# Developing Objective Measures of Foreign-Accent Conversion

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

***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 use phoneme-level match scores given by the HTK recognizer trained on a large number of English American speakers to obtain a measure of native accent. Finally, we propose a measure of speaker identity that projects acoustic vectors (e.g., Mel cepstral, F0) onto the linear discriminant that maximizes separability for a given pair of source and target speakers. 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.**

***Index Terms*—Accent conversion, foreign accent recognition, speaker recognition, voice conversion.**

## I. 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,[1] learners can more easily perceive differences between their accented utterances and their ideal accent-free counterparts. Accent usually manifests itself through a process known as phonological transfer [3]—L2 learners systematically substitute, delete, or insert phonemes in their second language as predicted by the phonological rules of their first language. Our method of accent conversion adopts techniques from voice conversion [4], [5], and is best suited to address the issue of phoneme substitution rather than insertion or deletion. We find that for speakers with a moderate accent, prosody modification and segmental substitution are sufficient to make their utterances sound more native.

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 [3] 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 [6]. 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 accent, 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

Manuscript received May 10, 2009; revised November 13, 2009. Current version published June 16, 2010. This work was supported by the National Science Foundation under Award 0713205. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Olivier Rosec.

The authors are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: dlfelps@cse.tamu.edu; rgutier@cse.tamu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2038818

[1]In this context, "voice quality" refers to the characteristics of a voice that make it unique. The contributing factors may be anatomical (e.g., length of the vocal tract or size of the larynx) or idiosyncratic (e.g., how hoarse, breathy, rough, or loud a person speaks). Acoustic quality, on the other hand, is a measure of goodness inherent in the speech signal; it is marked by the background noise level and any audible distortions.

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 [7], [8]. The work proposed here is related to, but distinct from, prior research in automatic accent classification and speaker identification/verification. The goal of accent classification is to assign a given speech recording to one of several accent/dialect categories. In contrast, our goal is to assign a foreign-accent score that correlates with ratings of perceived accent, which are continuous rather than discrete. Thus, our problem is one of regression rather than classification. The goal of speaker verification/identification 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).

## II. Related Work

### A. Foreign Accent Conversion

The relative contribution of various acoustic cues to the perception of a foreign accent has been extensively studied [3], [9]–[12]. 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 [9] and also by the availability of methods for time- and pitch-scaling [13]. Tajima *et al.* [14] 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 [15] 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 [16]. 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-accent score from 1.24 to 2.10 (on a seven-point scale), whereas resynthesizing Korean utterances with an American intonation reduced their foreign-accent score from 5.08 to 4.75. Prosodic conversion has also been explored in the context of CAPT. Nagano and Ozawa [17] 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 showed 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.* [18] investigated the use of prosodic modification to teach German prosody to Italian speakers. Their results were consistent with [17], 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.* [19] proposed an accent-synthesis method based on formant warping. First, the authors developed a formant tracker based on hidden Markov models (HMMs) and linear predictive coding (LPC), and applied it to a corpus containing several regional English accents (British, Australian, and American). Second, the authors resynthesized 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 resynthesized utterances were perceived as having the native accent. Kamiyama [20] 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 [21] and PRAAT [22] 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 accent, though prosody plays a more significant role. More recently, Huckvale and Yanagisawa [23] used an English text-to-speech (TTS) system to simulate English-accented Japanese utterances; a foreign accent 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 [24]. Our work differs from [19] in two respects. First, our accent conversion method (described in Section III-A) uses a spectral envelope vocoder, which makes it more suitable than formant tracking for unvoiced segments. Second, we evaluate not only the accent of the resynthesized 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

[23], our study was performed on natural speech, and focused on accent and identity rather than on intelligibility; as noted by Munro and Derwing [25], a strong foreign accent does not necessarily limit the intelligibility of the speaker.

### B. Acoustic Correlates of Acoustic Quality, Foreign Accent, and Speaker Identity

*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 [26], voice over IP, and cell phone networks [27]. The most common approach is to create a model of clean speech (e.g., with vector quantization [26]) 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 [28] and the human auditory system [29] have also been proposed. However, the prevailing non-intrusive measure is ITU-T recommendation P.563 [27], which is discussed in Section III-C.

*2) Foreign Accent:* Speaker adaptation is a prevalent topic in speech recognition research as it helps decrease speaker variability due to differences in gender, physiology, or accent [30]. Such investigations have also led to objective measures of accent [31]. These can be grouped into three categories [32]: 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 [33]. Classification is then achieved through pattern recognition, e.g., Gaussian mixture models (GMMs) [34], [35].

Accent-specific phone models have been explored by Arslan and Hansen [36]. 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 [37].

The final group takes a linguistic approach to accent classification, one which may be more sensitive than methods based on acoustic quality [32]. In one of the earlier papers, describing accent classification for speech recognition, Barry *et al.* [38] 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 [39].

*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 [40] indicate a stronger relationship between identity and spectral measures than between identity and pitch, other studies show a preference for pitch [41]. Malayath *et al.* [42] 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.* [43] 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 [44]. 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 [45].

## III. Methods

### A. Foreign Accent Conversion

According to the modulation theory of speech [46], 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 [47], 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 [24]. 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 [48]. From these phonetic segments, the ratio of source-to-target durations is used to specify a time-scaling factor $\alpha$ 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 $\beta$ 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 [49], while acquiring the pitch dynamics of the target speaker, which provides important cues to native accent [50]. Once the time- and pitch-scale modification parameters $(\alpha, \beta)$ are calculated, Fourier-domain PSOLA [16] 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 [51] 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. In order to reduce speaker-dependent information in the target's spectral envelope, we also perform vocal tract length normalization (VTLN) on the target's spectral envelop prior to multiplication with the source's flat excitation spectrum. Following [52], VTLN is performed by warping the target's spectral envelope according to a piecewise linear function defined by the average formant pairs of the two speakers; formant locations are estimated with PRAAT [22] over the entire corpus. The modified short-time spectra are transformed back to the time domain, and concatenated using a least-squared-error criterion [53]. 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.

TABLE I
STIMULUS CONDITIONS FOR THE 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 II
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 |

### B. Perceptual Evaluation

In [24], we performed a series of perceptual experiments to characterize the method in terms of 1) degradations in acoustic quality, 2) the degree of reduction in foreign accent, and 3) the extent to which the identity of the original speaker had been preserved. 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 I). 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 [54], 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 accent*. Following [55], participants were asked to rate the degree of foreign accent of utterances using a seven-point Empirically Grounded, Well-Anchored (EGWA) scale ($0 =$ not at all accented; $2 =$ slightly accented; $4 =$ quite a bit accented; $6 =$ extremely accented) [56].
- *Speaker identity*. Following [57], participants listened to a pair of linguistically different utterances, and were asked to 1) determine if the two sentences were produced by the same speaker, and 2) rate their confidence on a seven-point EGWA scale. These two responses were converted into a 15-point perceptual score (Table II). 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 [58].

### C. Objective Measures

*1) Acoustic Quality:* Our objective measure of acoustic quality is based on recommendation P.563 for single-ended (i.e., no reference) speech quality [27], 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, 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 100-Hz 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 [27]. 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 II-B1); 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 [27]; 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 20 sentences are averaged per condition.

*2) Foreign Accent:* The objective measure of accent adopted in this study is related to that in [36]. 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) [48]. We use acoustic models trained on 284 North American speakers [59] coupled with the CMU pronunciation dictionary [60]; the models contained approximately 7400 tied-state triphones, 16 Gaussian components per state (32 for silence), and were trained using the MFCC_0_D_A_Z acoustic features (i.e., 12 cepstra plus the 0th cepstra, delta and delta-delta, normalized using cepstral mean subtraction). 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. We call the HTK's general-purpose word recognizer "HVite" with flags to specify forced alignment (-a) and to output the calculated log likelihoods (-o); details of the procedure may be found in the HTK book [61], section 13.5. The objective score for an utterance is computed as the median value $\mu_{1/2}(\cdot)$ of the phoneme-level match scores (i.e., log likelihood) returned by HTK in the label file, excluding those associated with silences

$$\text{Accent} = -\mu_{1/2}\left\{logp(s_j|\lambda_i); s_j \leftarrow i; i = 1 \ldots n\right\}$$

where $s_j$ is the $j$th segment in the utterance (obtained by the force-alignment procedure), $i$ is the phoneme label to which segment $s_j$ is assigned, $logp(s_j|\lambda_i)$ is the log-likelihood of segment $s_j$ (i.e., the score returned by HVite), and $n$ is the number of phonemes ($n = 39$, not including silence).

To test the effectiveness of this measure against multiple dialects and accents of English we selected 108 speakers from the IDEA database [62]. Each of these speakers had read the "*Comma gets a cure*" passage and belonged to a country[2] with at least four such speakers, for a total of 13 countries. We employed spectral subtraction [63] to reduce background noise levels, which can vary significantly from speaker to speaker in the IDEA corpus. The resulting accent scores, summarized in Fig. 1(a), show a separation between countries that speak English as a first language and those that do not. A two-tailed $t$-test[3] found the means of the two groups to be significantly different; $t(107) = 7.16$, $p < 0.001$. Given that the HMM's acoustic models were trained on native speakers of American English, it is surprising that Australia outscored General American (genam), though the difference was not significant $t(22) = 0.16$, $p = 0.87$. In fact, the first country with a significantly different score from Australia was India; $t(20) = 2.22$, $p < 0.05$, the best scoring country where English is not the official language (English is considered a subsidiary official language in India).

[2]Since there may be multiple dialects within a single country, we selected speakers with the same dialect when possible. Namely, we chose Received Pronunciation (RP) from the U.K. and General American (genam) from the U.S.

[3]The $t$-test determines if the means between two groups are significantly different. Results are reported as $t(df) = \_, p < x$, where the $t$-score depends on the degrees of freedom ($df$). The difference is deemed significant if $p < 0.05$ (i.e., there is less than a 5% chance that the groups were drawn from the same distribution).

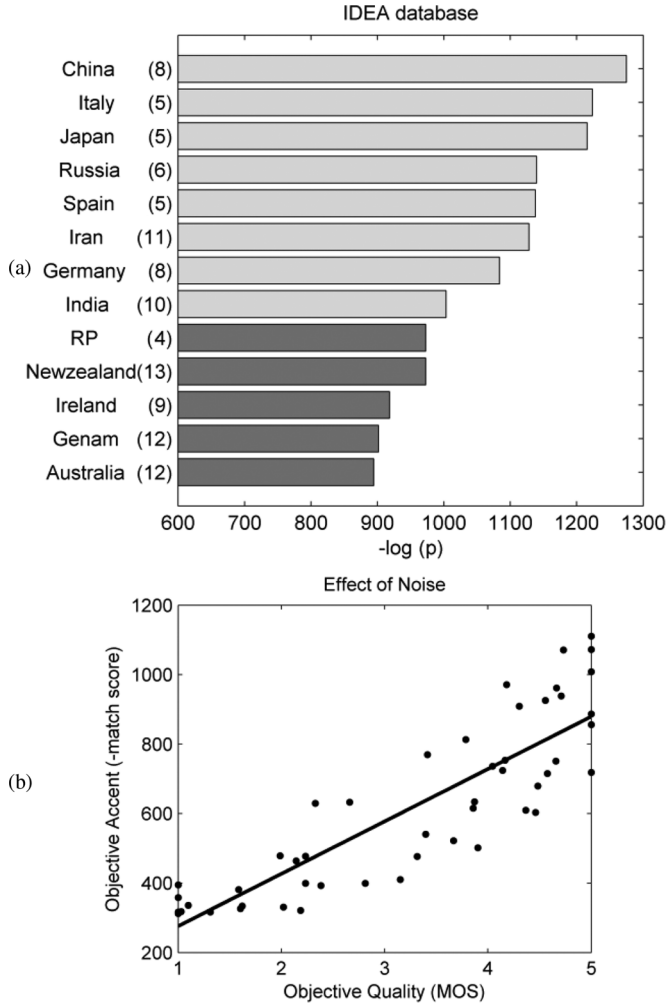


Fig. 1. (a) Average accent scores for 13 dialects and accents in the IDEA database. Dark colored bars represent countries that speak English as a first language. The number of speakers per country is given in parenthesis. (b) Effect of additive Gaussian noise on the HTK score. Ten sentences with ten levels of noise spanning the full range of MOS values were used for this purpose.

To determine the effect of residual noise on the likelihood scores, we selected seven speakers (with clean recordings) from the CMU ARCTIC dataset [64]. Ten sentences from each speaker were measured with ten levels of additive white noise spanning the range of objective quality measures. Results from one of the speakers (ksp_indianmale) are shown in Fig. 1(b); a strong correlation[4] between our objective measures of accent and quality indicates that this effect is significant; $r(98) = -0.87$, $p < 0.001$. We adjust all accent scores by the average trend of the seven speakers (120 accent points per quality point) to obtain a measure of accent that was (linearly) independent from acoustic quality.

*3) Speaker Identity:* Our objective measure of speaker identity is based on a signal discrimination criterion [65]. 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, which are generally 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.* [43], 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 [65], consider the problem of discriminating between two classes on the basis of a D-dimensional feature vector $x$. LDA seeks a linear projection vector $v$ such that the projected data $y = v^T x$ 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

$$v = S_W^{-1}(\mu_1 - \mu_2)$$

where $\mu_1$, $\mu_2$ are the sample mean of the two classes, and $S_W$ is the within-class scatter

$$S_W = \sum_{i=1}^{2} \sum_{n \in C_i} (x_n - \mu_i)(x_n - \mu_i)^T.$$

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 IV-B), and do not include any accent conversions. Once the Fisher's LDA solution $v$ has been computed from training data, each new test sentence is framed, each frame is analyzed to obtain acoustic vector $x$, and then projected onto the LDA solution. Each test sentence is then assigned an identity score $ID$ that corresponds to the average LDA projection of its frames

$$ID = E_x[v^T x].$$

Note that this identity score must be interpreted in relation to those of the reference speakers: if the identity score for a test utterance is closer to the score of source utterances $ID_S$ than to the score of target utterances $ID_T$, then it is evidence that the test utterance sounds more like the source than the target.

## IV. Results

We validated the proposed objective measures against results from a perceptual evaluation of our accent conversion model

[4]The Pearson product-moment correlation tests the linear dependence between two variables. Results are shown as $r(df) = \_, p < x$, where the magnitude and sign of $r$ indicate the strength and direction of the relationship; $r$ ranges from [−1,1].

previously reported in [24]. In what follows, we use a two-way analysis of variance[5] (ANOVA) to calculate statistical significance, with the two factors being the prosodic and segmental transformations.

### A. Acoustic Quality

Forty-three students participated in a 25-min test to rate the perceived quality of recorded/synthesized utterances. Results are summarized in Fig. 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, $F(1,76) = 57.31$ $p < 0.001$.

The objective measure also captured all 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. Some of these distortions can be traced back to errors in the original alignment in ARCTIC and to voicing mismatches between source and target. PSOLA can also introduce distortions; while we do impose upper and lower limits for the pitch- and time-scaling factors, these limits are currently defined globally as opposed to on a phone-by-phone basis (e.g., plosives should have lower time-stretching limits than vowels). 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.

### B. Foreign Accent

Thirty-nine students participated in a 25-min test to establish the degree of foreign accent of individual utterances. Results are summarized in Fig. 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], [17], [18]. 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 [64], since read speech is prosodically flat when compared to spontaneous or conversational speech [66]. The objective measure showed identical trends; the main effect of the segmental transformation was significant, $F(1,76) = 4.48$, $p < 0.05$, and the other effects were not.

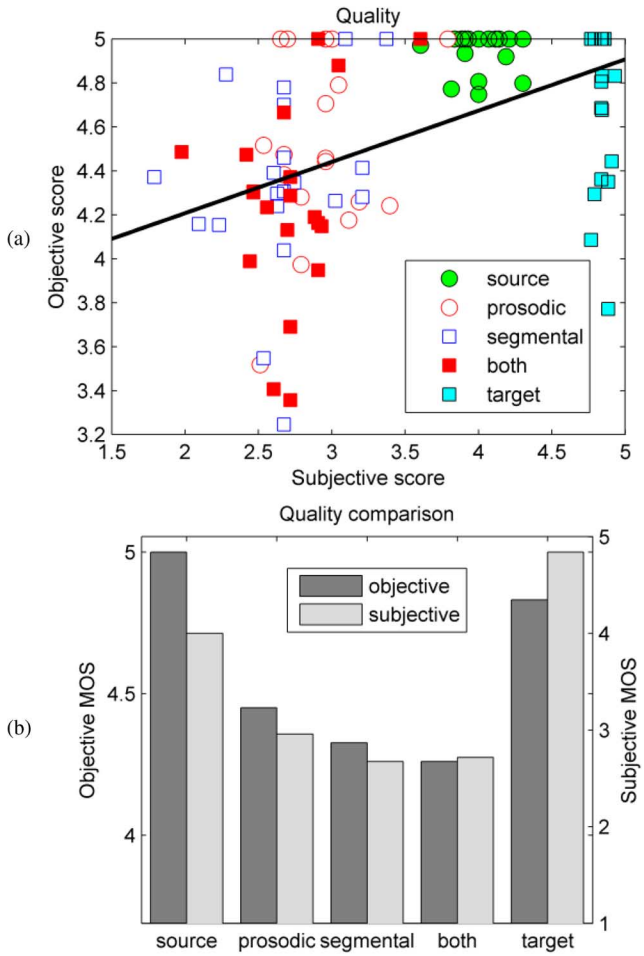


Fig. 2. (a) Correlation between objective and subjective measures of acoustic quality; each sample in the scatter plot represents an utterance. (b) Average scores for the two measures across experimental conditions. The objective measure follows a similar trend as the subjective scores.

### C. Speaker Identity

Forty-three students participated in a 25-min 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 [67], an MDS technique that attempts to preserve the geodesic distance between stimuli. ISOMAP visualizations of the identity tests, shown in Fig. 4, reveal two distinct clusters, one containing utterances from the target (native) speaker and a

[5]A two-factor ANOVA allows us to test the significance of the two manipulations (i.e., prosodic and segmental transformations) in the four conditions containing source excitation: source utterance [0,0], prosodic transformation [1,0], segmental transformation [0,1], and both transformations [1,1]. Results are reported as $F(df_A, df_{\text{error}}) = \_, p < x$, where the F-score for the independent variable A depends on its degrees of freedom ($df_A$) and those of the error ($df_{\text{error}}$). A complete analysis includes an F-score for each independent variable as well for all possible interactions.

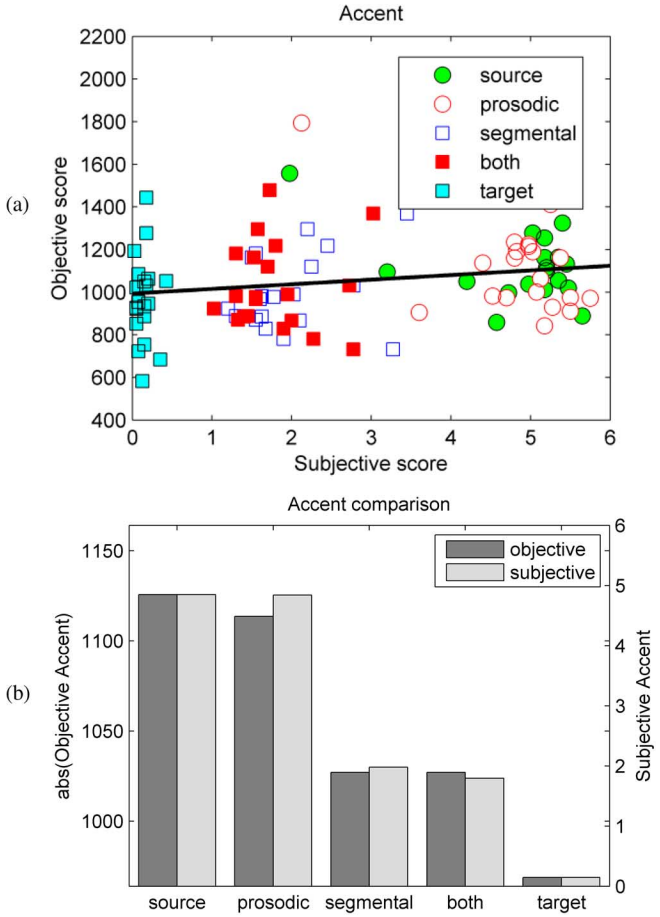


Fig. 3. (a) Correlation between objective and subjective measures of accent. (b) Average scores for the two measures across experimental conditions. The objective measure follows the same trend as the subjective scores. Left and right $y$-axes were aligned to facilitate the comparison.

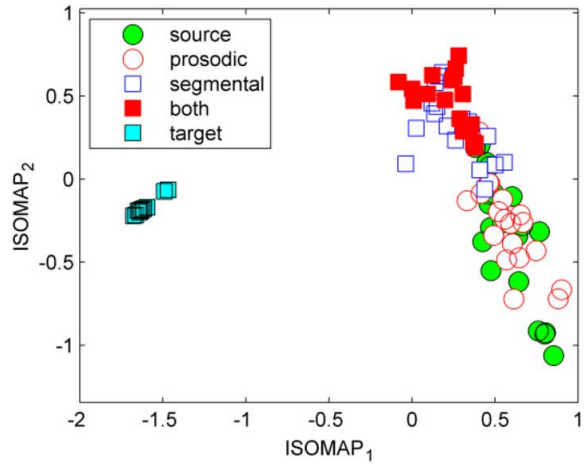


Fig. 4. Experimental results from the identity tests; ISOMAP reveals only two clusters: one for the source, and a second one for all other utterances.

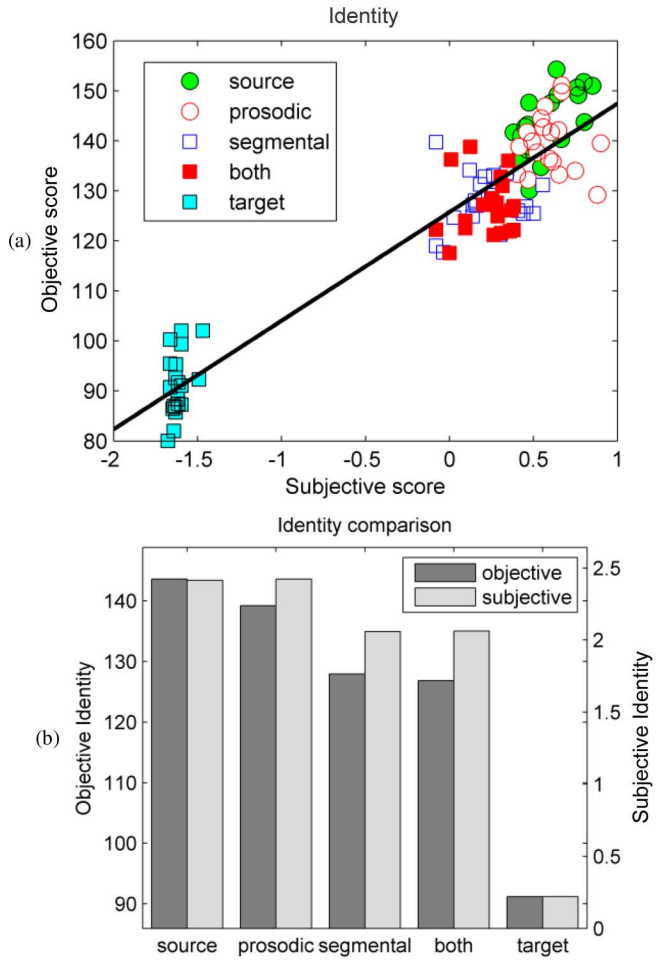


Fig. 5. (a) Correlation between objective and subjective measures of identity. (b) Average scores for the two measures across experimental conditions. (b) Experimental results from the identity tests. The objective measure follows the same trend as the subjective scores.

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). Fig. 5(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.

## V. Discussion

We have proposed objective measures that can be used to assess the acoustic quality, degree of foreign accent 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 Fig. 2(b) is $r(3) = 0.80$, $p = 0.11$; this coefficient increases to $r(3) = 0.997$, $p < 0.001$

for the accent measures shown in Fig. 3(b) and to $r(3) = 0.99$, $p < 0.005$ for the identity measures shown in Fig. 5(b). When computed across sentences, these correlation coefficients become $r(98) = 0.47$, $p < 0.001$ for the acoustic quality scores in Fig. 2(a), $r(98) = 0.21$, $p < 0.05$ for the accent scores in Fig. 3(a), and $r(98) = 0.94$, $p < 0.001$ for the identity scores in Fig. 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 Fig. 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 seven-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 Fig. 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 underpenalizes utterances resynthesized with our accent conversion model since P.563 focuses on degradations in narrowband 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 $ID = 128$ (arbitrary units) averaged across 20 utterances. This value can only be interpreted when compared against the scores for the source ($ID_S = 143$) and target ($ID_T = 91$) 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, 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 $d = e^{-(ID-ID_S)^2/(ID_S-ID_T)^2}$ may be used to transform the LDA projection into a measure that denotes how close an utterance is to the source. Alternatively, a posterior probability may be computed for identity scores as

$$p(S|ID) = \frac{p(ID|S)p(S)}{p(ID)} = \frac{1}{2K\sqrt{2\pi}\sigma_S} exp\left(\frac{-(ID-ID_S)^2}{2\sigma_S^2}\right)$$

where $\sigma_S$ is the standard deviation of identity scores for source utterances in the training set, $p(S)$ is the prior probability of source utterances (assumed equal to 1/2), and $p(ID) = K$ is a normalization constant to ensure that $p(S|ID)+p(T|ID) = 1$.

### A. Potential Applications

Development time 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 avoiding 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., [68] 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 [7], [8]. 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 [52]). 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 accent 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.

### B. 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, incorporating *a priori* information about a foreign speaker's first language (e.g., identifying minimal pairs) may lead to more sensitive measures of foreign accent, e.g., by placing higher emphasis on HTK scores of specific phonemes or words [69]. 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.

## ACKNOWLEDGMENT

The authors would like to thank Hart Blanton for suggestions regarding the EGWA scale and for making his laboratory available for perceptual tests.

## REFERENCES

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

[2] M. Jilka and G. Möhler, "Intonational foreign accent: Speech technology and foreign language teaching," in *Proc. ESCA Workshop Speech Tech. Lang. Learn.*, 1998, pp. 115–118.

[3] R. C. Major, *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. Mahwah, NJ: Erlbaum, 2001.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[5] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. ICASSP 2001*, Salt Lake City, UT, 2001, pp. 813–816.

[6] A. Ikeno and J. H. L. Hansen, "The effect of listener accent background on accent perception and comprehension," *EURASIP J. Audio, Speech, Music Process.*, vol. 2007, pp. 1–8, 2007.

[7] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: Technology push or demand pull?," in *Proc. CALL Conf.*, 2002, pp. 179–188.

[8] T. K. Hansen, "Computer assisted pronunciation training: The four 'K's of feedback," in *Proc. 4rth Int. Conf. Multimedia Inf. Technol. in Education*, 2006, pp. 342–346.

[9] T. van Els and K. de Bot, "The role of intonation in foreign accent," *Modern Lang. J.*, vol. 71, pp. 147–155, 1987.

[10] U. Gut, "Foreign accent," in *Speaker Classification I: Fundamentals, Features, and Methods*, J. G. Carbonell and J. Siekmann, Eds. New York: Springer, 2007, pp. 75–87.

[11] 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.

[12] M. Munro, "Non-segmental factors in foreign accent: Ratings of filtered speech," *Studies in Second Lang. Acquisition*, vol. 17, pp. 17–34, 1995.

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

[14] K. Tajima, R. Port, and J. Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," *J. Phon.*, vol. 25, pp. 1–24, 1997.

[15] K. Cho and J. G. Harris, "Towards an automatic foreign accent reduction tool," in *Proc. 3rd Int. Conf. Speech Prosody*, 2006.

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

[17] K. Nagano and K. Ozawa, "English speech training using voice conversion," in *Proc. ICSLP*, 1990, pp. 1169–1172.

[18] 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 Int. Conf. Speech Sci. Technol.*, 2006, pp. 24–29.

[19] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis by synthesis of acoustic correlates of British, Australian and American accents," in *Proc. ICASSP*, 2004, pp. 637–640.

[20] T. Kamiyama, "Perception of foreign accentedness in L2 prosody and segments: L1 Japanese speakers learning L2 French," in *Proc. Speech Prosody: ISCA*, 2004.

[21] 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 *Proc. ICSLP*, 1996, vol. 3, pp. 1393–1396.

[22] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (Version 4.5.15) Institute of Phonetics Sciences, Universiteit van Amsterdam, 2007.

[23] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in *Proc. ISCA Speech Synth. Workshop*, 2007, pp. 64–70.

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

[25] M. Munro and T. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang. Learn. Technol.*, vol. 45, pp. 73–97, 1995.

[26] L. Jin and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Veh. Technol. Conf.*, 1994, pp. 1719–1723.

[27] L. Malfait, J. Berger, and M. Kastner, "P.563-the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[28] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEE Proc. Vis, Image, Signal Process.*, vol. 147, pp. 493–501, 2000.

[29] K. Doh-Suk and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," in *Proc. ICASSP*, 2004, pp. 1060–1063.

[30] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[31] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *Int. J. Speech Technol.*, vol. 7, pp. 141–153, 2004.

[32] M. Huckvale, "ACCDIST: A metric for comparing speakers' accents," in *Proc. ICSLP*, 2004.

[33] 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, Lang. Process.*, vol. 15, no. 2, pp. 676–689, Feb. 2007.

[34] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Proc. 4th IEEE Workshop Autom. Identification Adv. Technol.*, 2005, pp. 139–143.

[35] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *Proc. ASRU*, 2001.

[36] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American english," *Speech Commun.*, vol. 18, pp. 353–367, 1996.
[37] Q. Yan and S. Vaseghi, "A comparative analysis of UK and US English accents in recognition and synthesis," in *Proc. ICASSP*, 2002, pp. 413–416.
[38] W. Barry, C. Hoequist, and F. Nolan, "An approach to the problem of regional accent in automatic speech recognition," *Comput. Speech, Lang.*, vol. 3, pp. 355–366, 1989.
[39] N. Minematsu and S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," in *Proc. Integration Speech Technol. Into Learn.*, 2000, pp. 130–137.
[40] H. Kuwabara and T. Takagi, "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method," *Speech Commun.*, vol. 10, pp. 491–495, 1991.
[41] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," *IEEE Trans. Audio Electroacoust.*, vol. AE-21, no. 5, pp. 428–436, Oct. 1973.
[42] N. Malayath, H. Hermansky, and A. Kain, "Towards decomposing the sources of variability in speech," in *Proc. Eurospeech*, 1997, pp. 497–500.
[43] Y. Lavner, I. Gath, and J. Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," *Speech Commun.*, vol. 30, pp. 9–26, 2000.
[44] F. Bimbot, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, 2004.
[45] L. M. Arslan, "Speaker Transformation Algorithm using Segmental Codebooks (STASC)," *Speech Commun.*, vol. 28, pp. 211–226, 1999.
[46] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, pp. 170–183, 1994.
[47] G. Fant, *Acoustic Theory of Speech Production*. s'Gravenhage, The Netherlands: Mouton, 1960.
[48] S. J. Young, *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. Cambridge, U.K.: Dept. of Eng., Cambridge Univ., 1993.
[49] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 2, pp. 176–182, Apr. 1975.
[50] B. Vieru-Dimulescu and P. B. d. Mareüil, "Contribution of prosody to the perception of a foreign accent: A study based on Spanish/Italian modified speech," in *Proc. ISCA Workshop Plasticity in Speech Perception*, London, U.K., 2005, pp. 66–68.
[51] D. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, pp. 786–794, 1981.
[52] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *Proc. ASRU*, 2003, pp. 676–681.
[53] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.
[54] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.
[55] M. Munro and T. Derwing, "Evaluations of foreign accent in extemporaneous and read material," *Lang. Testing*, vol. 11, pp. 253–266, 1994.
[56] B. Pelham and H. Blanton, *Conducting Research in Psychology, Measuring the Weight of Smoke*, 3rd ed. Belmont, CA: Thomson Higher Education, 2007.
[57] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Commun.*, vol. 10, pp. 265–275, 1991.
[58] 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. 1447–1469, 2002.
[59] K. Vertanen, *Baseline WSJ Acoustic Models for HTK and Sphinx: Training Recipes and Recognition Experiments*. Cambridge, U.K.: Univ. of Cambridge, 2006.
[60] R. Weide, *The CMU Pronunciation Dictionary, Release 0.6*. Pittsburgh, PA: Carnegie Mellon Univ., 1998.
[61] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Cambridge Univ., 1995, vol. 1996.
[62] P. Meier and S. Muller, "IDEA: International dialects of English archive," [Online]. Available: http://web.ku.edu/~idea/index.htm 2009
[63] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[64] J. Kominek and A. Black, *CMU ARCTIC Databases for Speech Synthesis*. Pittsburgh, PA: Carnegie Mellon Univ. Lang. Technol. Inst., 2003.
[65] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[66] O. P. Kenny, D. J. Nelson, J. S. Bodenschatz, and H. A. McMonagle, "Separation of non-spontaneous and spontaneous speech," in *Proc. ICASSP*, 1998, vol. 1, pp. 573–576.
[67] 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.
[68] O. Turk and L. M. Arslan, "Donor selection for voice conversion," in *Proc. EUSIPCO*, 2005.
[69] A. Harrison, W. Lau, H. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2787–2790.

**Daniel Felps** received the B.S. degree in computer engineering (honors) from Texas A&M University (TAMU), College Station, in 2005. He is currently working towards the Ph.D. degree at TAMU.

He been a Research Assistant in the Pattern Recognition and Intelligent Sensor Machines Laboratory, TAMU, since 2006. His research interests include speech processing, voice conversion, pattern recognition, and machine learning.

**Ricardo Gutierrez-Osuna** (M'00–SM'08) received the B.S. degree in electrical engineering from the Polytechnic University, Madrid, Spain, in 1992, and the M.S. and Ph.D. degrees in computer engineering from North Carolina State University, Raleigh, in 1995 and 1998, respectively.

From 1998 to 2002, he served on the faculty at Wright State University, Dayton, OH. He is currently an Associate Professor of Computer Engineering at Texas A&M University, College Station. His current research interests include voice and accent conversion, speech and face perception, wearable physiological sensors, and active sensing.