template that could be based on the resonant characteristics of the client's own vocal tract*," a view that is supportive of the work presented in this manuscript.

into CAPT. In behavioral shaping, the teacher asks the students to compare their utterances against their previous efforts rather than against a separate standard. This is accomplished by keeping track of the student's "best" utterances, and using them as a reference. Kewley-Port and Watson argue that using a normative reference can be detrimental in the early stages of 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 learner's performance and guide him towards the ultimate goal. Accent conversion provides a mechanism for implementing such behavioral shaping procedures. Namely, by convolving the learner's glottal excitation with a "morph" between the learner's and teacher's spectral envelope (as opposed to using the teacher's envelope), accent conversion provides a continuum of transformations. During the early stages of learning, the system would provide the learner with transformed utterances that have less ambitious prosodic and segmental goals. The rationale is that these intermediate teachers would provide more realistic (though still challenging) goals for the user for imitate. As each of these intermediate teachers was met, the transformation would be updated using the latest, best pronunciation of the user. Continuing in this iterative manner, the training processes can be seen as a trajectory in a two-dimensional imitation space composed of the increasingly better productions of the learner (behavioral shaping) and a continuum of accent transformations (morphing). This process is illustrated in Fig. 6.

We believe that accent conversion could play a significant role in the next generation of computer assisted pronunciation training tools. Our method is based on the assumption that accent is contained in the prosody and formant structure of an utterance, whereas speaker identity is captured by vocal tract length and glottal shape characteristics. Our method employs FD-PSOLA to adapt the speaking rate and pitch of the learner towards those of the teacher, and a segmental transformation to replace the spectral envelope of the learner with that of the normalized teacher. These techniques achieved a significant reduction in foreign accent while preserving the voice quality of the speaker. Our results also reveal a strong connection between accent and identity, which suggest a tradeoff between reducing accentedness and preserving speaker identity. Our perceptual results, coupled with previous research showing the benefit of prosodic manipulation in pronunciation training, suggests that full accent conversion (segmental and prosodic) can be a successful form of implicit feedback in computer assisted pronunciation training.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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." We would also like to acknowledge an anonymous reviewer for providing an alternative explanation to the results in Fig. 5b.

References

Abe, M.; Nakamura, S.; Shikano, K.; Kuwabara, H. Voice conversion through vector quantization. Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing; New York, NY. 1988. p. 655-658.

Anisfeld M, Bogo N, Lambert WE. Evaluational reactions to accented English speech. J Abnorm Social Psychol 1962;65:223–231.

Arslan LM, Hansen JHL. A study of temporal features and frequency characteristics in American English foreign accent. J Acoust Soc Amer 1997;102 (1):28–40.

Arslan, LM.; Talkin, D. Voice Conversion By Codebook Mapping Of Line Spectral Frequencies And Excitation Spectrum. Eurospeech' 97; Rhodes, Greece. 1997. p. 1347-1350.

Arthur B, Farrar D, Bradford G. Evaluation reactions of college students to dialect differences in the English of Mexican–Americans. Language Speech 1974;17 (3):255–270.

Auralog. Talk to Me. 2002. <http://www.auralog.com/en/Individuals_talktome.htm>

Bissiri, MP.; Pfitzinger, HR.; Tillmann, HG. Lexical Stress Training of German Compounds for Italian Speakers by means of Resynthesis and Emphasis. Proceedings of the 11th Australian International Conference on Speech Science & Technology; New Zealand: University of Auckland; 2006. p. 24-29.

Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer (Version 4.5.15). Universiteit van; 2007.

Bongaerts, T. Ultimate attainment in L2 pronunciation: the case of very advanced late L2 learners. In: Birdsong, D., editor. Second Language Acquisition and the Critical Period Hypothesis. Lawrence Erlbaum; Mahway, NJ: 1999. p. 133-159.

Celce-Murcia, M.; Brinton, D.; Goodwin, JM. Teaching pronunciation: a reference for teachers of English to speakers of other languages. Vol. xii. Cambridge University Press; Cambridge; New York: 1996. p. 435

Childers DG, Wu K, Hicks DM, Yegnanarayana B. Voice conversion. Speech Commun 1989;8 (2): 147–158.

Chun D. Signal analysis software for teaching discourse intonation. Language Learn Technol 1998;2 (1):61–77.

Compton AJ. Effects of filtering and vocal duration upon identification of speakers, aurally. J Acoust Soc Amer 1963;35 (11):1748–1755.

Derwing T, Munro M, Wiebe G. Evidence in favor of a broad framework for pronunciation instruction. Language Learn 1998a;48 (3):393–410.

Derwing T, Munro M, Wiebe G. Evidence in favour of a broad framework for pronunciation instruction. Language Learn 1998b;48:393–410.

Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959;1 (1):269–271.

Eskenazi M. Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype. Language Learn Technol 1999;2 (2):62–76.

Eskenazi, M.; Hansma, S. The fluency pronunciation trainer. Proc. STiLL Workshop on Speech Technology in Language Learning; 1998.

Fant, G. Mouton s' Gravenhage. 1960. Acoustic Theory of Speech Production.

GenevaLogic. SpeedLingua. 2007. <http://www.speedlingua.com>

Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Trans Acoust Speech Signal Process 1984;32 (2):236–243.

Hansen, TK. Computer assisted pronunciation training: the four 'K's of feedback. 4th Internat. Conf. on Multimedia and Information and Communication Technologies in Education; Seville, Spain. 2006. p. 342-346.

Hincks R. Speech technologies for pronunciation feedback and evaluation. ReCALL 2003;15 (1):3–20.

Huckvale, M.; Yanagisawa, K. Spoken language conversion with accent morphing. Proc. ISCA Speech Synthesis Workshop; Bonn, Germany. 2007. p. 64-70.

Jilka, M.; Möhler, G. Intonational foreign accent: speech technology and foreign language teaching. Proc. ESCA Workshop on Speech Technology in Language Learning; 1998. p. 115-118.

Kain, A.; Macon, MW. Spectral voice conversion for text-to-speech synthesis. Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing; 1998. p. 285-288.

Kenny, OP.; Nelson, DJ.; Bodenschatz, JS.; McMonagle, HA. Separation of non-spontaneous and spontaneous speech. Proc. 1998 IEEE Internat. Conf. on Acoustics, Speech and Signal Processing; 1998. p. 573-576.

Kewley-Port, D.; Watson, CS. Computer assisted speech training: practical considerations. In: Syrdal, RBA.; Greenspan, S., editors. Applied Speech Technology. CRC Press; Boca Raton, FL: 1994. p. 565-582.

Kominek, J.; Black, A. CMU ARCTIC databases for speech synthesis. Carnegie Mellon University Language Technologies Institute; 2003.

Kounoudes, A.; Naylor, PA.; Brookes, M. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing; Orlando, FL, USA. 2002. p. I-349-I-352.

Kreiman J, Papcun G. Comparing discrimination and recognition of unfamiliar voices. Speech Commun 1991;10 (3):265–275.

Lenneberg, EH. Biological Foundations of Language. Vol. xvi. Wiley; New York: 1967. p. 489

Levy, M. Computer-assisted Language Learning: Context and Conceptualization. Vol. xv. Clarendon Press; Oxford University Press; Oxford, New York: 1997. p. 298

Lippi-Green, R. English With an Accent: Language, Ideology, and Discrimination in the United States. Routledge; 1997.

Lyster R. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. Language Learn 2001;51 (s1):265–301.

Major, RC. Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology. Vol. ix. L. Erlbaum; Mahwah, NJ: 2001. p. 209

Makhoul, J.; Berouti, M. In: Berouti, M., editor. High-frequency regeneration in speech coding systems; Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing; 1979. p. 428-431.

Markham, D. Travaux de l'Institut de linguistique de Lund. Lund University Press; Lund: 1997. Phonetic Imitation, Accent, and the Learner; p. 269

Martin, P. WinPitch LTL II, a Multimodal Pronunciation Software. ISCA; 2004.

Matsumoto H, Hiki S, Sone T, Nimura T. Multidimensional representation of personal quality of vowels and its acoustical correlates. IEEE Trans Audio Electroacoust 1973;21 (5):428–436.

McAllister, R. Second language perception and the concept of foreign accent. Proc. ESCA Workshop on Speech Technology in Language Learning (STiLL 98); Stockholm, Sweden. 1998. p. 155-158.

Menzel, W.; Herron, D.; Bonaventura, P.; Morton, R. Automatic detection and correction of non-native English pronunciations. InSTILL; Dundee, Scotland: 2000. p. 49-56.

Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun 1990;9 (5–6):453–467.

Moulines E, Laroche J. Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Commun 1995;16 (2):175–205.

Munro M. Non-segmental factors in foreign accent: ratings of filtered speech. Studies Second Language Acquisit 1995;17:17–34.

Munro M, Derwing T. Evaluations of foreign accent in extemporaneous and read material. Language Testing 1994;11:253–266.

Munro M, Derwing T. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learn Technol 1995;45 (1):73–97.

Murray GL. Autonomy and language learning in a simulated environment. System 1999;27 (3):295–308.

Nagano, K.; Ozawa, K. English speech training using voice conversion. 1st Internat. Conf. on Spoken Language Processing (ICSLP 90); Kobe, Japan. 1990. p. 1169-1172.

Neri A, Cucchiarini C, Strik H, Boves L. The pedagogy–technology interface in computer assisted pronunciation training. Comput Assisted Language Learn 2002;15 (5):441–467.

Neri, A.; Cucchiarini, C.; Strik, H. Automatic speech recognition for second language learning: how and why it actually works. Proc. 15th Internat. Congress of Phonetic Sciences; Barcelona, Spain. 2003. p. 1157-1160.

Paul D. The spectral envelope estimation vocoder. IEEE Trans Acoust Speech Signal Process 1981;29 (4):786–794.

Peabody M, Sene S. Towards automatic tone correction in nonnative mandarin. Chinese Spoken Language Process 2006:602–613.

Pelham, B.; Blanton, H. Conducting Research in Psychology, Measuring the Weight of Smoke. Thomson Higher Education; Belmont, CA: 2007.

Penfield, W.; Roberts, L. Speech and brain-mechanisms. Princeton University Press; Princeton, NJ: 1959. p. 286

Pennington MC. Computer-aided pronunciation pedagogy: promise, limitations, directions. Comput Assisted Language Learn 1999;12:427–440.

Probst K, Ke Y, Eskenazi M. Enhancing foreign language tutors – in search of the golden speaker. Speech Commun 2002;37 (3–4):161–173.

Repp BH, Williams DR. Categorical tendencies in imitating self-produced isolated vowels. Speech Commun 1987;6 (1):1–14.

Rogers CL, Dalby JM. Prediction of foreign-accented speech intelligibility from segmental contrast measures. J Acoust Soc Amer 1996;100 (4):2725–2726.

Ryan EB, Carranza MA. Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English. J Personality Social Psychol 1975;31 (5):855–863.

Sambur M. Selection of acoustic features for speaker identification. IEEE Trans Acoust Speech Signal Process 1975;23 (2):176–182.

Schairer KE. Native speaker reaction to non-native speech. Mod Language J 1992;76 (3):309–319.

Scovel, T. Issues in Second Language Research. Vol. ix. Newbury House; Cambridge (England), New York: 1988. A Time to Speak: A Psycholinguistic Inquiry into the Critical Period for Human Speech; p. 206

Sheffert SM, Pisoni DB, Fellowes JM, Remez RE. Learning to recognize talkers from natural, sinewave, and reversed speech samples. J Exp Psychol Hum Percept Perform 2002;28 (6):1447–1469. [PubMed: 12542137]

Sphinx. SphinxTrain: Building Acoustic Models for CMU Sphinx. Carnegie Mellon University; 2001.

Sundermann, D.; Ney, H.; Hoge, H. VTLN-based cross-language voice conversion. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding; St. Thomas, US Virgin Islands. 2003. p. 676-681.

Sundström, A. Automatic prosody modification as a means for foreign language pronunciation training. Proc. ISCA Workshop on Speech Technology in Language Learning (STILL 98); Marholmen, Sweden. 1998. p. 49-52.

Tang, M.; Wang, C.; Sene, S. Voice Transformations: From Speech Synthesis to Mammalian Vocalizations. Eurospeech; 2001; Aalborg, Denmark. 2001.

Tenenbaum JB, Silva Vd, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290 (5500):2319–2323. [PubMed: 11125149]

Traunmüller H. Conventional, biological and environmental factors in speech communication: a modulation theory. Phonetica 1994;51:170–183. [PubMed: 8052672]

Turk, O.; Arslan, LM. Donor Selection for Voice Conversion. EUSIPCO; 2005; Antalya, Turkey. 2005.

Turk O, Arslan LM. Robust processing techniques for voice conversion. Comput Speech Language 2006;20 (4):441–467.

van Lancker D, Kreiman J, Emmory K. Familiar voice recognition: pattern and parameters. Part I: recognition of backward voices. J Phonetics 1985;13:19–38.

Vieru-Dimulescu, B.; Mareüil, PBd. Contribution of prosody to the perception of a foreign accent:a study based on Spanish/Italian modified speech. Proc. ISCA Workshop on Plasticity in Speech Perception; London, UK. 2005. p. 66-68.

Wachowicz KA, Scott B. Software that listens: it's not a question of whether, it's a question of how. CALICO J 1999;16 (3):253–276.

Watson C, Kewley-Port D. Advances in computer-based speech training: Aids for the profoundly hearing impaired. Volta-Review 1989;91:29–45.

Yan, Q.; Vaseghi, S.; Rentzos, D.; Ho, C-H. Analysis by synthesis of acoustic correlates of British, Australian and American accents. Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, 2004 (ICASSP'04); Montreal, Quebec, Canada. 2004. p. I-637-I-640.

Young, SJ. Tech Rep TR. Department of Engineering, Cambridge University; UK: 1993. The HTK Hidden Markov Model Toolkit: Design and Philosophy; p. 153

Appendix A

To perform multi-dimensional scaling, we first create a (100 × 100) matrix containing the average perceptual distance between any two of the 100 utterances. Shown in Fig. 7a as an image (darker colors indicate larger perceptual distances between the corresponding pair of utterances), this matrix is sparse due to the large number of utterance pairs (10,000) relative to the number of participants. To guard against outliers, we eliminate any utterance pairs that have been rated by only one participant. We use an ε-neighborhood with a radius of 7 perceptual units[9] to define a local connectivity graph; the resulting local distance matrix is shown in Fig. 7b. Geodesic distances between every pair of utterances are then estimated using Dijkstra's shortest paths algorithm (Dijkstra, 1959), which results in the fully connected distance matrix D shown in Fig. 7c.

Following Tenenbaum et al. (2000), we apply an operator $\tau(\cdot)$ to matrix D, which converts distances into inner products:

$$\tau(D) = -\frac{HSH}{2} \tag{3}$$

where S is a matrix containing the squared distances found in D (i.e. $S_{ij}=D_{ij}^2$), H is the centering matrix

$$H = I_N - \frac{1}{N} \tag{4}$$

I_N is an identity matrix, and N is =100. The ith component y_i of the d-dimensional embedding (i.e. the coordinates of the N utterances on the ith dimension of the embedding) is found by

$$y_i = \sqrt{\lambda_p} v_p^i \tag{5}$$

where λ_p is the pth eigenvalue of the matrix $\tau(D)$ and v_p^i is the ith component of the pth eigenvector. Each of the 100 samples is then represented in two dimensions as y_1 and y_2. A two-dimensional embedding of the distance matrix in Fig. 7c is shown in Fig. 5b.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.specom.2008. 11.004.

[9]Scores of 0–7 indicates pairs of utterances that participants believed to have been produced by the same speaker.

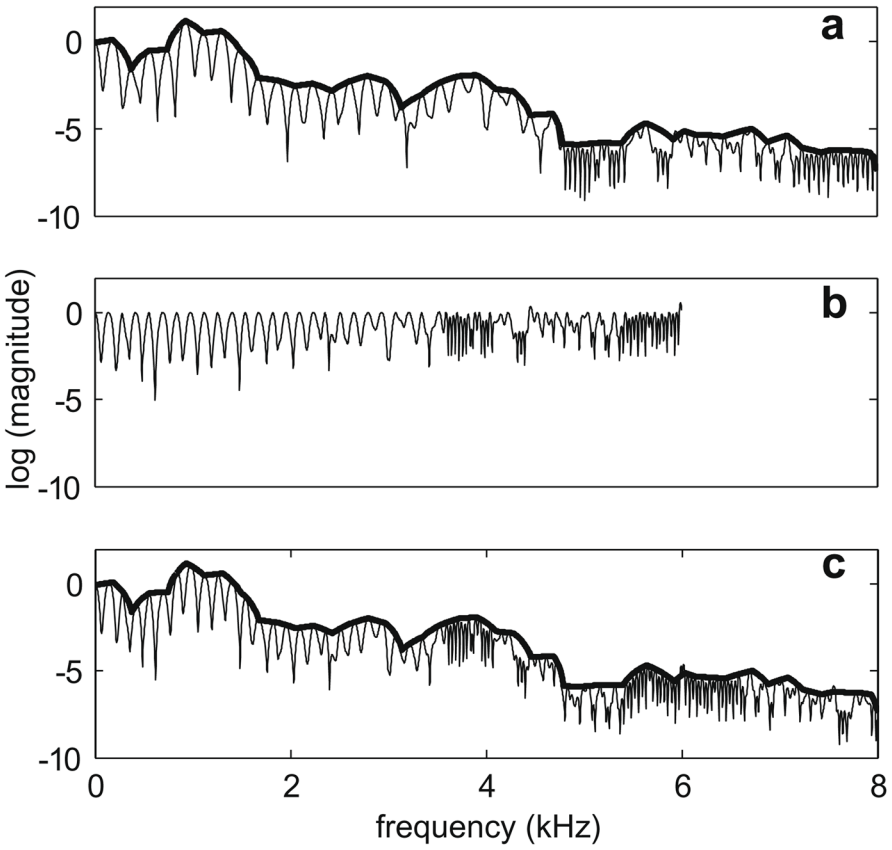

Fig. 1.

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).

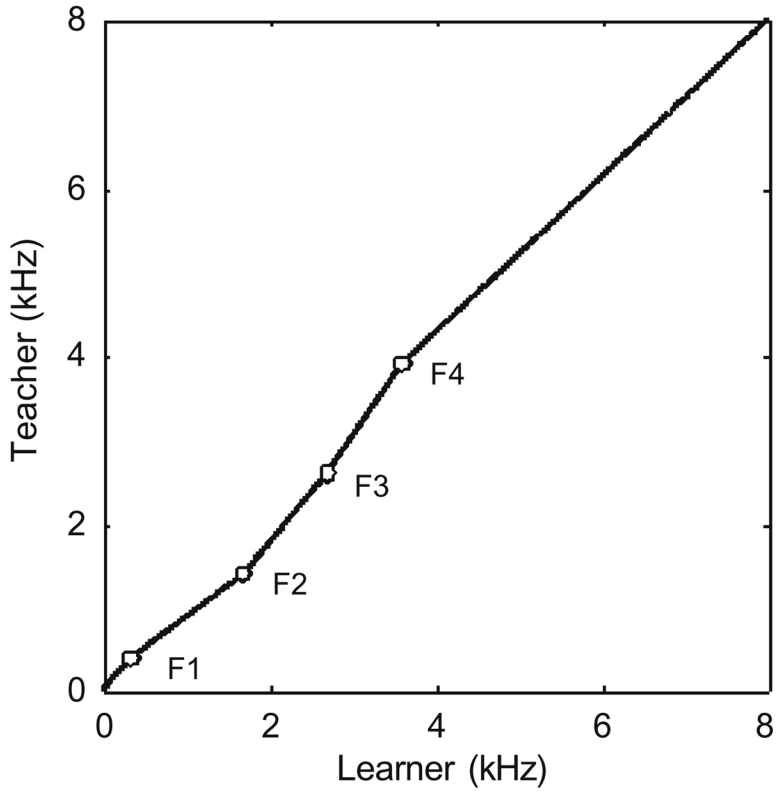

Fig. 2.

The VTLN frequency mapping is created by linearly interpolating between average formant locations for the learner and teacher. This physically-motivated transformation preserves acoustic cues associated with the vocal tract length of the learner.

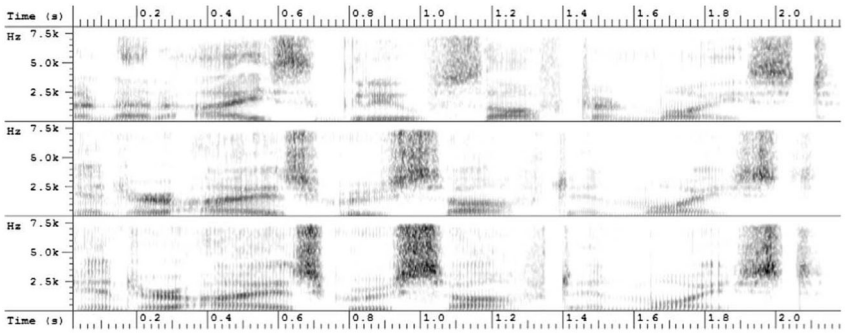

Fig. 3.
Wideband spectrograms of the utterance "…*and her eyes grew soft and moist*." From top to bottom—learner, learner with prosodic and segmental transformation, and teacher. Video samples are available as Supplemental material (1–5.mpg).

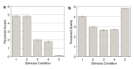

Fig. 4.

(a) Accent ratings showing mean ± standard error for each stimulus category. The segmental transformation significantly reduced accent. (b) Quality ratings showing mean ± standard error for each stimulus category. The transformations significantly reduce quality; note that utterances from the (unmodified) foreign speaker were rated as having lower quality than those from the (unmodified) native speaker. (1 = foreign speaker, 2 = prosodic transformation, 3 = segmental transformation, 4 = prosodic and segmental transformations, 5 = native speaker).

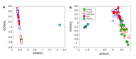

Fig. 5.
Experimental results from the identity tests. Tight clusters may appear as a single point (e.g. as in the case of utterances from the original teacher). (a) Forward speech; ISOMAP reveals three distinct clusters: one for the teacher, one for the learner (with prosodic transformation), and a third cluster with the segmental transformations. (b) Reverse speech; ISOMAP reveals only two clusters: one for the teacher, and a second one for all other utterances.

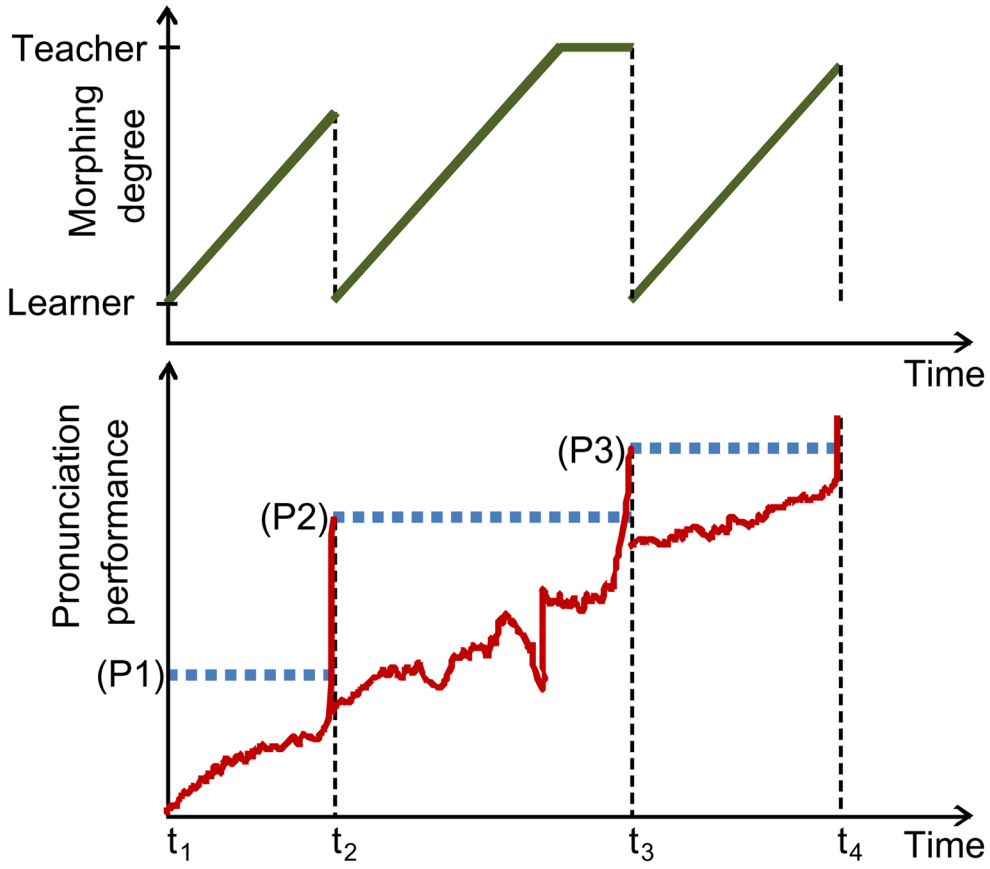

Fig. 6.

Illustration of a training procedure incorporating behavioral shaping and morphing. At time t_1, the learner produces utterance P1 as their first attempt to imitate a native utterance. P1 becomes the initial learner utterance. Between times t_1 and t_2, feedback to the user is in the form of increasingly higher degrees of morphing (see top trace in the figure) between the learner P1 and its full accent converted version. At time t_2, the user produces P2, which exceeds their best production thus far (P1). As a result, P2 becomes the next learner utterance. Between times t_2 and t_3, feedback to the user is again in the form of increasingly higher degrees of morphing between the learner P2 and its full accent converted version. At time t_3, the user produces (P3), which becomes the next learner utterance. In this fashion, utterances fed back to the learner have increasingly higher degrees of native accentedness; first, as a result of the steady increase in morphing; second, because of the increasingly better productions of the learner.

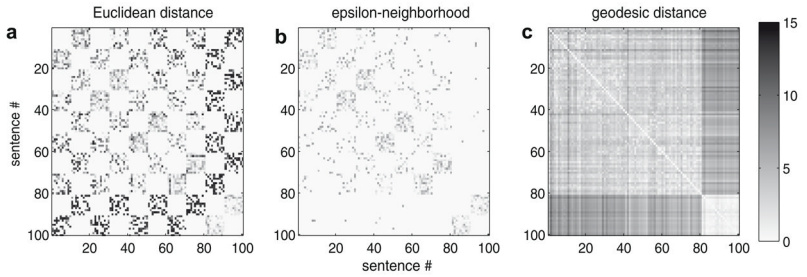

Fig. 7.

Estimating the geodesic distance between utterances; utterances are displayed in groups of 20, corresponding to their stimulus condition (i.e. utterances 1–20 are from condition 1, 21–40 from condition 2, etc.) Dark pixels indicate that (on average) the corresponding pair of utterances was perceived as having been produced by different speakers; the grayscale is shown on the far right. (a) Raw average distances for the identity experiment with reversed speech. A checkerboard pattern appears due to the testing procedure (refer to footnote 3). (b) Data is thresholded to remove distances greater than seven; this separates pairs of utterances that were perceived as "from the same speaker" from those perceived as "from different speakers". Utterance pairs for which data was scarce (less than two examples) were also removed to avoid potential problems with outliers. (c) Fully connected graph reconstructed by Dijkstra's shortest path algorithm. Notice the block structure showing low geodesic distance within utterances from condition 5 (teacher) and within utterances from conditions 1 through 4 (learner's glottal excitation). It is this distribution of geodesic distances that leads to the clusters observed in Fig. 5.

Table 1

Stimulus conditions for the perceptual studies.

#	Stimulus
1	Student utterance
2	Student w/ prosodic-conversion
3	Student w/ segmental conversion
4	Student w/ prosodic & segmental conversion
5	Teacher utterance

Table 2

Combined identity score.

Value	Equivalent meaning
0	Same speaker, very confident
6	Same speaker, not at all confident
7	N/A
8	Different speaker, not at all confident
14	Different speaker, very confident