

Developing objective measures of foreign-accent conversion

Daniel Felps and Ricardo Gutierrez-Osuna, *Senior Member, IEEE*

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

{dlfelps, rgutier}@cse.tamu.edu

Abstract

Various methods have recently appeared to transform foreign-accented speech into its native-accented counterpart. Evaluation of these accent conversion methods requires extensive listening tests across a number of perceptual dimensions. This article presents three objective measures that may be used to assess the acoustic quality, degree of foreign accent, and speaker identity of accent-converted utterances. Accent conversion generates novel utterances: those of a foreign speaker with a native accent. Therefore, the acoustic quality in accent conversion cannot be evaluated with conventional measures of spectral distortion, which assume that a clean recording of the speech signal is available for comparison. Here we evaluate a single-ended measure of speech quality, ITU-T recommendation P.563 for narrow-band telephony. We also propose a measure of foreign accent that exploits a weakness of automatic speech recognizers: their sensitivity to foreign accents. Namely, we employ the HTK recognizer trained on a large number of English American speakers to obtain a measure of nativeness for different accent-converted utterances. Finally, we propose a measure of speaker identity that extracts the unique discriminatory acoustic cues for a given pair of source and target speakers through Linear Discriminant Analysis. The three measures are evaluated on a corpus of accent-converted utterances that had been previously rated through perceptual tests. Our results show that the three measures have a high degree of correlation with their corresponding subjective ratings, suggesting that they may be used to accelerate the development of foreign-accent conversion tools. Applications of these measures in the context of computer assisted pronunciation training and voice conversion are also discussed.

Keywords: accent conversion, voice conversion, speaker recognition, foreign accent recognition.

1 INTRODUCTION

Older learners of a second language (L2) typically speak with a so-called “foreign accent,” sometimes despite decades of immersion in a new culture. During the last two decades, a handful of studies have suggested that it would be beneficial for these learners to be able to listen to their own voices producing native-accented utterances [1, 2]. The rationale is that, by stripping away information that is only related to the teacher’s voice quality, learners can more easily perceive differences between their accented utterances and their ideal accent-free counterparts. A foreign accent is manifested primarily as deviations from the expected segmental (e.g. formants) and prosodic (e.g. intonation, duration, and rate) norms of a language [3]. This suggests that signal processing techniques similar to those employed in voice conversion [4-9] may also be used to transform foreign-accented utterances into their native-accented versions. As a result, a number of studies have recently appeared on the subject of foreign accent conversion [10-12]. Notwithstanding their similarities, voice conversion and foreign accent conversion have orthogonal goals. Voice conversion seeks to transform utterances from a speaker so they sound as if another speaker had produced them. In contrast, 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. Thus, foreign accent conversions must be evaluated according to multiple criteria, including not only the degree of foreign accent reduction and acoustic quality of the transformation but also the extent to which the voice quality of the foreign speaker has been preserved. These evaluations are challenging for multiple reasons. First, some of the above criteria can be conflicting; as an example, the perceived identity of a foreign speaker may be inextricably coupled with his/her accent [13] to where removal of the foreign accent leads to the perception of a different speaker. Moreover, whether or not a speaker is perceived to have a foreign accent depends on the dialect and exposure of the listener [14]. Finally, and more importantly, because foreign accent conversion seeks to generate utterances that have never been produced (i.e. those of an L2 learner having a native accent), no ground truth is available against which the transformations can be tested, and perceptual studies must be employed at every stage in the process.

The specific objective of this work is to develop objective measures of acoustic quality, foreign accentedness and speaker identity that are consistent with perceptual evaluations. Such measures would be invaluable in a number of scenarios. As an example, the ability to objectively rate synthesized utterances may be used to search and fine-tune parameters in accent conversion systems –our immediate motivation. Objective measures may also be used in computer assisted pronunciation training (CAPT) to match the voice of the L2 learner with a suitable voice from a pool of native speakers, or to provide feedback to the learner, which is a critical issue in CAPT [15, 16]. The work proposed here is related to but distinct from prior research in automatic accent classification, speaker identification and verification, and speech enhancement. In accent classification, the goal is to assign a given speech recording to one of several accent/dialect categories. In contrast, our goal is to assign a foreign accentedness score that correlates with ratings of perceived accentedness, which are continuous rather than discrete. Thus, our problem is one of regression rather than classification. In speaker verification/identification, the goal is to determine the veracity of a speaker’s claimed identity or assign a speaker’s voice to one of several known speakers. In contrast, our goal is to obtain a measure of similarity between a synthesized voice (i.e. accent converted) and that of two reference speakers (i.e. a native and a foreign speaker). Finally, most measures of acoustic quality in speech enhancement rely on the existence of a ground truth to compute a measure of spectral distortion. In contrast, and as mentioned earlier, accent conversion (if successful) leads to utterances that have never been produced, and for which a ground truth does not exist.

2 RELATED WORK

2.1 Foreign accent conversion

The relative contribution of various acoustic cues to the perception of a foreign accent has been extensively studied [3, 13, 17-19]. Due to space constraints, however, this review will focus on studies that have manipulated acoustic cues for the specific purpose of converting a foreign accent into its native counterpart (or vice versa). These studies have been organized according to whether they have concentrated on prosodic features or have also considered transformation of segmental cues.

Prosodic transformations are, by far, the most common approach to foreign accent conversion. This is motivated by the fact that prosody plays a very significant role in foreign accents [17] and also by the availability of methods for time- and pitch-scaling [20]. Tajima et al. [21] investigated the effect of temporal patterning on the intelligibility of speech. The authors used dynamic time warping and LPC resynthesis to modify the timing of English phrases spoken by native Chinese speakers and native English speakers. When utterances by Chinese speakers were distorted to match the timing of English speakers, intelligibility increased from 39% to 58%, as measured by native English *listeners*. Likewise, when utterances by English speakers were distorted to match the timing of Chinese speakers, intelligibility declined from 94% down to 83%. Cho and Harris [22] developed an automatic tool to transform the prosody of L2 speech. The authors collected a corpus of English utterances produced by native speakers of American English and by Korean speakers. Utterances were time aligned through dynamic time warping, and then transformed in duration and pitch by means of PSOLA [23]. Four types of stimuli were evaluated by native listeners of American English: utterances by Korean speakers with (1) Korean intonation and (2) American intonation, and utterances by American speakers with (3) American intonation and (4) Korean intonation. Resynthesizing American utterances with Korean intonation increased their foreign-accentedness from 1.24 to 2.10 (on a 7-point scale), whereas resynthesizing Korean utterances with an American intonation reduced their foreign-accentedness from 5.08 to 4.75. Prosodic conversion has also been explored in the context of CAPT. Nagano and Ozawa [24] evaluated a prosodic-conversion method to teach 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. Their results show that 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. [25] investigated the use of prosodic modification to teach German prosody to Italian speakers. Their results were consistent with [24], and indicate that the learner's own voice (with corrected prosody) was a more effective form of

feedback than prerecorded utterances from a German native speaker.

Segmental techniques for foreign accent conversion have been investigated only recently. Yan et al. [26] proposed an accent-synthesis method based on formant warping. First, the authors developed a formant tracker based on hidden Markov models (HMM) and linear predictive coding (LPC), and applied it to a corpus containing several regional English accents (British, Australian, and American). Second, the authors re-synthesized utterances by warping formants from a foreign accent onto the formants of a native accent; pitch- 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. Kamiyama [27] investigated the perception of French utterances produced by Japanese learners. The study consisted of eight short phrases read by Japanese and French speakers. Six types of resynthesized utterances were evaluated, involving all combinations of three segmental conditions (European French, Canadian French, and Japanese phonemes) with two prosodic conditions (French and Japanese). MBROLA [28] and PRAAT [29] were used to perform segmental and prosodic modifications, respectively. Results from this study indicate that both segmental and prosodic characteristics contribute to the perceptual rating of accentedness, though prosody plays a more significant role. More recently, Huckvale and Yanagisawa [12] 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. We have also investigated the role of prosodic and segmental information on the perception of foreign accents [11]. Our work differs from [26] in two respects. First, our accent conversion method (described in section 3.1) 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. As discussed earlier, the latter is critical

because a successful accent-conversion model should preserve the identity of the foreign-accented speaker. In contrast with [12], our study was performed on natural speech, and focused on accentedness and identity rather than on intelligibility; as noted by Munro and Derwing [30], a strong foreign accent does not necessarily limit the intelligibility of the speaker.

2.2 Acoustic correlates of foreign accent, acoustic quality, and speaker identity

2.2.1 Acoustic quality

Objective measures of quality can be broadly described as either *intrusive* or *non-intrusive*. Intrusive measures evaluate the quality of modified speech against the original, high-quality reference speech. The International Telecommunication Union (ITU-T) recommendation for end-to-end speech quality assessment is P.862, which achieves an average correlation of 0.94 with subjective Mean Opinion Scores (MOS). Such intrusive models are ideal for testing coding or transmission systems because the original, unmodified speech is available for comparison. However, they are not appropriate for voice conversion systems; though a well-defined ground truth exists in this case (i.e. the voice of the target speaker), it is unrealistic to expect a transformed utterance to match the target exactly. For that matter, intrusive models are even more questionable for accent conversion systems because the latter lack a well-defined target.

Non-intrusive measures of speech quality must be used when reference signals are too costly or impossible to obtain, in which case one must predict quality based on the test speech itself. Non-intrusive measures are well suited for testing satellite systems [31], voice over IP, and cell phone networks [32]. The most common approach is to create a model of clean speech (e.g. with vector quantization [31]) to serve as a pseudo-reference signal. The average distance to the nearest reference centroid provides an indication of speech degradation, which can then be used to estimate subjective quality. Models of the vocal tract [33] and the human auditory system [34] have also been proposed. However, the prevailing non-intrusive measure is ITU-T recommendation P.563 [32], which is discussed in section 3.3.1.

2.2.2 Foreign accentedness

Speaker adaptation is a prevalent topic in speech recognition research as it helps decrease speaker variability due to differences in gender, physiology, or accent [35]. Such investigations have also led to objective measures of accent [36]. These can be grouped into three categories [37]: methods that model the global acoustic distribution, methods based on accent-specific phone models, and analysis of pronunciation systems. The first approach models the distribution of acoustic vectors from speakers of a particular accent; e.g. formant frequencies of standard English vowels [10]. Classification is then achieved through pattern recognition, e.g. Gaussian mixture models (GMM) [38, 39].

Accent-specific phone models have been explored by Arslan and Hansen [40]. Their method evaluated words sensitive to accent on separate HMM word recognizers trained for each accent (e.g. English, Turkish, German, or Chinese). The accent chosen was the one associated with the HMM that yielded the highest likelihood; their method compared favorably against classification performance by human listeners. Other researchers have also taken a similar approach [41].

The final group takes a linguistic approach to accent classification, one which may be more sensitive than methods based on acoustic quality [37]. In one of the earlier papers, describing accent classification for speech recognition, Barry et al. [42] compared acoustic realizations *within* a particular speaker. By analyzing systematic differences (or similarities), the authors were able to separate four regional English accents; e.g. Northern English uses the same vowel for “pudding” and “butter,” but American English uses different vowels. Once such phonemic relations are established, it is then sufficient to evaluate the accent of a speaker based on a single sentence that exploits this information. Related approaches analyze a speaker’s phonetic tree (created through cluster analysis) to determine accent [43].

2.2.3 Speaker identity

As objective measures of accent have been derived from speaker adaptation methods, so have objective measures of identity come from research in speaker recognition. Early investigations focused on

identifying suitable features for discrimination (e.g. pitch and formant frequencies), but recent advances have come from improved machine learning techniques (e.g. Gaussian mixture models). While some results on feature selection [44] indicate a stronger relationship between identity and spectral measures than between identity and pitch, others show a preference for pitch [45]. Malayath et al. [46] proposed a multivariate method to separate the two main sources of variability in speech: speaker identity and linguistic content. Namely, the authors used oriented PCA to project an acoustic feature vector (LPC-cepstrum) into a subspace that minimized speaker-dependent information while maximizing linguistic information; this method may also be used for the opposite problem: capturing speaker variability while reducing linguistic content. Lavner et al. [47] investigated the relative contributions of various acoustic features (glottal waveform shape, formant locations, F0) to the identification of familiar speakers. Their results indicate that shifting the higher formants (F3, F4) has a more significant effect than shifting the lower formants, and that the shape of the glottal waveform is of minor importance provided that F0 is preserved. More interestingly, the study found that the very same acoustic manipulations had different effects on different speakers, which suggests that the acoustic cues of identity are speaker-dependent.

Once the foundation for appropriate acoustic features was established, researchers turned their attention toward classification techniques [48]. Recent text-independent speaker identification systems often employ GMMs. Typically, a separate GMM is trained for each of the speakers in question, and an unknown speaker is identified when the likelihood of a given utterance exceeds a threshold for one of the models. GMMs have also been used as objective measures of identity for voice conversion [49].

3 METHODS

3.1 Foreign accent conversion

According to the modulation theory of speech [50], a speaker’s utterance results from the modulation of a voice-quality carrier with linguistic gestures. 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 [51], 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. Such is the underlying motivation behind our accent-conversion method, which is briefly reviewed here; further details may be found in [11]. Our method proceeds in two distinct steps. First, prosodic conversion is performed by modifying the phoneme durations and pitch contour of the (foreign-accented) source utterance to follow those of the (native-accented) target. Second, formants from the source utterance are replaced with those from the target.

To perform *time scaling*, we assume that the speech has been phonetically segmented by hand or with a forced-alignment tool [52]. From these phonetic segments, the ratio of source-to-target durations is used to specify a time-scaling factor α for the source on a phoneme-by-phoneme basis ($0.25 \leq \alpha \leq 4$). Our *pitch-scaling* 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, limited to pitch-scale factors β in the range $0.5 \leq \beta \leq 2$. This process allows us to preserve the identity of the source speaker by maintaining the pitch baseline and range [53], while acquiring the pitch dynamics of the target speaker, which provides important cues to native accentedness [54]. Once the time- and pitch-scale modification parameters (α, β) are calculated, Fourier-domain PSOLA [23] is used to perform the prosodic 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 content. Thus, our strategy consists of combining the target's spectral envelope (filter) with the source's glottal excitation. For each source and target analysis window, we first apply SEEVOC [55] to decompose the signal into its spectral envelope and a flattened excitation spectrum, and then multiply the spectral envelope of the target with the flattened spectral excitation of the source. The modified short-time spectra are transformed back to the time domain, and concatenated using a least-squared-error criterion [56]. In order to reduce speaker-dependent information in the target's spectral envelope, we also perform vocal tract length normalization using a piecewise linear function defined by the average formant pairs of the two speakers [7]; formant locations are estimated with PRAAT [29] over the entire corpus. The result is a signal that contains the source's excitation and the target's spectral envelope normalized to the source's vocal tract length.

3.2 Perceptual evaluation

In [11], we performed a series of perceptual experiments to characterize the method in terms of (1) the degree of reduction in foreign accentedness, (2) the extent to which the identity of the original speaker had been preserved, and (3) degradations in acoustic quality. 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 audio files for the five conditions are available as supplemental material (1-5.wav and rev1-5.wav). Perceptual evaluation consisted of three independent experiments:

- *Acoustic quality.* Following [6], participants were asked to rate the acoustic quality of utterances on a standard MOS scale from 1 (bad) to 5 (excellent). Before the test began, participants listened to examples of sounds with various accepted MOS values.
- *Foreign accentedness.* Following [57], participants were asked to rate the foreign accentedness of utterances using 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) [58].

- *Speaker identity.* Following [59], participants listened to a pair of linguistically different utterances, and were asked to (i) determine if the two sentences were produced by the same speaker, and (ii) rate their confidence on a 7-point EGWA scale. These two responses were converted into a 15-point perceptual score (Table 2). To prevent participants from using accent as a cue to identity, utterances were played backwards. This 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 [60].

Table 1. Stimulus conditions for perceptual studies

#	Stimulus
1	Source utterance
2	Source w/ prosodic conversion
3	Source w/ segmental conversion
4	Source w/ prosodic & segmental conversion
5	Target utterance

Table 2. Combined score for identity ratings

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

3.3 Objective measures

3.3.1 Acoustic quality

Our objective measure of acoustic quality is based on recommendation P.563 for single-ended (i.e. no reference) speech quality [32], which is freely available for download from the ITU-T website. The algorithm operates in three stages: preprocessing, distortion estimation, and perceptual mapping. During preprocessing, the signal level is normalized to -26.0 dBov. Next, an additional version of the speech is created by filtering it with a response similar to the properties of a standard telephone. A third version is filtered with a fourth-order Butterworth high-pass filter with a 100Hz cutoff. Finally, voiced areas are detected using ITU-T recommendation P.56.

P.563 makes use of several distortion measures to identify the various types of distortions that may be present in the signal. The first measure, based on speech production, approximates the vocal tract area function. This is accomplished by transforming the coefficients of an eighth order pitch-synchronous LP

analysis (via their reflection coefficients) into an area function with eight tubes [33]. These eight tubes are then divided into three groups: front, middle, and rear cavity (corresponding to tubes 1-3, 4-6, and 7-8). Sudden changes in the areas of any of the three cavities are indications of distortion. A second measure of distortion simulates an intrusive quality measure with a reference signal being provided by a speech reconstruction module. The module is designed to remove or modify the noise in the distorted speech signal. The reconstructed speech is then compared with the distorted speech using a psychoacoustic model similar to the one found in ITU-T P.862 (see section 2.2.1); this step measures the amount of distortion removed by the speech reconstruction module. The final measures of distortion include estimation of SNR and detection of robotization, temporal clipping, and signal correlated noise.

The final stage, perceptual mapping, takes the above measures of distortion and calculates the final MOS score with a classifier followed by a regression model. The classifier identifies which of seven types of degradations are most likely to be present (i.e. robotization, interruption and clipping, signal correlated noise, low SNR, unnaturally low pitch, unnaturally high pitch, or [default] general distortion). Quality is then estimated on a standard MOS scale using a regression model trained on examples from that class.

P.563 shows an average correlation of 0.85 with subjective MOS [32]; as a comparison, its intrusive counterpart P.862 achieves a correlation of 0.94. The ITU-T further recommends that, when evaluating a system, multiple speech files be tested and their scores averaged. Suggested applications for P.563 include live network monitoring using digital or analog connections, live network end-to-end testing using digital or analog connections, and live network end-to-end testing with unknown speech sources at the far end side. Despite the fact that P.563 is not intended to measure the quality of transformed speech, we find that it yields reasonable results when at least twenty sentences are averaged per condition.

3.3.2 Foreign accentedness

The objective measure of accent adopted in this study is related to that in [40]. In our method, however, we evaluate a test utterance on a continuous speech HMM (trained on acoustic models from native speakers of American English), and the match score is used as an estimate of its degree of *nativeness*. The

primary advantage to this approach is that, as long as the target accent remains American English, then one need not train a separate HMM for an arbitrary source accent. Our implementation of the continuous speech recognizer is based on the freely-available Hidden Markov Model Toolkit (HTK) [52]. We use acoustic models trained on 284 North American speakers [61] coupled with the CMU pronunciation dictionary [62]. There is no need for a language model since we operate in forced-alignment mode. We choose this over standard speech recognition in order to constrain the desired pronunciation to that of an American English dialect. The objective score for an utterance is given as the median value of the phoneme-level match scores contained in the label file, excluding those associated with silences.

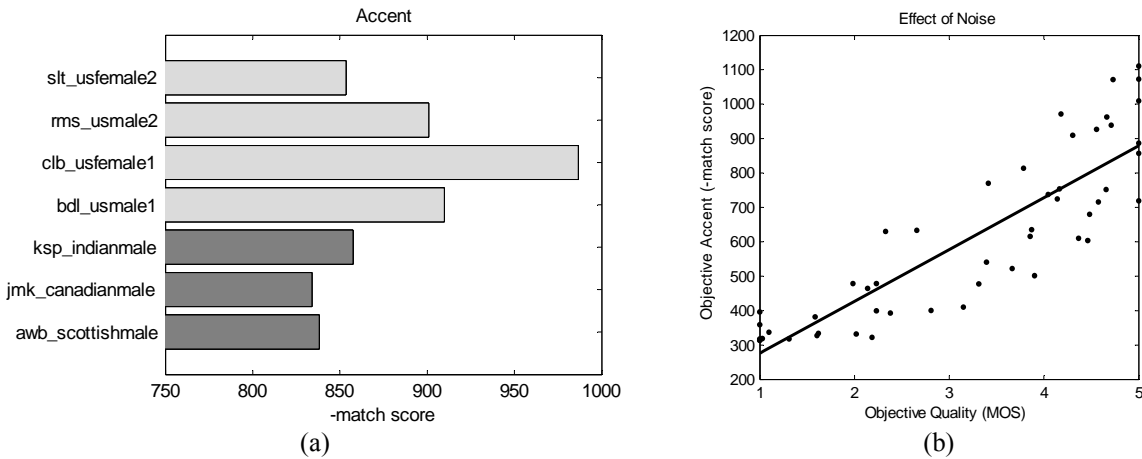


Figure 1. (a) Average HTK match scores for the seven speakers in ARCTIC. Native speakers are represented by darker color bars. (b) Effect of additive Gaussian noise on the HTK score. Ten sentences with ten levels of noise spanning the full range of MOS values were used for this purpose.

To test the effectiveness of this measure against multiple speakers (both native and foreign) we evaluated the seven speakers (four native and three foreign) of the CMU ARCTIC database [63]. Twenty utterances from each speaker were averaged for a final score. Results are summarized in Figure 1(a); differences between the two groups were found to be significant; $t(5)=-2.1$, $p<0.05$. To determine whether the speakers' dialect was the primary factor affecting the HTK match score, we also performed a preliminary experiment with ten sentences, each replicated with ten levels of additive white noise spanning the range of objective quality measures. Results in Figure 1(b) show a strong correlation $r(98)=-0.87$, $p<0.001$

between our objective measures of accent (HTK match scores) and quality (ITU-T P.862), which represents a decrease of about 150 points in match score with each level of quality. Accordingly, HTK scores were be corrected according to this trend in order to obtain a measure of accentedness that was (linearly) independent from acoustic quality.

3.3.3 Speaker identity

Our objective measure of speaker identity is based on a signal discrimination criterion [64]. Namely, given a corpus of acoustic features from source and target speakers, we find a projection that maximizes the separability between both speakers by means of Fisher's Linear Discriminant Analysis (LDA). This approach compares favorably against conventional methods for speaker recognition based on GMMs. The primary advantage stems from the fact that LDA is a supervised method, whereas GMMs are unsupervised. Namely, GMMs are trained to model the distribution of data in feature space without regard to a feature's discriminatory ability or noise level. LDA, on the other hand, finds the subspace with the highest discriminatory information. This is particularly advantageous when acoustic features are poorly selected or when the source and target have broadly overlapping distributions in feature space. In addition, as demonstrated by Lavner et al. [47], discriminatory features may also change for each source/target pair; LDA will automatically adapt in such a situation. Moreover, the computational requirements for LDA are also significantly lower; as shown below, a solution is found through a single matrix inversion, whereas EM is a fixed-point method. Finally, for binary discrimination problems, the LDA solution is a single dimension, which facilitates interpretation. In summary, we find LDA to be a powerful yet efficient solution for determining an objective measure of speaker identity.

Following [64], consider the problem of discriminating between two classes on the basis of a D -dimensional feature vector x . LDA seeks a linear projection vector w such that the projected data $y = w^T x$ maximizes the distance between classes relative to the variance within each class. It can be shown that, for Gaussian distributed classes with equal covariance, the optimal linear projection is

$$w = S_W^{-1}(m_2 - m_1) \quad (1)$$

where S_W is the between-class scatter, and m_1, m_2 are the sample mean of the two classes:

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T \quad (2)$$

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n \quad m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

For accent/voice conversion, the two classes correspond to the source and target speakers, and the feature vector x is a vector of acoustic parameters for each speech frame (F0 and 13 MFCCs in this work). In our implementation, one hundred sentences from each speaker are analyzed (in 20 ms frames) to generate a training set. To avoid overfitting, these sentences are different from those later used for testing (refer to section 4.2), and do not include any accent conversions. Once the Fisher's LDA solution w has been computed from training data according to (1), each new test sentence is framed, each frame is analyzed to obtain acoustic vector x and this vector is projected to obtain a score ($y = w^T x$). Each test sentence is then assigned an identity score that corresponds to the average score across its frames.

4 RESULTS

We validated the proposed objective measures against results from a perceptual evaluation of our accent conversion model previously reported in [11]. In what follows, we use a two-way ANOVA to calculate statistical significance, with the two factors being the prosodic and segmental transformations.

4.1 Acoustic quality

Forty-three students participated in a 25-minute test to rate the perceived quality of recorded/synthesized utterances. Results are summarized in Figure 2. Original recordings from the target (native) speaker received the highest average rating (4.84), followed by those from the source (foreign) speaker (4.0); this difference was statistically significant, $t(19)=-21.42$, $p<0.001$ (two-tailed). Though recording conditions may have been different for both speakers, it is also possible that subjects penalized the “quality” of non-native speech because it was less intelligible. All transformations lowered quality ratings with respect to the original recordings. Two-way ANOVA found all effects significant: main prosodic, $F(1,76)=48.48$, $p<0.001$; main segmental, $F(1,76)=119.14$, $p<0.001$; and interaction, $F(1,76)=57.31$, $p<0.001$.

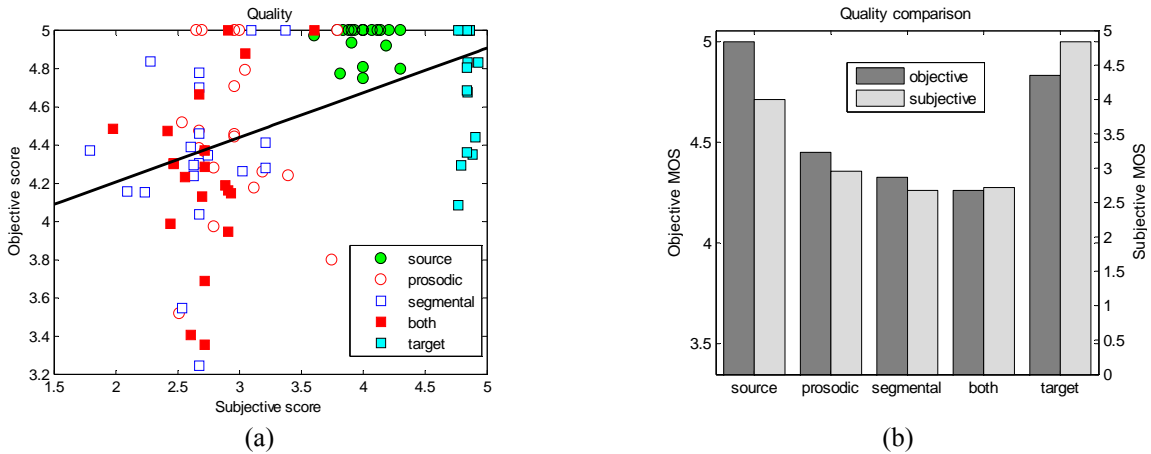


Figure 2. (a) Correlation between objective and subjective measures of acoustic quality; each sample in the scatterplot represents an utterance. (b) Average scores for the two measures across experimental conditions. The objective measure follows a similar trend as the subjective scores.

The objective measure also captured the effects: main prosodic, $F(1,76)=11.30$, $p<0.005$; main segmental, $F(1,76)=23.76$, $p<0.001$; interaction, $F(1,76)=5.26$, $p<0.05$. In other words, the modifications induced by the prosodic and segmental conversions create detectable distortions in the output. Interestingly, the objective measure shows that the source recordings are of higher quality than those of the target, $t(19)=3.18$, $p<0.005$, which gives support to the hypothesis that it is not the quality of the recording but the lower intelligibility of the foreign-accented speech that prompted listeners to give it

lower ratings than to the native-accented speech.

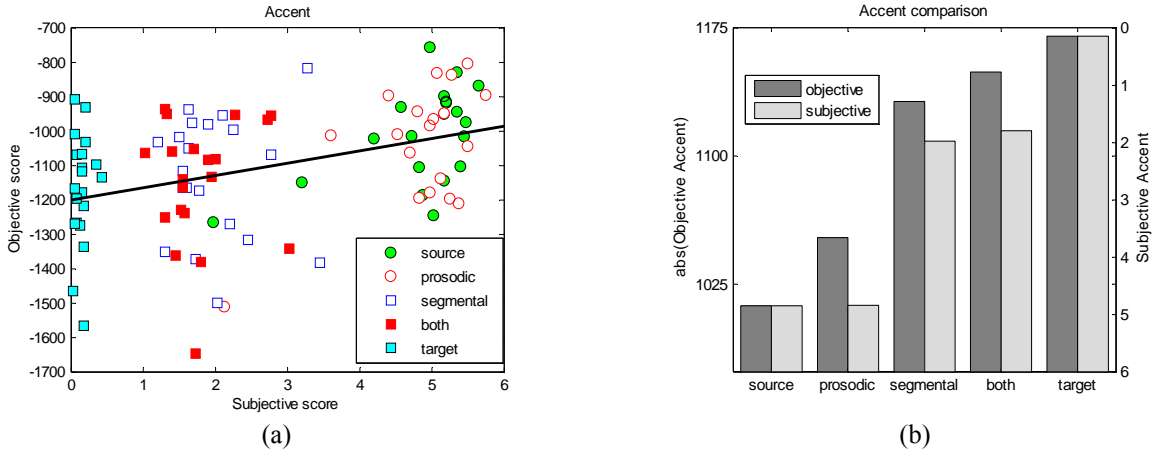


Figure 3. (a) Correlation between objective and subjective measures of accent. (b) Average scores for the two measures across experimental conditions. The objective measure follows the same trend as the subjective scores. Left and right y-axes were aligned to facilitate the comparison. The subjective scale (foreign accentedness) was also inverted to match the sign of the objective score (nativeness).

4.2 Foreign accentedness

Thirty-nine students participated in a 25-minute test to establish the degree of accentedness of individual utterances. Results are summarized in Figure 3. As expected, original recordings from the source (foreign) speaker received the highest average accent rating (4.85), while target (native) recordings scored the lowest (0.15). The main effect of the segmental transformation was significant, $F(1,76)=343.03$, $p<0.001$, indicating that subjects detected a noticeable difference between the accent of the source (4.85) and the accent of the segmental conversion (1.97). The other effects (i.e. a main effect of prosody and interaction effects) were not significant. This was an unexpected result, as previous studies with altered prosody have shown significant effects [2, 24, 25]. One possible explanation for this finding is that the prosody of the source speaker was close to that of a native speaker, when compared to his segmental productions. An alternative explanation is offered by the elicitation procedure in ARCTIC (Kominek and Black, 2003), since read speech is prosodically flat when compared to spontaneous or conversational speech [65]. The objective measure showed identical trends; the main effect of the segmental

transformation was significant, $F(1,76)=7.84$, $p<0.01$, and the other effects were not.

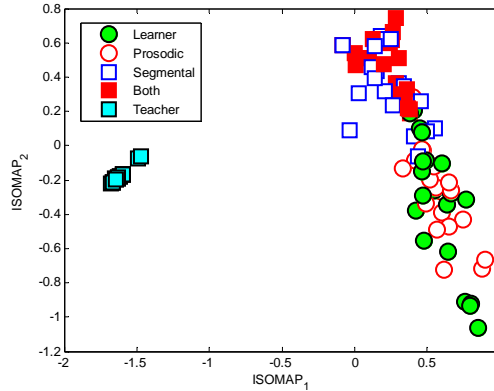


Figure 4. (a) Experimental results from the identity tests; ISOMAP reveals only two clusters: one for the source, and a second one for all other utterances.

4.3 Speaker identity

Forty-three students participated in a 25-minute speaker identification test, which yielded a collection of perceptual distances between pairs of utterances. Because only the relative distance between stimuli is available, we employ multidimensional scaling (MDS) to find a low-dimensional visualization that preserves those pair-wise distances. Namely, we use ISOMAP [66], an MDS technique that attempts to preserve the geodesic distance between stimuli. ISOMAP visualizations of the identity tests, shown in Figure 4(a), reveal two distinct clusters, one containing utterances from the target (native) speaker and a second cluster containing all other conditions. These results indicate that participants could clearly discriminate target utterances from the rest, and that all accent conversions (segmental, prosodic, both) were perceived as being closer to the source (foreign) speaker than to the target speaker. Analysis of the first ISOMAP dimension shows a main effect for the segmental conversion, $F(1,76)=103.79$, $p<0.001$, but no main effect for the prosodic conversion or interaction effects. This indicates that participants were also able to perceive differences between conditions 1 (source) and 3 (segmental), but not between 1 (source) and 2 (prosodic) or between 3 (segmental) and 4 (prosodic + segmental). Figure 4(b) plots the first ISOMAP projection against the LDA projection, which reveals a high correlation (-0.94) between

subjective and objective measures. Moreover, the objective measure is found to be more sensitive as it shows a main effect not only for the segmental conversion, $F(1,76)=114.16$, $p<0.001$, but also for the prosodic conversion, $F(1,76)=4.35$, $p<0.05$. The objective measure shows no interaction effects.

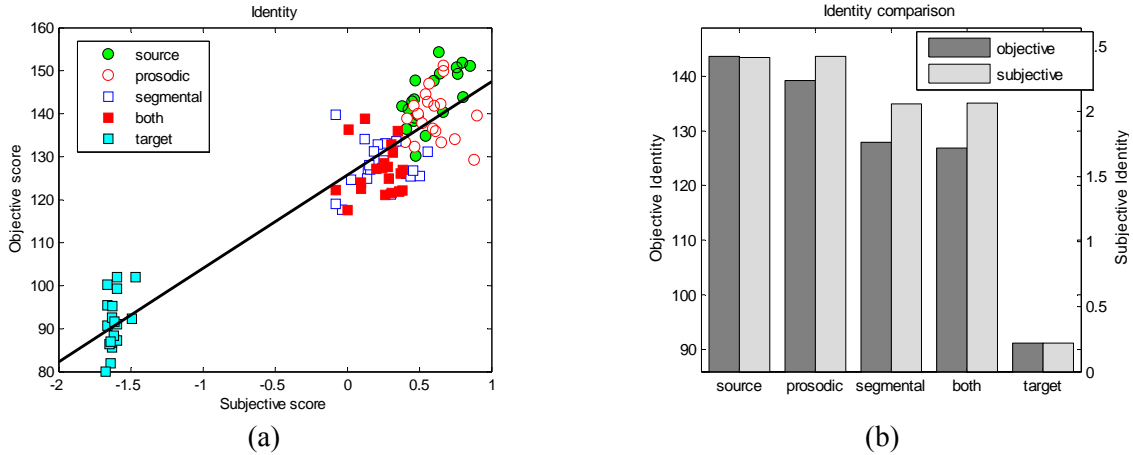


Figure 5. (a) Correlation between objective and subjective measures of identity. (b) Average scores for the two measures across experimental conditions. (b) Experimental results from the identity tests. The objective measure follows the same trend as the subjective scores.

5 DISCUSSION

We have proposed objective measures that can be used to assess the acoustic quality, degree of foreign accentedness and speaker identity of utterances. The three measures show a high degree of correlation across conditions with their corresponding subjective ratings. The correlation coefficient between both measures of acoustic quality shown in Figure 2(b) is $r(3)=0.80$, $p=0.11$; this coefficient increases to $r(3)=0.96$, $p<0.01$ for the accent measures shown in Figure 3(b) and to $r(3)=0.99$, $p<0.005$ for the identity measures shown in Figure 5(b). When computed across sentences, these correlation coefficients become $r(98)=0.47$, $p<0.001$ for the acoustic quality scores in Figure 2(a), $r(98)=0.39$, $p<0.001$ for the accent scores in Figure 3(a), and $r(98)=0.94$, $p<0.001$ for the identity scores in Figure 5(a). These results indicate that the acoustic quality and identity measures need to be computed across multiple sentences; this finding is not surprising since the ITU-T recommends that quality scores with the P.563 standard be obtained as the average across a number of recordings.

No attempts were made in our study to match the scales between objective and subjective measures. As an example, the foreign accent ratings in Figure 3(b) have a different scale on the objective and subjective measures. This issue may be easily addressed by mapping the HTK scores into the 7-point perceptual scale with a regression model. HTK scores may also be converted into an absolute scale by normalizing relative to a corpus of native and foreign-accented speech; see Figure 1(a). However, these extra steps become unnecessary if all one needs are relative measurements, as is the case when optimizing model parameters in an accent conversion system. In this case, it is not the absolute value of the accent measure that is important, but whether it is higher (or lower) than the accent measure for a different set of model parameters; such information is sufficient to guide the optimization engine. Our results also show scaling differences between objective and subjective measures of acoustic quality, despite the fact that P.563 provides a measure in a MOS scale. These results may indicate a downward bias in our perceptual experiments, despite the fact that participants were provided speech samples with various accepted MOS scores. It seems more likely, however, that P.563 under-penalizes utterances resynthesized with our accent conversion model since P.563 focuses on degradations in narrow-band telephony rather than in speech transformations. Sidestepping these scaling differences, however, our results indicate that the three objective measures are remarkably consistent with perceptual ratings when averaged across sentences.

Unlike the acoustic quality and foreign accent ratings (both objective and subjective), which have a monotonic scale, speaker identity ratings must be interpreted relative to the source and target speakers. As an example, consider the identity score for segmental conversions reported by LDA, a value of $y = 128$ (arbitrary units) averaged across 20 utterances. This value can only be interpreted when compared against the scores for the source ($\mu_S = 143$) and target ($\mu_T = 91$) speakers: segmental conversions are significantly closer to the source than to the target. When projected on the LDA solution, utterances from our three accent conversions (segmental, prosodic, and segmental + prosodic) lie somewhere between source and target utterances, a reasonable result considering that these conversion combine elements from both speakers (glottal excitation, prosody, formants, and vocal tract length). However, it is possible that

an utterance may project outside of the bounds defined by the source and the target. This suggests that the LDA scores should eventually be mapped into a measure of distance relative to the source and/or to the target. As an example, a radially symmetric kernel of the form $d = e^{(y-\mu_S)^2/(\mu_S-\mu_T)^2}$ may be used to transform the LDA projection into a measure that denotes how close an utterance is to the source.

5.1 Potential applications

Speed is an important factor that has not yet been discussed, though it was a major motivation for this work. The ability to evaluate converted utterances in a rapid, unbiased manner is extremely useful for research and development in foreign accent conversion. Time invested in developing these objective measures is quickly returned through time saved by more rapid prototyping and parameter tuning. Admittedly, intermediate development steps are rarely evaluated by formal listening tests, but sidestepping subjective evaluations (even informal ones) is necessary to be able to perform an online optimization of parameters.

Beyond this specific motivation, we believe that the proposed objective measures would be invaluable in various scenarios. Take for instance the case of pronunciation training, where traditional methods involve exercise, practice, and feedback between a student and a human teacher. Although not as effective as human instruction, CAPT offers several advantages in applied settings, such as allowing users to follow personalized lessons, at their own pace, and to practice as often as they like while reducing potential sources of anxiety and embarrassment. In this context, Probst et al. (2002) have shown that learners who imitate a well-matched speaker improve their pronunciation more than those who imitate a poor match, suggesting the existence of a user-dependent “golden speaker.” Thus, accent conversion may provide learners with the optimal “golden speaker”: their native-accented selves. This would require matching the voice of the learner (the source speaker) with the voice of a teacher (the target). Because the success of the accent conversions depend on the source-target pair (see e.g. [67] for the “donor selection” problem in voice conversion), objective measures may be used to a suitable accent donor from a pool of native speakers. Objective measures may also be used to provide feedback to the learner, which is a critical issue

in CAPT [15, 16]. As an example, measures of foreign accent may be used to track the learner’s progress over time and adapt the CAPT tool accordingly, for instance by increasing the complexity of the exercises as the learner improves her pronunciation; these strategies are known as “behavioral shaping” [1].

Our objective measures have been tailored for accent conversion, but they can be adapted for voice conversion with slight modification of the methods and interpretations. Though the goals of accent conversion and voice conversion with respect to identity are diametrically opposed, LDA is equally useful in both problems. In voice conversion, for instance, a positive result would find the converted utterances projected closer to the target than to the source. Voice conversion does not make a distinction between accent and identity because most methods implicitly model “segmental” accent, and cross-accent conversion is rarely performed (though cross-language conversion is an active research topic [7]). However, in cases involving source/target pairs with different accents, it is reasonable to assume that a converted voice with the target accent would be preferred over one with the source accent. In such a case it may be worthwhile to measure accent as an additional component of identity. The objective measure of accentedness may be more relevant in voice conversion if interpreted as a measure of intelligibility. In the case of acoustic quality, P.563 should be as appropriate for voice conversion as it is for accent conversion.

5.2 Future work

Further research may be required to determine the extent to which these objective measures work across different pairs of source and target speakers. In terms of acoustic quality, there is no reason to believe that it would perform differently on other speakers given that P.563 is an ITU-T standard. With reference to accent, results in Figure 1(a) on the ARCTIC corpus and the fact that HTK was trained on a large population of American English speakers support the view that our accent measure can discriminate L2 from L1 speech, though further tests could be performed on datasets containing a broad range of accents, such as the Speech Accent Archive [68]. Finally, our identity measure relies on a sound statistical method (LDA) to identify directions of maximum discrimination among pairs of speakers, on a pair-by-pair basis. Thus, the LDA scores can be expected to adapt to different pairs of source and target speakers.

6 ACKNOWLEDGMENTS

Hart Blanton is acknowledged for suggestions regarding the EGWA scale and for making his laboratory available for perceptual tests. This work was supported by NSF award 0713205.

7 REFERENCES

- [1] C. Watson and D. Kewley-Port, "Advances in computer-based speech training: Aids for the profoundly hearing impaired," *Volta-Review*, vol. 91, pp. 29-45, 1989.
- [2] M. Jilka and G. Möhler, "Intonational Foreign Accent: Speech Technology and Foreign Language Teaching," *ESCA Workshop on Speech Tech. Lang. Learn.*, pp. 115-118, 1998.
- [3] U. Gut, "Foreign Accent," in *Speaker Classification I: Fundamentals, Features, and Methods*, 2007, pp. 75-87.
- [4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP*, New York, NY 1988, pp. 655-658.
- [5] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, vol. 8, pp. 147-158, 1989.
- [6] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, 1998, pp. 285-288.
- [7] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *ASRU*, 2003, pp. 676-681.
- [8] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Computer Speech & Language*, vol. 20, pp. 441-467, 2006.
- [9] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 131-142, 1998.
- [10] Q. Yan, S. Vaseghi, D. Rentzos, and C. H. Ho, "Analysis and Synthesis of Formant Spaces of

- British, Australian, and American Accents," *IEEE Trans Audio, Speech, and Lang. Proc.*, vol. 15, pp. 676-689, 2007.
- [11] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, in press in press.
- [12] M. Huckvale and K. Yanagisawa, "Spoken Language Conversion with Accent Morphing," in *Proc. ISCA Speech Synth. Workshop*, 2007, pp. 64-70.
- [13] R. C. Major, *Foreign accent: the ontogeny and phylogeny of second language phonology*: Erlbaum, 2001.
- [14] A. Ikeno and J. H. L. Hansen, "The effect of listener accent background on accent perception and comprehension," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 1-8, 2007.
- [15] A. Neri, C. Cucchiaroni, and H. Strik, "Feedback In Computer Assisted Pronunciation Training: Technology Push Or Demand Pull? ," in *CALL Conference*, 2002, pp. 179-188.
- [16] T. K. Hansen, "Computer Assisted Pronunciation Training: The four 'K's of feedback," in *Fourth Intl. Conf. on Multimedia Information Communication Technologies in Education*, 2006, pp. 342-346.
- [17] T. van Els and K. de Bot, "The role of intonation in foreign accent," *Modern Language Journal*, vol. 71, pp. 147-155, 1987.
- [18] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *JASA*, vol. 102, pp. 28-40, 1997.
- [19] M. Munro, "Non-segmental factors in foreign accent: ratings of filtered speech," *Studies in Second Language Acquisition*, vol. 17, pp. 17-34, 1995.
- [20] 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.
- [21] K. Tajima, R. Port, and J. Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," *Journal of Phonetics*, vol. 25, pp. 1-24, 1997.
- [22] K. Cho and J. G. Harris, "Towards an Automatic Foreign Accent Reduction Tool," in *Proc. 3rd Intl. Conf. on Speech Prosody*, 2006.
- [23] 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.
- [24] K. Nagano and K. Ozawa, "English Speech Training Using Voice Conversion," in *ICSLP*, 1990, pp. 1169-1172.
- [25] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical Stress Training of German Compounds for Italian Speakers by means of Resynthesis and Emphasis," in *Proc 11th Australian Intl Conf Speech Science & Technology*, 2006, pp. 24-29.
- [26] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis by synthesis of acoustic correlates of British, Australian and American accents," in *ICASSP*, 2004, pp. 637-640.
- [27] T. Kamiyama, "Perception of Foreign Accentedness in L2 Prosody and Segments: L1 Japanese Speakers Learning L2 French," in *Speech Prosody: ISCA*, 2004.
- [28] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. v. d. Vreken, "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes," in *ICSLP*. vol. 3, 1996, pp. 1393-1396.
- [29] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.5.15)," Universiteit van Amsterdam, Institute of Phonetics Sciences, 2007.
- [30] M. Munro and T. Derwing, "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," *Language Learning & Technology*, vol. 45, pp. 73-97, 1995.

- [31] L. Jin and R. Kubichek, "Output-based objective speech quality," in *IEEE Vehicular Technology Conference*, 1994, pp. 1719-172.
- [32] L. Malfait, J. Berger, and M. Kastner, "P.563-The ITU-T Standard for Single-Ended Speech Quality Assessment," *IEEE Trans Audio, Speech and Lang. Proc.*, vol. 14, pp. 1924-1934, 2006.
- [33] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEEE Proc. Vision, Image, and Signal Processing*, vol. 147, pp. 493-501, 2000.
- [34] K. Doh-Suk and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," in *ICASSP*, 2004, pp. 1060-3.
- [35] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, pp. 171-185, 1995.
- [36] C. Huang, T. Chen, and E. Chang, "Accent Issues in Large Vocabulary Continuous Speech Recognition," *Intl. Journal of Speech Technology*, vol. 7, pp. 141-153, 2004.
- [37] M. Huckvale, "ACCDIST: A Metric for Comparing Speakers' Accents," in *ICSLP*, 2004.
- [38] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, 2005, pp. 139-143.
- [39] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *ASRU*, 2001.
- [40] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, pp. 353-367, 1996.
- [41] Q. Yan and S. Vaseghi, "A comparative analysis of UK and US English accents in recognition and synthesis," in *ICASSP*, 2002.
- [42] W. Barry, C. Hoequist, and F. Nolan, "An approach to the problem of regional accent in

- automatic speech recognition," *Computer Speech and Language*, vol. 3, pp. 355-366, 1989.
- [43] N. Minematsu and S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," in *Proc. Integration of Speech Technology into Learning*, 2000, pp. 130-137.
- [44] H. Kuwabara and T. Takagi, "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method," *Speech Communication*, vol. 10, pp. 491-495, 1991.
- [45] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," *IEEE Trans Audio Electroacoustics*, vol. 21, pp. 428-436, 1973.
- [46] N. Malayath, H. Hermansky, and A. Kain, "Towards decomposing the sources of variability in speech," in *Eurospeech*, 1997, pp. 497-500.
- [47] Y. Lavner, I. Gath, and J. Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," *Speech Communication*, vol. 30, pp. 9-26, 2000.
- [48] F. Bimbot, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430-451, 2004.
- [49] L. M. Arslan, "Speaker Transformation Algorithm using Segmental Codebooks (STASC)," *Speech Communication*, vol. 28, pp. 211-226, 1999.
- [50] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: a modulation theory," *Phonetica*, vol. 51, pp. 170-183, 1994.
- [51] G. Fant, *Acoustic theory of speech production*. s'Gravenhage: Mouton, 1960.
- [52] S. J. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," Department of Engineering, Cambridge University 1993.

- [53] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 23, pp. 176-182, 1975.
- [54] B. Vieru-Dimulescu and P. B. d. Mareuil, "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.
- [55] D. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, pp. 786-794, 1981.
- [56] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236-243, 1984.
- [57] M. Munro and T. Derwing, "Evaluations of foreign accent in extemporaneous and read material," *Language Testing*, vol. 11, pp. 253-266, 1994.
- [58] B. Pelham and H. Blanton, *Conducting Research in Psychology, Measuring the Weight of Smoke*, 3rd ed. Belmont, CA: Thomson Higher Education, 2007.
- [59] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, pp. 265-275, 1991.
- [60] S. M. Sheffert, D. B. Pisoni, J. M. Fellowes, and R. E. Remez, "Learning to recognize talkers from natural, sinewave, and reversed speech samples," *J Exp Psychol Hum Percept Perform*, vol. 28, pp. 1447-69, 2002.
- [61] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments," University of Cambridge, United Kingdom 2006.
- [62] R. Weide, "The CMU pronunciation dictionary, release 0.6," Carnegie Mellon University, 1998.
- [63] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University Language Technologies Institute 2003.

- [64] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.
- [65] O. P. Kenny, D. J. Nelson, J. S. Bodenschatz, and H. A. McMonagle, "Separation of non-spontaneous and spontaneous speech," in *ICASSP*, 1998, pp. 573-576 vol.1.
- [66] 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.
- [67] O. Turk and L. M. Arslan, "Donor Selection for Voice Conversion," in *EUSIPCO*, 2005.
- [68] S. H. Weinberger, "Speech Accent Archive," in *Online at* < <http://accent.gmu.edu/index.php> > (*accessed May 10, 2009*), 2008.