Reduction of non-native accents through statistical parametric articulatory synthesis

Sandesh Aryal and Ricardo Gutierrez-Osuna[a)]
Department of Computer Science and Engineering, Texas A&M University, College Station, Texas 77843

(Received 13 March 2014; revised 24 November 2014; accepted 1 December 2014)

This paper presents an articulatory synthesis method to transform utterances from a second language (L2) learner to appear as if they had been produced by the same speaker but with a native (L1) accent. The approach consists of building a probabilistic articulatory synthesizer (a mapping from articulators to acoustics) for the L2 speaker, then driving the model with articulatory gestures from a reference L1 speaker. To account for differences in the vocal tract of the two speakers, a Procrustes transform is used to bring their articulatory spaces into registration. In a series of listening tests, accent conversions were rated as being more intelligible and less accented than L2 utterances while preserving the voice identity of the L2 speaker. No significant effect was found between the intelligibility of accent-converted utterances and the proportion of phones outside the L2 inventory. Because the latter is a strong predictor of pronunciation variability in L2 speech, these results suggest that articulatory resynthesis can decouple those aspects of an utterance that are due to the speaker's physiology from those that are due to their linguistic gestures.
 [http://dx.doi.org/10.1121/1.4904701]

[SSN] Pages: 433–446

I. INTRODUCTION

Speakers who acquire a second language (L2) after a certain age—the so-called "critical period"—rarely acquire native-like pronunciation. Though having a non-native accent does not necessarily limit intelligibility, non-native speakers can be subjected to negative stereotypes (Rosini, 1997). Thus, by improving their pronunciation, adult L2 learners stand more to gain than mere intelligibility. A number of studies in pronunciation training (Felps *et al.*, 2009, and references therein) have shown that learners can benefit from imitating a native (L1) speaker with a similar voice as their own. However, finding such a "golden speaker" for each learner is impractical. To address this issue, previous work (Felps *et al.*, 2009) has suggested the use of speech modification methods to provide the ideal "golden speaker" for each L2 learner: their own voice, but with a native accent. The rationale is that, by stripping away information that is only related to the teacher's voice quality, accent conversion makes it easier for students to perceive differences between their accented utterances and their ideal accent-free counterparts.

Several accent-conversion methods based on acoustic modifications have reported reductions in non-native accentedness for L2 speech (Huckvale and Yanagisawa, 2007; Yan *et al.*, 2007; Felps *et al.*, 2009; Aryal *et al.*, 2013). Despite their appeal (audio recordings are easy to obtain), acoustic modification methods lack the robustness needed for pronunciation training applications. As an example, the degree of accent reduction and overall synthesis quality depends largely on the accuracy of time alignments between pairs of L1 and L2 utterances. This is problematic because L2 speech may contain a high number of phoneme deletions, insertions and substitutions, making time alignment errors all but inevitable. More importantly, linguistic content and voice quality characteristics interact in complex ways in the acoustic domain. As a result, acoustic-based accent conversion can oftentimes lead to the perception of a "third-speaker" (Felps *et al.*, 2009), one with a voice quality different from either speaker.

This paper presents an articulatory method that avoids the inherent complexity of accent conversion in the acoustic domain. The method consists of building an articulatory synthesizer of the L2 speaker, then driving it with articulatory gestures[1] from an L1 speaker. As illustrated in Fig. 1, the approach requires (1) a flexible articulatory synthesizer that can capture subtle accent-related changes in articulators and (2) an articulatory normalization method that can account for physiological differences between the two speakers. The approach builds on our prior work on data-driven articulatory synthesis (Felps *et al.*, 2012), which illustrated the limitations of unit-selection techniques when used with small articulatory corpora.[2] For this reason, the method proposed here uses the Gaussian mixture model (GMM) of Toda *et al.* (2008) to generate a forward mapping from L2 articulators to L2 acoustics. Compared to unit selection, this statistical parametric articulatory synthesizer does not require a large articulatory corpus and provides a continuous mapping from articulators to acoustics, so it can interpolate phonemes that do not exist in L2 inventory. Given the differences in vocal tract physiology between the two speakers and in articulatory measurement procedures [e.g., pellet placement in electromagnetic articulography (EMA)], driving the resulting model with L1 articulators is unlikely to produce intelligible speech. To address this issue, in our earlier study

[a)]Author to whom correspondence should be addressed. Electronic mail: rgutier@cse.tamu.edu

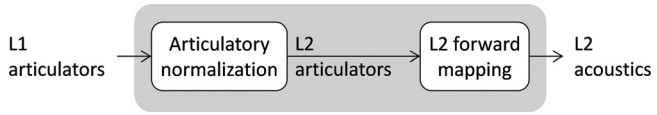

FIG. 1. Articulatory accent conversion is a two-step process consisting of L1-L2 articulatory normalization and L2 forward mapping.

(Felps *et al.*, 2012) we had mapped L1 and L2 articulators (EMA positions) into the six-point Maeda parameter approximations of Al Bawab *et al.* (2008). While this parameterization can reduce individual speaker differences, it also reduces synthesis quality because it removes important information available in the raw EMA positions. For this reason, in the present work we achieve articulatory normalization by transforming EMA articulators between the two speakers by means of a pellet-dependent Procrustes transformation derived from articulatory landmarks of the two speakers, as proposed by Geng and Mooshammer (2009).

We evaluate the proposed accent conversion method through a series of listening tests of intelligibility, non-native accentedness and speaker identity. Our results show that driving the L2 articulatory model with (normalized) articulators from an L1 speaker improves intelligibility, reduces non-native accentedness, and preserves voice identity.

The rest of this manuscript is organized as follows. Section II reviews previous research on accent conversion in both domains (acoustic and articulatory) and their limitations. Section III describes the statistical articulatory synthesizer and articulatory normalization methods of Fig. 1, whereas Sec. IV describes the acoustic-articulatory corpus and experimental protocol used in our study. Section V presents results from articulatory registration and listening tests of intelligibility, accentedness, and speaker identity. The article concludes with a discussion of results and directions for future work.

II. BACKGROUND

A. Foreign accent conversion

Accent conversion is closely related to voice conversion but seeks a more elusive goal. In voice conversion, the objective is to convert an utterance from a source speaker to sound as if it had been produced by a different (but known) target speaker (Sundermann *et al.*, 2003; Turk and Arslan, 2006). To do so, voice conversion techniques attempt to transform the two main dimensions of a speaker's voice individuality: physiological characteristics (e.g., voice quality, pitch range), and linguistic gestures (e.g., speaking style, accent, emotional state, etc.) Because the target speaker is known, evaluation of voice conversion results is relatively straightforward. In contrast, accent conversion seeks to combine the vocal tract physiology of an L2 learner with the linguistic gestures of an L1 teacher. This is a far more challenging problem because it requires separating both sources of information; it also seeks to synthesize speech for which there is no ground truth—the L2 voice with a native accent—which also makes evaluation more challenging than in the case of voice conversion.

What is a foreign accent? A foreign accent can be defined as a systematic deviation from the expected phonetic and prosodic norms of a spoken language. Some aspects of accent are acoustically realized as prosodic features such as pitch trajectory, phoneme durations, and stress patterns. In these cases, a simple prosody modification alone can significantly reduce the perceived accent of an L2 utterance. As an example, modification of vowel durations can reduce the foreign accentedness in Spanish-accented English (Sidaras *et al.*, 2009) because there is a significant difference in vowel durations between both languages. Modifying the prosody of an L2 utterance is straightforward because the target pitch and energy patterns and phoneme durations can be directly obtained from an L1 utterance of the same sentence. Once these prosodic features have been extracted, various techniques such as time-domain pitch synchronous overlap (PSOLA) (Yan *et al.*, 2007), frequency-domain PSOLA (Felps *et al.*, 2009), and speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) (Aryal *et al.*, 2013) have been found effective in modifying prosodic cues to foreign accents.

In most cases, though, prosodic modifications are not sufficient to achieve accent conversion. As an example, a few studies have shown that modification of phonetic realizations (i.e., segmental modification) is far more effective for accent reduction than prosody modification alone, both within regional accents of the same language (Yan *et al.*, 2007) and different languages (Felps *et al.*, 2009). However, segmental modifications are more involved than prosodic modification, especially in the acoustic domain (Aryal *et al.*, 2013). Because these acoustic realizations have an articulatory origin, one may argue that segmental conversion requires direct modification of the underlying articulatory gestures. In the following subsections we review prior work on accent conversion in the acoustic and articulatory domains.

1. Accent conversion in the acoustic domain

In early work, Yan *et al.* (2007) developed an accent-conversion method that exploited differences in vowel formant trajectories for three major English accents (British, Australian, and General American). The authors learned a speaker-independent cross-accent mapping of formant trajectories by building a statistical model [a two-dimensional hidden Markov model (HMM)] of vowel formant ratios from multiple speakers, and then extracting empirical rules to modify pitch patterns and vowel durations across the three regional accents. Once these 2D-HMMs and empirical rules had been learned from a corpus, the authors then adjusted the formant frequencies, pitch patterns and vowel durations of an utterance to match a target accent. In an ABX test, 78% of Australian-to-British accent conversions were perceived as having a British accent. Likewise, 71% of the British-to-American accent conversions were perceived to have an American accent. In both evaluations, changing prosody alone (pitch and duration pattern) led to noticeable

changes in perceived accent, though not as significantly as incorporating formant modifications as well. The method hinged on being able to extract formant frequencies, so it cannot be easily extended to larger corpora because formant frequencies are ill-defined for unvoiced phones and cannot be tracked reliably even in voiced segments.

A segmental modification method for accent conversion suitable for voiced and unvoiced phones was proposed by Felps *et al.* (2009). The authors split short-time spectra into a spectral envelope and flat glottal spectra. Then, they replaced the spectral envelope of an L2 utterance with a frequency-warped spectral envelope of a parallel L1 utterance and recombined it with the L2 glottal excitation; frequency warping was performed using a vocal tract length normalization function that matched the average formant frequencies of the two speakers (Sundermann *et al.*, 2003). Modification of prosodic cues (phone duration and pitch contour) was performed via PSOLA. Listening tests showed a significant reduction in accent following segmental modification: when listeners were asked to rate accentedness in a seven-point Likert scale,[3] accent-converted utterances were rated as being "somewhat" accented (1.97 numeric rating) whereas original L2 utterances were rated as being "quite a bit" accented (4.85 numeric rating). In contrast, prosodic modification did not achieve a significant reduction in accent (4.83 numeric rating). Listening tests of speaker identity with forward speech showed that segmental transformations (with or without prosodic transformation) were perceived as a third speaker, though the effect disappeared when participants were asked to discriminate reversed speech. The authors concluded that listeners used not only organic cues (voice quality) but also linguistic cues (accentedness) to discriminate speakers, which suggests that something is inevitably lost in the identity of a speaker when accent conversion is performed.

A few studies have attempted to blend L2 and L1 vocal tract spectra instead of completely replacing one with the other, as was done in (Felps *et al.*, 2009). In one such study, Huckvale and Yanagisawa (2007) reported improvements in intelligibility for Japanese utterances produced by an English text-to-speech (TTS) after blending their spectral envelope with that of an utterance of the same sentence produced by a Japanese TTS. More recently, we presented a voice morphing strategy that can be used to generate a continuum of accent transformations between a foreign speaker and a native speaker (Aryal *et al.*, 2013). The approach performs a cepstral decomposition of speech into spectral slope and spectral detail. Accent conversions are then generated by combining the spectral slope of the foreign speaker with a morph of the spectral detail of the native speaker. Spectral morphing is achieved by first representing the spectral detail through pulse density modulation and then averaging pulses in a pair-wise fashion. This morphing technique provides a tradeoff between reducing the accent and preserving the voice identity of the L2 learner, and may serve as a behavioral shaping strategy in computer assisted pronunciation training.

A limitation of both of our previous acoustic methods for accent conversion (Felps *et al.*, 2009; Aryal *et al.*, 2013) is that they require careful alignment (at the frame level) between parallel recordings from an L1 and L2 speaker. Given the common occurrence of deletion, substitution, and insertion errors in L2 speech, however, obtaining a good alignment is not always possible. As mentioned earlier, acoustic methods are also limited by the complex interaction of linguistic gestures and vocal tract physiology when looking at a spectrogram. As a result, accent conversions tend to be perceived as if they had been produced by a "third-speaker," one who is different from the original L1 and L2 speakers. Both of these issues disappear by operating in the articulatory domain. First, once an articulatory synthesizer has been built, there is no need for further alignment between L1 and L2 utterances: new accent conversions can be generated by driving the synthesizer directly from L1 articulators. Second, and more importantly, the linguistic gestures are readily available via the measured L1 articulators, whereas the voice identity is captured by the mapping from L2 articulators to L2 acoustics.

2. Accent conversion in the articulatory domain

Despite their theoretical advantages, articulatory techniques have not been widely used for accent conversion; one exception is the abovementioned study by Felps *et al.* (2012) on unit-selection synthesis. The approach consisted of identifying mispronounced diphones in an L2 utterance, and then replacing them with other L2 diphones (from an L2 corpus) with similar articulatory configurations as a reference L1 utterance. The target articulatory feature vector consisted of six Maeda parameter approximations (all but larynx height, which could not be measured with EMA), velocity for each of those parameters, normalized pitch, normalized loudness, and diphone duration. By replacing mispronounced diphones with other diphone units from the same speaker, the approach ensured that speaker identity was maintained. Unfortunately, the unit-selection synthesizer lacked the flexibility needed for accent conversion. First, the articulatory corpus contained 20 000 phones (or about 60 min of active speech) which, despite being larger than other articulatory databases [e.g., MOCHA-TIMIT (Wrench, 2000), X-Ray Microbeam (Westbury, 1994)], is considered small for unit-selection synthesis.[2] Second, the unit-selection framework does not have a mechanism to interpolate between units, so it cannot produce sounds that have not been already produced by the L2 learner. Finally, the approach requires that L2 utterances be segmented and transcribed phonetically, which makes it impractical for pronunciation training settings. Based on these findings, we decided to explore other methods for articulatory synthesis that may have the flexibility and low-data requirements needed for accent conversion; a review of these methods is provided next.

B. Articulatory synthesis

Articulatory synthesizers have had a long tradition in speech research, starting with the electrical vocal tract analogs of Stevens *et al.* (1953). These models have improved our understanding of the speech production mechanism and in recent years have also provided alternative speech

representations to improve the performance of automatic speech recognition systems (King *et al.*, 2007; Ghosh and Narayanan, 2011; Arora and Livescu, 2013).

Articulatory synthesis methods can be grouped into two broad categories, physics-based models and data-driven models. Physics-based models approximate vocal tract geometry using a stack of cylindrical tubes with different cross section areas. Speech waveforms are then generated by solving the wave propagation equation in the approximated tube model. In a classical study, Mermelstein (1973) analyzed midsagittal X-ray tracings to extract ten parameters that represented the configuration of the lips, jaw, tongue, velum, and larynx. This parameterization was then geometrically converted into a vocal tract area function and the corresponding all-pole filter model, This study showed that the midsagittal position of a few critical articulators is sufficient to generate intelligible speech, and served as the basis for the articulatory synthesizer of Rubin *et al.* (1981). The midsagittal representation of articulators was also emphasized in another classical articulatory model by Maeda (1990). The author analyzed X-ray motion pictures of the vocal tract from two speakers to extract seven articulatory parameters, and found that 88% of the variance during phonetic articulation could be explained with only four articulatory parameters (three tongue points and jaw position). These early studies cemented the use of the vocal tract midsagittal plane as an articulatory representation in speech production research. Later research addressed the issue of generating articulatory trajectories from text using principles from articulatory phonology (Browman and Goldstein, 1990), leading to the development of the Task Dynamic Model (Saltzman and Munhall, 1989), and that of speech motor skill acquisition, resulting in the DIVA (Directions Into Velocities of Articulators) model (Guenther, 1994). A concern with articulatory synthesis models is the large number of parameters that need to be specified in order to produce an utterance, and the lack of guarantees that the resulting trajectories correspond to the actual articulatory gestures of a speaker. This makes it difficult to determine whether poor synthesis results are due to the generated articulatory gestures or the underlying articulatory-to-acoustic model. To address this issue, Toutios and Maeda (2012) coupled Maeda's model with articulatory positions measured from EMA and real-time magnetic resonance imaging (rtMRI). Visual alignment between EMA pellet positions, the standard Maeda vocal tract grid, and rtMRI was performed manually; from this, two geometrical mappings were computed: (a) a mapping from EMA to standard Maeda control parameters and (b) a mapping from the standard Maeda control parameters to a set of speaker-specific vocal tract grid variables. The authors were able to synthesize "quite natural and intelligible" VCV words; a subsequent study (Toutios and Narayanan, 2013) using the same procedure reported successful synthesis of French connected speech. However, voice similarity between the original speaker and the articulatory synthesis was not assessed as part of the study.

In contrast with physics-based models, data-driven models use machine learning techniques to build a forward mapping from simultaneous recordings of articulators and acoustics (Kaburagi and Honda, 1998; Toda *et al.*, 2008; Aryal and Gutierrez-Osuna, 2013). Because these models are generally trained on individual speakers, the resulting forward model automatically captures the voice characteristics of the speaker, making them ideally suited for accent conversion. In an early study, Kaburagi and Honda (1998) used a k-nearest-neighbors method to predict acoustic observations from articulatory positions. Given a target articulatory frame, estimating its (unknown) acoustic observation consisted of finding a few closest articulatory frames in the corpus, and then computing a weighted average of their acoustic observations. The authors found that synthesis quality improved when the search for the closest articulator frames was limited within phoneme category. In an influential study, Toda *et al.* (2008) proposed a statistical parametric approach to learn the forward mapping. The approach consisted of modeling the joint distribution of articulatory-acoustic vectors with a GMM. Given a target articulatory frame, its acoustic observation was estimated from the GMM using a maximum likelihood estimate of the acoustic trajectory considering the dynamic features. This model was the basis of our prior work (Aryal and Gutierrez-Osuna, 2013), which sought to determine the effect on synthesis quality of replacing measured articulators (i.e., from EMA) with inverted articulators (i.e., predicted from acoustics).

In conclusion, data-driven articulatory synthesizers are better suited for accent conversion than their physics-based counterparts since they can effortlessly capture the unique voice characteristics of each speaker. Among data-driven models, statistical parametric models are also preferable to unit-selection approaches synthesis (Felps *et al.*, 2012) due to their interpolation capabilities and reduced data requirements. As we review next, statistical parametric models have also been found to be flexible enough for various types of speech modification.

1. Parametric statistical model in speech modification

Though we are not aware of any accent conversion studies based on statistical parametric synthesis, a few studies have illustrated the capabilities of these models for speech modification through articulatory control (Toda *et al.*, 2008; Ling *et al.*, 2009). As an example, Toda *et al.* (2008) used a GMM to estimate acoustic parameters (mel-cepstral coefficients) from articulatory parameters (seven EMA positions, pitch, and loudness). Then they manipulated EMA positions to simulate the effect of speaking with the mouth wide open. As a result of this manipulation, the authors observed a loss of high frequency components in fricatives. Similarly, Ling *et al.* (2009) showed the flexibility of an HMM-based articulatory synthesizer to modify vowels. The authors used context-dependent phoneme HMMs to model the joint dynamics of articulatory (six EMA positions) and acoustic (line spectral frequencies) parameters; acoustic outputs were modeled as a Gaussian distribution with the mean value given by a linear function of articulatory and state-specific parameters. The authors then modified vowels by manipulating articulatory parameters alone. As an example, increasing the tongue-height parameters led to a clear shift in vowel

perception from [ɛ] to [ɪ] in synthesis. Similarly, decreasing the tongue-height parameters led to a shift from [ɛ] to [æ]. Such capabilities suggest that statistical articulatory synthesis is ideally suited for the purpose of accent conversion.

C. Articulatory normalization

Accent conversion in the articulatory domain (Felps *et al.*, 2012) involves driving an articulatory synthesizer of the L2 learner with measured articulators from an L1 speaker. This requires articulatory normalization to account for anatomical differences between the two speakers. One approach is to parameterize the measured articulatory positions into a speaker-independent representation. Several such representations have been suggested in literature. As an example, Maeda (1990) proposed a set of relative measurements of the vocal tract that explain the majority of articulatory variance. In Maeda's representation, the vocal tract is represented by seven parameters: lips opening, jaw opening, lip protrusion, tongue tip height, tongue body shape, tongue dorsum position, and velum position. Al Bawab *et al.* (2008) developed a method to approximate Maeda parameters from EMA pellet positions; to remove individual differences, the method performed within-speaker z-score normalization of the approximated Maeda parameters. This normalized representation was then used for automatic speech recognition from articulatory positions derived from acoustics via analysis-by-synthesis. Hashi *et al.* (1998) proposed a normalization procedure to generate speaker-independent average articulatory postures for vowels. Using data from the X-ray microbeam corpus (Westbury, 1994), the authors scaled articulatory positions relative to a standard vocal tract, and then expressed the tongue surface relative to the palate. This procedure was able to reduce cross-speaker variance in the average vowel postures. In the context of articulatory inversion, McGowan (1994) and Mitra *et al.* (2011) have shown that the tract constriction variables (TVs) of Browman and Goldstein (1990) have fewer non-uniqueness problems compared to EMA pellet trajectories, which suggests they may be a better speaker-independent articulatory representation. As an example, Ghosh and Narayanan (2011) converted EMA articulatory positions into TVs, which were then used as the articulatory representation in a subject-independent articulatory inversion model. The authors reported inversion accuracies close to subject-dependent models, particularly for the lip aperture, tongue tip, and tongue body articulators.

A second approach to account for individual differences is to learn a cross-speaker articulatory mapping. As an example, Geng and Mooshammer (2009) used the Procrustes transform, learned from a parallel corpus containing articulatory trajectories of multiple speakers during vowel production. The objective of the study was to unveil speaker-independent strategies for vowel production by removing speaker-specific variations. The authors reported a 30% improvement in subject-independent articulatory classification of vowels following Procrustes normalization. Qin *et al.* (2008) described a method to predict tongue contours (as measured via ultrasound imaging) from a few landmarks (EMA pellet positions). Using a radial basis function network, the authors were able to reconstruct full tongue contours with 0.3–0.2 mm errors using only 3 or 4 landmarks. In a follow-up study (Qin and Carreira-Perpinán, 2009), the authors proposed an articulatory mapping to adapt the previous predictive model to a new speaker using a 2D-wise linear alignment mapping. Their results show that a small adaptation corpus (about ten full tongue contours) is sufficient to recover very accurate (0.5 mm) predictive models for each new speaker. These studies suggest that a linear mapping can model a significant amount of inter-speaker differences in vocal tract geometry.

More recently, Felps *et al.* (2014) extended the Procrustes transformation of EMA position data by allowing independent local translation at each articulatory fleshpoint and observed further reduction in the inter-speaker differences. The independent translation parameters for each fleshpoint allowed the transform to adjust for the non-uniform positioning of the articulatory fleshpoints across speakers. Additional reduction in inter-speaker differences may be achieved by allowing independent scaling and rotation parameters for each fleshpoint.

III. METHODS

Our proposed articulatory method for accent conversion follows the generic outline shown in Fig. 1. The method takes an acoustic-articulatory trajectory from an L1 test utterance and transforms it to match the voice quality of the L2 speaker. In a first step, the method normalizes the L1 articulatory trajectory to the L2 articulatory space (Sec. III A). Then, it uses the normalized L1 trajectories as an input to a GMM-based articulatory synthesizer trained on an L2 acoustic-articulatory corpus. The result is an utterance that has the articulatory gestures and prosody of the L1 speaker but the voice quality of the L2 speaker. Both procedures are described in detail in the following sections.

A. Cross-speaker articulatory mapping

The articulatory mapping transforms a vector $\mathbf{x}_{\mathrm{L1}}$ of EMA articulatory coordinates for the L1 speaker into the equivalent articulatory positions $\hat{\mathbf{x}}_{\mathrm{L2}} = f_{12}(\mathbf{x}_{\mathrm{L1}})$, where $f_{12}(\cdot)$ denotes a set of Procrustes transforms, one for each fleshpoint. Namely, given an L1 fleshpoint with anteroposterior and superoinferior coordinates $(x_{\mathrm{L1},a}, x_{\mathrm{L1},s})$, the function estimates the L2 fleshpoint coordinates $(\hat{x}_{\mathrm{L2},a}, \hat{x}_{\mathrm{L2},s})$ as

$$[\hat{x}_{\mathrm{L2},a}, \hat{x}_{\mathrm{L2},s}] = [c_a, c_s] + \rho\,[x_{\mathrm{L1},a,\mathrm{L1},s}]\,\mathbf{A}, \tag{1}$$

where $[c_a, c_s]$ is the translation vector, ρ is the scaling factor and, $\mathbf{A}$ is a 2×2 matrix representing the rotation and reflection. We estimate the Procrustes parameters $\{c_a, c_s, \rho, \mathbf{A}\}$ by solving the minimization problem

$$\{c_a, c_s, \rho, \mathbf{A}\} \overset{\min}{\sum_{\text{all landmarks}}} \|[x_{\mathrm{L2},a}, x_{\mathrm{L2},s}] - ([c_a, c_s] + \rho[x_{\mathrm{L1},a}, x_{\mathrm{L1},s}]\mathbf{A})\|, \tag{2}$$

where $[x_{\mathrm{L2},a}, x_{\mathrm{L2},s}]$ and $[x_{\mathrm{L1},a}, x_{\mathrm{L1},s}]$ are the coordinates of corresponding landmarks in the L2 and L1 speaker,

respectively. These parameters are learned for each pellet in the articulatory corpus.

Following Geng and Mooshammer (2009), we select a set of articulatory landmarks from the phonetically transcribed corpus. Namely, for each phone in the L1 inventory and for each speaker, we calculate the centroid of the EMA articulatory coordinates as the average across all frames that belong to the phone (according to the phonetic transcription). These pairs of phone centroids (one from the L1 speaker, one from the L2 speaker) are then used as the corresponding landmarks in Eq. (2). The overall approach is summarized in Fig. 2.

B. Forward mapping

To generate acoustic observations from articulatory positions, we use a GMM-based forward mapping model (Toda *et al.*, 2008) which incorporates global variance (GV) of the acoustic features (Toda *et al.*, 2007). The forward mapping estimates the temporal sequence of static acoustic parameters, (MFCC_{1-24}), from the trajectory of articulatory features $\mathbf{x}$. For each frame at time t the articulatory feature vector $\mathbf{x}_t$ consists of 15 parameters: the anteroposterior and superoinferior coordinate of six EMA pellets, pitch $(\log f_0)$, loudness (MFCC_0), and nasality. Since the velum position is not available in our EMA corpus, we used the text transcription of the utterances to generate a binary feature that represented nasality. In the absence of a transcription, the nasality feature may be derived from acoustic features—see Pruthi and Espy-Wilson (2004)—as in the case for fundamental frequency and loudness.

For completeness, we include a detailed description of the forward mapping in (Toda *et al.*, 2007, 2008). In a first step, we model the joint distribution of articulatory-acoustic features $\mathbf{Z}_t = [\mathbf{x}_t, \mathbf{Y}_t]$, where $\mathbf{x}_t$ is the articulatory feature vector at time t, and $\mathbf{Y}_t = [\mathbf{y}_t, \Delta\mathbf{y}_t]$ is an acoustic feature vector containing both static and delta MFCCs. Using a Gaussian mixture, the joint distribution becomes

$$p(\mathbf{Z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{3}$$

where α_m is the scalar weight of the mth mixture component and $\mathcal{N}(\mathbf{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ is the Gaussian distribution with mean $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$. We use symbol $\boldsymbol{\lambda}^{(z)} = \{a_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\}$ to denote the full parameter set for the GMM. The mean vector $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$ denote the joint statistics of articulatory and acoustic features for the mth mixture

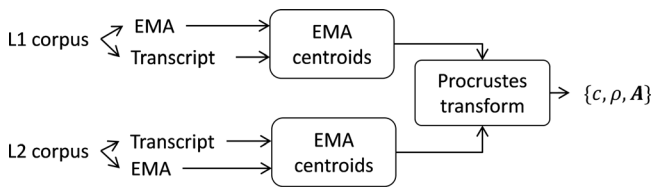

FIG. 2. Overview of the cross-speaker articulatory normalization procedure. A separate set of parameters is obtained for each EMA pellet.

$$\boldsymbol{\mu}_m^{(z)} = [\boldsymbol{\mu}_m^{(x)}\ \boldsymbol{\mu}_m^{(Y)}],\ \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xY)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}. \tag{4}$$

In a second step, we model the GV of predicted acoustics to account for over-smoothing effects of the GMM. Consider the within-sentence variance of the dth acoustic feature $y_t(d)$, given by $v(d) = E[(y_t(d) - E[y_t(d)])^2]$. The GV of these features in an utterance $\mathbf{y}(= [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T])$ is then given by a vector $\mathbf{v}(\mathbf{y}) = [v(1), v(2), \ldots, v(D)]$, where D is the dimension of acoustic vector $\mathbf{y}_t$. We model the distribution of GVs for all the utterances in the training set, $P(\mathbf{v}(\mathbf{y})|\boldsymbol{\lambda}^{(v)})$, with a single Gaussian distribution,

$$p(\mathbf{v}(\mathbf{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\mathbf{v}(\mathbf{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}), \tag{5}$$

where model parameters $\boldsymbol{\lambda}^{(v)} = \{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\}$ denote the vector of average global variance $\boldsymbol{\mu}^{(v)}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}^{(vv)}$, learned from the distribution of $\mathbf{v}(\mathbf{y})$ in the training set.

At synthesis time, given the trained models $[\boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}]$ and a test sequence of articulatory vectors $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_T]$, we obtain the maximum-likelihood acoustic (static only) trajectory $\hat{\mathbf{y}}$,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{Y}|\mathbf{x}, \boldsymbol{\lambda}^{(z)})^{\omega} P(\mathbf{v}(\mathbf{y})|\boldsymbol{\lambda}^{(v)}), \tag{6}$$

where $\mathbf{Y} = [\mathbf{y}_1, \Delta\mathbf{y}_1, \mathbf{y}_2, \Delta\mathbf{y}_2, \ldots, \mathbf{y}_T, \Delta\mathbf{y}_T]$ is the time sequence of acoustic vectors (both static and dynamic) and $\mathbf{v}(\mathbf{y})$ is the variance of static acoustic feature vectors. The power term $\omega\,(= 1/2T)$ in Eq. (6) provides a balance between the two likelihoods. We solve for $\hat{\mathbf{y}}$ in Eq. (6) via expectation-maximization; for details, refer to Toda *et al.* (2007).

C. Pitch modification

As discussed earlier (see Sec. II A), prosody modification is an important part of accent conversion. Following Toda *et al.* (2007), we use the pitch trajectory of the L1 speaker, which captures the native intonation pattern, but normalize it to the pitch range of the L2 speaker to preserve his or her natural vocal range. More specifically, given an L1 pitch trajectory $p_{\mathrm{L1}}(t)$, we generate the pitch trajectory $p_{\mathrm{L2}}(t)$ in L2 pitch range as

$$\log(p_{\mathrm{L2}}(t)) = [\log(p_{\mathrm{L1}}(t)) - \mu_{\mathrm{L1}}] \frac{\sigma_{\mathrm{L2}}}{\sigma_{\mathrm{L1}}} + \mu_{\mathrm{L2}}, \tag{7}$$

where $(\mu_{\mathrm{L1}}, \sigma_{\mathrm{L1}})$ and $(\mu_{\mathrm{L2}}, \sigma_{\mathrm{L2}})$ are the mean and standard deviation of log-scaled pitch of the L1 and L2 speakers, respectively, as estimated from the training corpus. The estimated pitch trajectory $p_{\mathrm{L2}}(t)$ is used as an input to the GMM and in the final STRAIGHT waveform generation, as discussed next.

D. Acoustic processing

We use STRAIGHT (Kawahara, 1997) to extract acoustic features and synthesize the resulting speech waveform.

Given a L1 utterance, we extract pitch (f_0) aperiodicity and spectral envelope with STRAIGHT. For each frame (sampled at 200 Hz to match the EMA recordings), we then compute MFCC_{0-24} by warping the STRAIGHT spectral envelope according to the mel frequency scale (25 mel filterbanks, 8 kHz cutoff frequency) and applying a type-II discrete cosine transformation (DCT); the first coefficient (MFCC_0) becomes the energy of the frame. Next, we modify the L1 pitch trajectory to match the L2 pitch range, as described in Sec. III C. The frame energy (MFCC_0) and the modified pitch (p_{L2}) are combined with the normalized L1 EMA positions (described in Sec. III A) and the binary nasality feature to form an input articulatory feature vector for the L2 forward mapping (described in Sec. III B), which generates an estimate of the L2 spectral coefficients (MFCC_{1-24}).

Following Aryal and Gutierrez-Osuna (2013), we then reconstruct the STRAIGHT spectral envelope from the estimated L2 spectral coefficients (MFCC_{1-24}) and the L1 energy (MFCC_0). Specifically, given a vector of predicted MFCCs, the least-squares estimate of the spectral envelope is $\hat{\mathbf{s}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{e}$, where $\mathbf{F}$ is the mel frequency filter bank matrix used to extract MFCCs from the STRAIGHT spectrum, and $\mathbf{e}$ is the exponential of the inverse DCT of MFCCs. In a final step, we use the STRAIGHT synthesis engine to generate the waveform using the estimated spectral envelope $\hat{\mathbf{s}}$, the L1 aperiodicity and the modified pitch. The overall process is summarized in Fig. 3.

IV. EXPERIMENTAL PROTOCOL

We performed a series of perceptual listening experiments to evaluate the proposed method in terms of its ability to improve intelligibility, reduce non-native accentedness, and preserve voice individuality. For this purpose, we used a corpus of audio and articulatory recordings from a native speaker of American English, and a non-native speaker whose first language was Spanish (Felps *et al.*, 2012; Aryal and Gutierrez-Osuna, 2013) collected at the University of Edinburgh by means of EMA (Carstens AG500). Both speakers recorded the same 344 sentences chosen from the Glasgow Herald corpus. The non-native speaker recorded an additional 305 sentences from the same corpus. Out of the 344 common sentences, we randomly selected 50 sentences (220 s of active speech in total; 4.40 s/sentence on average) for testing, and used the remaining 294 sentences (1290 s total; 4.39 s/sentence) to train the forward mapping and the articulatory mapping. Six standard EMA pellets positions were recorded: upper lip, lower lip, lower jaw, tongue tip, tongue body, and tongue dorsum. Four additional pellets (placed behind the ears, the upper nasion, and the upper jaw) were used to cancel head motion and provide a frame of reference. EMA pellet positions were recorded at 200 Hz. From each acoustic recording, we also extracted pitch, aperiodicity and spectral envelope using STRAIGHT (Kawahara, 1997). MFCCs were then estimated from the STRAIGHT spectrum and resampled to match the EMA recordings. The result was a database of articulatory-acoustic feature vectors containing pitch, MFCC_{0-24} and six EMA positions per frame.

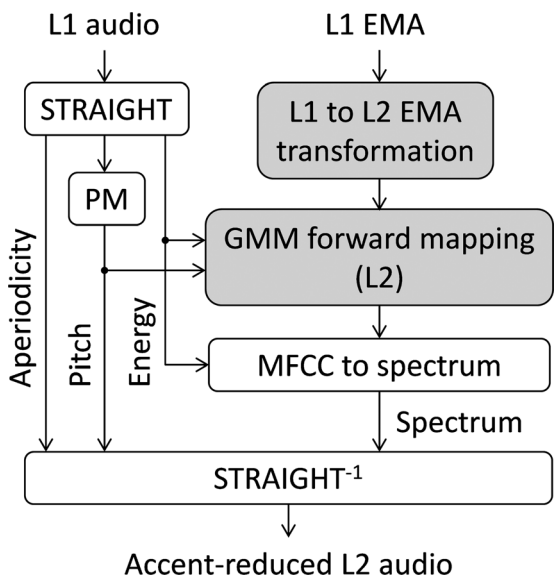

FIG. 3. Block diagram of accent conversion method (PM: pitch modification).

A. Experimental conditions

We considered five different experimental conditions for the listening tests: the proposed accent conversion method (AC), articulatory synthesis of L2 utterances (L2_{EMA}), articulatory synthesis of L1 utterances (L1_{EMA}), MFCC compression of L2 speech (L2_{MFCC}), and normalization of L1 utterances to match the vocal tract length and pitch range of L2 (L1_{GUISE}). The conditions are summarized in Table I.

The first experimental condition (AC) was the proposed accent conversion method, illustrated in Fig. 1. Namely, we built an L2 forward mapping by training a GMM with 128 mixtures on L2 articulatory-acoustic frames, and the Procrustes articulatory registration model by training on the articulatory landmarks of Eq. (2); only non-silent frames in the 294 training sentences were used for this purpose. Once the cross-speaker articulatory mapping and L2 forward mapping had been trained, we performed accent-conversion for each of the L1 utterances not used for training, following the procedure outlined in Fig. 3.

The second experimental condition (L2_{EMA}) consisted of articulatory synthesis of L2 utterances, obtained by driving the L2 forward model with L2 articulators. This

TABLE I. Four experimental conditions for listening test.

Experimental condition	Aperiodicity and energy	Pitch	Articulators	Spectrum
AC	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping
L2_{EMA}	L2	L2	L2	L2 forward mapping
L1_{EMA}	L1	L1	L1	L1 forward mapping
L2_{MFCC}	L2	L2	N/A	L2 MFCC
L1_{GUISE}	L1	L1 scaled to L2	N/A	L1 warped to L2

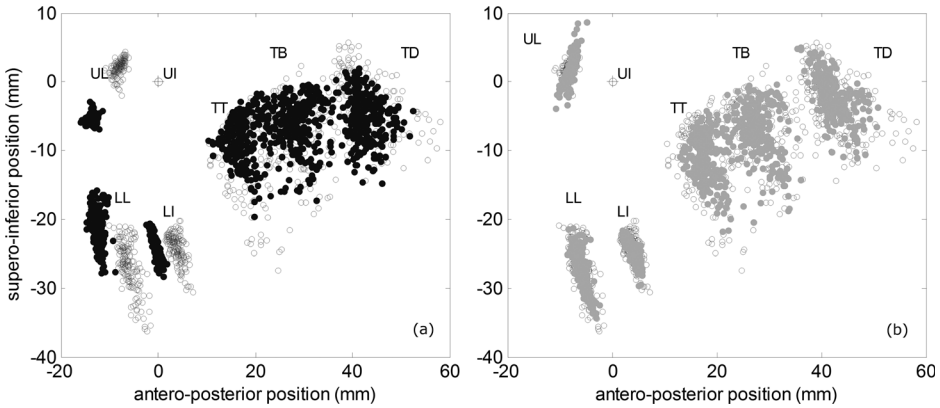

FIG. 4. (a) Distribution of six EMA pellet positions from the L1 speaker (solid markers) and L2 speaker (hollow markers) from a parallel corpus. Large differences can be seen in the span of the measured positions of articulators (UL: upper lip; LL: lower lip; LI: lower incisor; TT: tongue tip; TB: tongue blade; and TD: tongue dorsum). The upper incisor (UI) was used as a reference point. (b) Distribution of EMA pellet positions for the L1 speaker (solid markers) and L2 speaker (hollow markers) following articulatory normalization.

condition was used as the baseline for non-native production of the utterances since it had similar acoustic quality as AC. Because articulatory synthesis results in a loss of acoustic quality, and considering that acoustic quality interacts with accent perception (Felps and Gutierrez-Osuna, 2010), comparing AC against the L2 original utterances would have been problematic.

The third experimental condition ($L1_{EMA}$) consisted of articulatory synthesis of L1 utterances, obtained by driving an L1 forward model with L1 articulators. This condition served as the baseline for native production of the utterances, accounting for the loss of quality due to articulatory synthesis. This condition may also be taken as an upper bound of what accent conversion may be able to achieve in terms of intelligibility and accentedness.

The fourth experimental condition ($L2_{MFCC}$) consisted of re-synthesizing the original L2 utterances following compression into MFCCs. Utterances in this condition underwent a four-step process: (1) STRAIGHT analysis, (2) compression of STRAIGHT smooth spectra into MFCCs, (3) reconstruction of STRAIGHT smooth spectra from MFCCs, and (4) STRAIGHT synthesis; refer to Sec. III D for details on steps (2) and (3). This modification enabled a fair comparison against AC utterances by factoring out losses in acoustic quality caused by the MFCC compression step in Fig. 3.

The fifth experimental condition ($L1_{GUISE}$) consisted of modifying L1 utterances to match the pitch range and vocal tract length of the L2 speaker. This condition allowed us to test whether a simple guise could achieve similar accent-conversion performance as the proposed AC method: as shown in a number of studies (Lavner *et al.*, 2000, and references therein), pitch range and formant frequencies are good indicators of voice identity. Utterances in the $L1_{GUISE}$ condition were synthesized as follows. First, the L1 pitch trajectory was rescaled and shifted to match the pitch range of L2 speaker using Eq. (7). Then, we performed vocal tract length normalization by warping the L1 STRAIGHT spectrum to match the global statistics of the L2 speaker. Following Sundermann *et al.* (2003), we used a piecewise linear warping function governed by the average formant pairs of the two speakers, estimated over the training corpus; formants were extracted from the roots of the LPC coefficients of non-silent frames. For similar reasons as those described above, $L1_{GUISE}$ utterances also underwent the same MFCC compression procedure of $L2_{MFCC}$ utterances.

B. Participant recruitment

We evaluated the proposed method (AC) by comparing against the other four experimental conditions ($L2_{EMA}$, $L1_{EMA}$, $L2_{MFCC}$, $L1_{GUISE}$) in terms of intelligibility, accentedness, and speaker individuality through a series of perceptual listening tests. Following our prior work (Felps *et al.*, 2012; Aryal *et al.*, 2013; Aryal and Gutierrez-Osuna, 2013), participants for the perceptual studies were recruited through Mechanical Turk, Amazon's online crowdsourcing tool. In order to qualify for the studies, participants were required to reside in the United States and pass a screening test that consisted of identifying various American English accents: Northeast (i.e., Boston, New York), Southern (i.e., Georgia, Texas, Louisiana), and General American (i.e., Indiana, Iowa). Participants who did not pass this qualification task were not allowed to participate in the studies. In addition, participants were asked to list their native language/dialect and any other fluent languages that they spoke. If a subject was not a monolingual speaker of American English then their responses were excluded from the results. In the quality and accent evaluation tests, participants were asked to transcribe the utterances to ensure they paid attention to the recordings. Participants with incomplete responses were excluded from the study.

V. RESULTS

A. Accuracy of articulatory normalization

In a first experiment, we analyzed the effect of the Procrustes transforms on the distribution of articulatory configurations. First, we compared the spatial distribution of the six EMA pellets for the L1 and L2 speakers before and after articulatory normalization. Figure 4(a) shows the distribution before articulatory normalization; differences between the two speakers are quite significant, not only in terms of the average position of each pellet but also in terms of its spatial distribution (e.g., variance). These discrepancies can be attributed largely to differences in vocal tract geometry between the two speakers, though inconsistencies in pellet placement during the EMA recordings also play a role. Regardless of the source of these discrepancies, the results in Fig. 4(b) shows that the articulatory normalization step achieves a high degree of consistency in the spatial distribution of pellets between the two speakers.

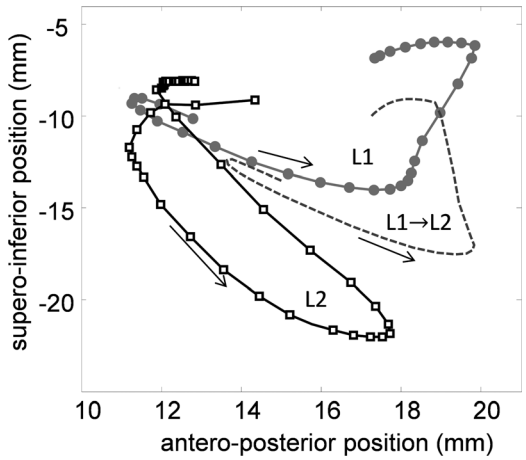

FIG. 5. Trajectory of the tongue-tip pellet in L1 and L2 utterances of the word "that." The L1 trajectory normalized to the L2 articulatory space is also shown. Arrows indicate the direction of trajectories.

Next, we compared articulatory trajectories for the L1 speaker, the L2 speaker, and the L1 after articulatory normalization. Figure 5 shows the trajectory of tongue tip for the word "that" in a test utterance. As a result of the normalization step, the L1 articulatory trajectory becomes closer to the L2 trajectory but also preserves the dynamics of the L1 production; this makes it easier to spot articulatory errors in the L2 utterance. Namely, the fig shows a noticeable difference between the L2 trajectory and the L1-normalized trajectory in antero-posterior position towards the end of the word. This discrepancy can be traced back to a typical phonetic substitution of alveolar stop [t] with the dental one [t̪] in L2 speakers whose mother tongue is Spanish, which results from moving the tongue tip forward to make a constriction at the teeth instead of the alveolar ridge. Such display of normalized trajectories may also be used as supplementary feedback mechanism to the learner in computer-assisted pronunciation training.

Finally, we analyzed the effect of articulatory normalization on the distribution of articulatory configurations at the phonetic level; the middle frame of vowel segments was used for this purpose. Figure 6(a) shows the centroid and half-sigma contour (i.e., half standard deviation) of the tongue tip pellet position, a critical articulator for the frontal vowels ([ɪ], [i], [ɛ], [e], and [æ]), for the two speakers (L1 and L2). As shown in Fig. 6(a), the half-sigma contours for corresponding vowels in the two speakers have no overlap, with the exception of [ɪ] and [ɛ]. Notice also the larger spread in articulatory configurations for the L2 speaker compared to the L1 speaker, a result that is consistent with prior studies showing larger acoustic variability and phonemic overlap in non-native speech productions (Wade *et al.*, 2007). Figure 6(b) shows the articulatory configurations following the articulatory normalization step; vowel centroids for the normalized L1 speaker are within the half-sigma contour of the corresponding vowel for the L2 speaker.

B. Assessment of intelligibility

In a first listening test we assessed the intelligibility of AC as compared to $L1_{EMA}$ and $L2_{EMA}$ utterances. Three independent groups of native speakers of American English ($N = 15$ each) transcribed the 46 test utterances[4] for the three experimental conditions (AC, $L1_{EMA}$, $L2_{EMA}$). From each transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly identified words to the total number of words in the utterance. Participants also rated the (subjective) intelligibility of the utterances (S_{intel}) using a seven-point Likert scale (1: not intelligible at all, 3: somewhat intelligible, 5: quite a bit intelligible, and 7: extremely intelligible).

Figure 7 shows the word accuracy and intelligibility ratings for the three experimental conditions. Accent conversions ($AC : W_{acc} = 0.64, S_{intel} = 3.83$) were rated as being significantly more intelligible ($p < 0.01$; t–test) than L2 articulatory synthesis ($L2_{EMA} : W_{acc} = 0.50, S_{intel} = 3.30$), a result that supports the feasibility of our accent-conversion approach, though not as intelligible ($p < 0.01$; t–test) as the upper bound of L1 articulatory synthesis ($L1_{EMA} : W_{acc} = 0.90, S_{intel} = 4.96$). In all three conditions, the two intelligibility measures (W_{acc}, S_{intel}) were found to be significantly correlated ($\rho > 0.89, N = 46$); for this reason, in what follows we will focus on W_{acc} as it is the more objective of the two measures.

The scatter plot in Fig. 8 shows the AC and $L2_{EMA}$ word accuracies for the 46 test sentences. In 70% of the cases (32 sentences; those above the main diagonal in the figure)

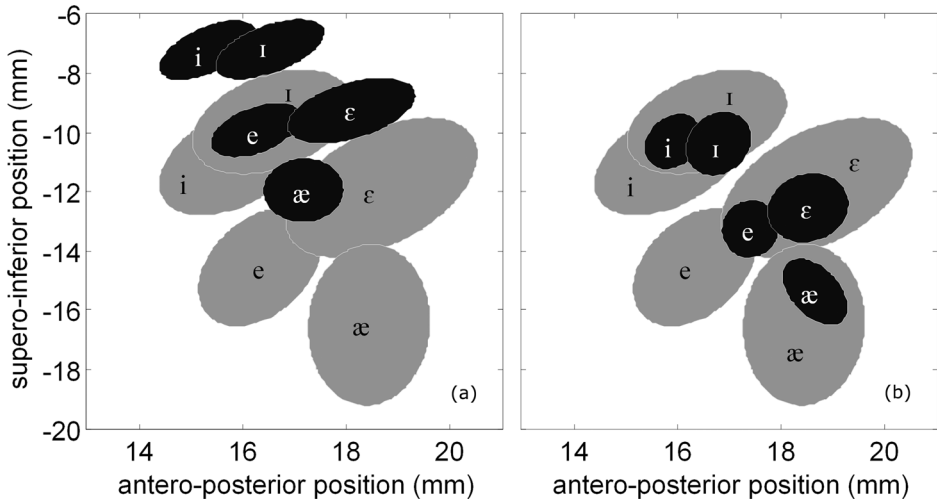

FIG. 6. (a) Distribution of tongue tip position in frontal vowels for the L1 speaker (dark ellipses) and L2 speaker (light) speaker; ellipses represent the half-sigma contour of the distribution for each vowel. (b) Distribution of tongue tip position in frontal vowels for the L1 speaker after articulatory mapping (dark) and the L2 speaker (light).

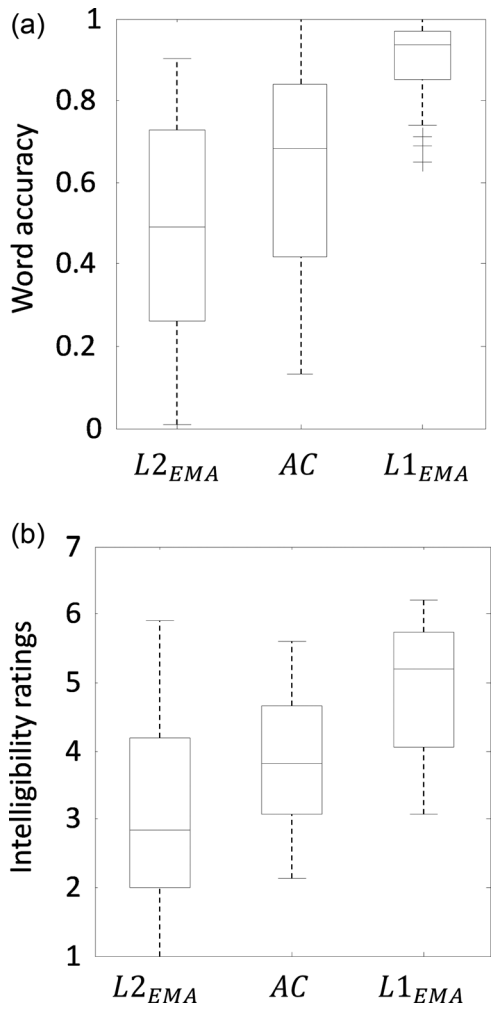

FIG. 7. Box plot of (a) word accuracy and (b) subjective intelligibility ratings for $L1_{EMA}$, $L2_{EMA}$, and AC utterances.

accent conversion improved word accuracy compared to that obtained on $L2_{EMA}$ utterances, further supporting our approach. Notice, however, the lack of correlation between the two conditions, an unexpected result since one would expect that the initial word accuracy (i.e., on $L2_{EMA}$ utterances) would have a strong influence on word accuracy following accent conversion. As will be discussed next, this result suggests the presence of two independent factors affecting intelligibility in the two conditions.

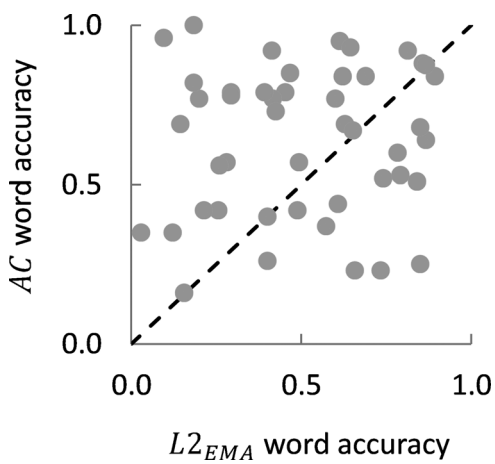

FIG. 8. Word accuracy for AC and $L2_{EMA}$ for the 46 test sentences. The diagonal dashed line represents the sentences for which $W_{acc}(AC) > W_{acc}(L2_{EMA})$ are above the dashed line and vice versa.

The results in Fig. 7(a) also show a large variance in word accuracy for L2 articulatory synthesis ($L2_{EMA}$) compared to L1 articulatory synthesis ($L1_{EMA}$). In previous work (Aryal and Gutierrez-Osuna, 2014b) we found that accent conversions are most beneficial when used on utterances that are difficult to produce by L2 speakers based on differences between the L1 and L2 phonetic inventories. Accordingly, we examined whether the variance in word accuracy for $L2_{EMA}$ could be explained by the phonetic complexity of each sentence, measured as the number of L1 phones in the sentence that do not exist in the L2 inventory. Differences in phonetic inventories are a known reason behind non-native accents; see, e.g., You *et al.* (2005). In our case, the English language includes a number of consonants that do not exist in Spanish (our L2 speaker's mother tongue), most significantly the fricatives [v], [z], [θ], [ʃ], [ʒ] and [ð], the affricate [ʤ], the pseudo-fricative [h], and the liquid [ɹ]. Spanish also does not have lax vowels, the schwa as well as r-colored vowels. Thus, for each test sentence we computed the number of phones that did not exist in Spanish ($N_{p\notin L2}$) and compared it against the $L2_{EMA}$ word accuracy. Both variables ($N_{p\notin L2}$, $W_{acc}(L2_{EMA})$) are significantly correlated ($\rho = 0.44, N = 46, p < 0.01$), suggesting that variance in intelligibility for $L2_{EMA}$ utterances can be explained by differences in the L1 and L2 phonetic inventory. We found, however, no significant correlation ($\rho = -0.2; p = 0.09$) between $N_{p\notin L2}$ and word accuracy for AC utterances, which suggests that the accent conversion process is able to cancel out the main source of (poor) intelligibility: phonetic complexity from the perspective of the L2 learner.

What then, if not sentence complexity, drives the intelligibility of AC utterances? Since both conditions (AC, $L2_{EMA}$) use the same articulatory synthesizer, we hypothesized that interpolation issues would be at fault. To test this hypothesis, for each frame in an AC utterance we computed the Mahalanobis distance between the L1 registered articulators and the centroid of the corresponding L2 phone, then averaged the distance over all non-silent frames in the utterance. The larger this measure, the larger the excursion of the registered L1 articulatory trajectory from the L2 articulatory space. We found, however, no significant correlation ($\rho = -0.21; p = 0.08$) between this measure and word accuracy on AC utterances, which suggests that the total amount of interpolation present in an AC utterance does not explain its lack of intelligibility.

In a final analysis we then decided to test whether the phonetic content of the utterance would explain its intelligibility, our rationale being that the acoustic effect of interpolation errors is not uniform across phones. As an example, due to the presence of critical articulators, a small error in the tongue tip height can transform a stop into a fricative whereas the same amount of error in tongue tip height may not make much of a difference in a vowel. Accordingly, we calculated the correlation between word accuracy and the proportion of phones in an utterance with a specified

TABLE II. Correlation between word accuracy and the proportion of phones in a sentence containing a particular articulatory-phonetic feature.

	Articulatory features	AC	$L2_{EMA}$	$L1_{EMA}$
Manner	Stops	**−0.43**	0.22	−0.21
	Fricatives	−0.01	0.04	−0.02
	Affricates	0.05	−0.17	0.05
	Nasals	0.31	−0.10	0.17
	Liquids	−0.19	−0.08	−0.22
	Glides	**0.40**	0.01	0.17
Place	Bilabials	−0.07	0.28	−0.07
	Labiodentals	0.14	−0.11	−0.03
	Lingual dental	−0.18	−0.03	−0.12
	Lingual alveolar	−0.04	−0.12	0.10
	Lingual palatal	0.02	−0.21	−0.04
	Lingual velar	0.01	0.25	−0.08
	Glottal	0.01	0.14	−0.09
Voicing	Voiced	0.01	−0.07	−0.18
	Unvoiced	−0.10	0.14	0.07

phonetic-articulatory feature. Results are shown in Table II for six features of manner of articulation, seven features of place of articulation, and voicing. Correlation coefficients found to be significant ($p < 0.01$) are shown in bold. In the case of $L1_{EMA}$ and $L2_{EMA}$ utterances, we found no significant effect on intelligibility for any of the articulatory features, an indication that the GMM articulatory synthesizer was trained properly. In the case of AC utterances, however, we found a strong negative correlation between intelligibility and the proportion of stops in the sentence. Thus, it appears that small registration errors, to which stops are particularly sensitive, are largely responsible for the loss of intelligibility in accent-converted utterances.[5]

C. Assessment of non-native accentedness

In a second listening experiment we sought to determine whether the proposed accent-conversion method could also reduce the perceived non-native accent of L2 utterances. Following our previous work (Aryal and Gutierrez-Osuna, 2014a,b), participants were asked to listen to $L2_{EMA}$ and AC utterances of the same sentence and select the most native-like[6] among them. For this test, we focused on the subset of sentences for which AC and $L2_{EMA}$ utterances had higher intelligibility ($W_{acc} > 0.5$); i.e., those on the upper-right quadrant in Fig. 8. In this way, we avoided asking participants to rate which of two unintelligible utterances was less foreign-accented (a questionable exercise) or whether an unintelligible utterance was more foreign-accented than an intelligible one (an exercise of predictable if not obvious results).

Participants ($N = 15$) listened to 30 pairs of utterances (15 AC-$L2_{EMA}$ pairs, and 15 $L2_{EMA}$-AC pairs) presented in random order to account for order effects. Their preferences are summarized in Fig. 9. AC utterances were rated as being more native than $L2_{EMA}$ utterances in 62% of the sentences (SE 4%), which is significantly higher than the 50% chance level [$t = 3.2$, degrees of freedom (DF) = 14, $p = 0.003$, single tail]. This result indicates the proposed

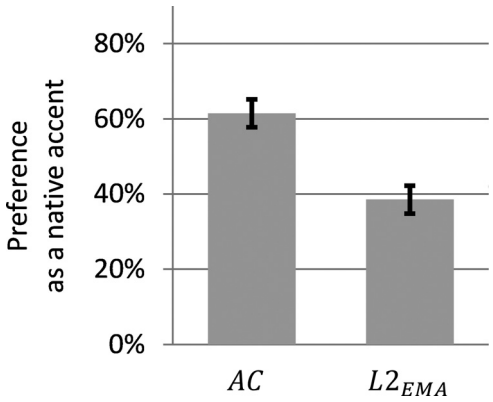

FIG. 9. Subjective evaluation of non-native accentedness. Participants were asked to determine which utterance in a pair was more native-like.

accent-conversion method can be effective in reducing the perceived non-native accent of L2 utterances. To verify that these results were not accidental (e.g., caused by the lower acoustic quality of articulatory synthesis), we performed an additional listening test to compare accent ratings for native ($L1_{EMA}$) and nonnative ($L2_{EMA}$) articulatory synthesis. In this test, a different group of listeners ($N = 15$) compared 30 pairs of utterances (15 $L1_{EMA}$-$L2_{EMA}$ pairs, and 15 $L2_{EMA}$-$L1_{EMA}$ pairs), and selected the most native-like utterance in a pair. As expected, $L1_{EMA}$ utterances were rated as more native than $L2_{EMA}$ in 96% of the cases, which indicates that articulatory syntheses do retain dialect/accent information.

Closer inspection of the listeners' responses to the accent perception comparisons showed an influence of presentation order within pairs. Namely, AC was rated as more native than $L2_{EMA}$ 53% of the times whenever AC appeared first, but the proportion increased to 70% if AC was the second utterance in the pair; this difference was statistically significant ($t = 4.0$, DF = 14, $p = < 0.001$, single tail). This bias is consistent with the "pop-out" effect (Davis *et al.*, 2005), according to which a degraded utterance is perceived as being less degraded if presented after a clean version of the same utterance, i.e., when the lexical information is known. Extending this result to the perception of native accents, $L2_{EMA}$ may then be treated as the degraded utterances relative to the AC condition, which would explain why $L2_{EMA}$ utterances were rated as less accented if they were presented after AC.

D. Assessment of voice individuality

In a third and final listening experiment we tested the extent to which the accent conversion method was able to preserve the voice identity of the L2 speaker. For this purpose, we compared AC utterances against $L2_{MFCC}$ utterances (MFCC compressions of the original L2 recordings) and $L1_{GUISE}$ utterances (a simple guise of L1 utterances to match the vocal tract length and pitch range of the L2 speaker).

Following previous work (Felps *et al.*, 2009), we presented participants with a pair of linguistically different utterances from two of the three experimental conditions. Presentation order was randomized for conditions within

each pair and for pairs of conditions. Participants ($N = 15$) rated 40 pairs, 20 from each group (AC-$L2_{MFCC}$, $L1_{GUISE}$-$L2_{MFCC}$) randomly interleaved, and were asked to (1) determine if the utterances were from the same or a different speaker (forced choice) and (2) rate how confident they were in their assessment using a seven-point Likert scale. Once the ratings were obtained, participants' responses and confident levels were combined to form a voice similarity score (VSS) ranging from −7 (extremely confident they are different speakers) to +7 (extremely confident they are the same speaker).

Figure 10 shows the mean VSS between pairs of experimental conditions. Listeners were "quite" confident that AC and $L2_{MFCC}$ utterances were from the same speaker [VSS = 4.2, standard error (s.e.) = 0.5]. This result suggests that the method is able to preserve the voice-identity of the L2 learner. Likewise, listeners were very confident (VSS = 5.9, s.e. = 0.3) that $L1_{GUISE}$ and $L2_{MFCC}$ utterances were from different speakers, which indicates that a simple guise of the L1 speaker cannot capture the voice quality of the L2 learner.

VI. DISCUSSION

We have presented an accent-conversion method that transforms non-native utterances to match the articulatory gestures of a reference native speaker. Our approach consists of building a GMM-based articulatory synthesizer of a non-native learner, then driving it with measured articulatory gestures from a native speaker. Results from listening tests show that accent conversion provides statistically significant increases in intelligibility, as measured by objective scores (word recognition), subjective ratings, and overall preference (70%) when compared to synthesis driven by L2 articulators. More importantly, unlike in the case of synthesis driven by L2 articulators, the intelligibility of accent conversions is not affected by the proportion of phones outside the phonetic inventory of the L2 speaker. This result suggests that the method can successfully remove one of the primary causes of non-native accents. Subsequent pairwise listening tests of native accentedness also show a preference towards accent conversions (62%) when compared to synthesis driven by L2 articulators. Finally, listening tests of speaker identity indicate that driving the L2 articulatory synthesizer with (registered) articulatory gestures from a different speaker does not change the perceived voice quality of the resulting synthesis. When combined with our results on intelligibility and accentedness, this finding suggests that our overall approach (L1 → L2 articulatory normalization followed by L2 articulatory synthesis) is an effective strategy to decouple those aspects of an utterance that are due to the speaker physiology from those that are due to the language.

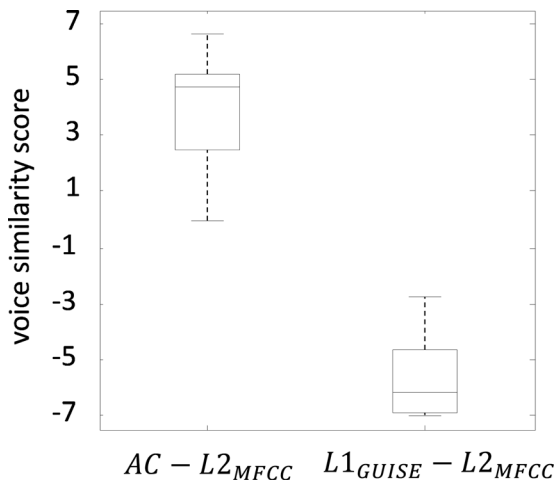

FIG. 10. Average pairwise voice similarity scores. Scores range from −7 (different speaker with high confidence) to 7 (same speaker with high confidence).

Further analysis indicates that the intelligibility of accent-converted utterances decreases with the proportion of stop consonants in the sentence. Given that stops require the formation of a complete constriction, small articulatory registration errors can have a significant effect on the acoustic output of the model; as an example, a small error in tongue-tip height may cause a lingua-alveolar stop to become fricative (e.g., from [t] to [s]). A potential solution to this problem may be to incorporate knowledge of critical articulators by replacing the mapping in Eq. (1) with one that is context-dependent. To this end, Felps *et al.* (2010) have shown that the accuracy of articulatory-acoustic mappings can be increased by using phone-specific weights for the EMA coordinates of critical articulators. Likewise, context-dependent articulatory mappings could be used to minimize errors in EMA pellet positions that are critical to each phone, in this fashion improving synthesis quality and accent-conversion performance. Additional information on vocal tract geometry may also be used to improve synthesis performance. As an example, having access to the palate contour may be used to compute the distance (or contact) between passive and active articulators, or to extract tract constriction variables, which are known to have less variability than EMA pellet positions (McGowan, 1994; Mitra *et al.*, 2011).

The reader may question whether an articulatory representation other than the raw (x,y) position of EMA pellets could have been used here to reduce speaker differences. Indeed, in prior work (Felps *et al.*, 2012; Aryal and Gutierrez-Osuna, 2013) we had chosen to transform EMA pellet positions into the six-point Maeda parameter approximations of Al Bawab *et al.* (2008), then z-score normalized them to further reduce differences in speaker vocal-tract geometry. During the early stages of the present study, however, we observed that acoustic quality was markedly better when raw EMA pellet positions, rather than z-score normalized Maeda parameters approximations, were used as an input to the articulatory synthesizer. This can be partly attributed to the fact that projecting a 12-dimensional representation (x and y position for six EMA pellets), into a six-dimensional representation (Maeda parameters) is a lossy step. More importantly, although applying the Maeda parameterization and within speaker z-scoring increases the correlation between pairs of articulatory parameters from the two

speakers compared to the EMA positions (TT: from 0.68 to 0.71, TD: from 0.61 to 0.64), the correlation coefficient for the tongue-body-shape (TBS) Maeda parameter, which involves a non-linear transformation, becomes negligible ($\rho = 0.15$). This makes it more difficult for the cross-speaker articulatory mapping in Eq. (1) to bring Maeda parameter approximations from the two speakers into registration. TBS is a critical parameter in articulatory synthesis and the mapping error has significant impact on synthesis quality. In summary, our results suggest that, while Maeda parameter approximations are certainly less speaker-dependent, accurate registration (as needed for cross-speaker articulatory synthesis) is more easily achieved by transforming the raw EMA pellet positions. It is also possible that standard Maeda control parameters may show fewer inter-speaker differences than the Maeda approximations of Al Bawab *et al.* (2008); computing the standard Maeda parameters would require registering EMA pellet positions with the vocal tract geometry and the Maeda vocal tract grid; see Toutios and Narayanan (2013).

At present, our approach uses L1 aperiodicity spectra and therefore does not consider speaker individuality cues that may be contained in the L2 aperiodicity (Kawahara, 1997). Thus, further improvements in voice similarity may be obtained by replacing the L1 aperiodicity with its L2 equivalent. One possibility is to estimate L2 aperiodicity from the estimated L2 spectra by exploiting the relation between both signals.

The L2 speaker in our study had lived in the United States for 16 years at the time of recordings so, while he maintained a noticeable non-native accent, he was functionally bilingual. Additional work is needed to assess the effectiveness of our accent conversion method for L2 speakers with different levels of L1 proficiency and from different L1 backgrounds. Additional work is also needed to make our approach more practical for CAPT by avoiding the need to measure L2 articulators. As described in our recent work (Aryal and Gutierrez-Osuna, 2014a), one promising alternative is to build a cross-speaker articulatory synthesizer (i.e., a mapping from L1 articulators to L2 acoustics). Because such mapping would not require L2 articulators, it would also avoid issues stemming from inaccuracies in articulatory registration or missing articulatory configurations in the L2 corpus. Additional strategies consist of predicting articulatory features from speech acoustics, either in the form of articulatory phonetics features (King *et al.*, 2007), or through speaker-independent articulatory inversion (Ghosh and Narayanan, 2011).

ACKNOWLEDGMENTS

This work was supported by NSF Award No. 0713205. We are grateful to Professor Steve Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during the sabbatical stay of R.G.O. at CSTR (University of Edinburgh), and to Dr. Christian Geng for his assistance in performing the EMA recordings. We are also grateful to Professor Kawahara for permission to use the STRAIGHT analysis-synthesis method.

[1]We used the term "articulatory gestures" in a broader sense to represent the dynamics of vocal tract configurations. Not to be confused with "gestures" and "gestural scores" in the gestural framework of articulatory phonetics developed at Haskins Laboratories (Browman and Goldstein, 1990).

[2]Previous work on text-to-speech unit-selection synthesis shows that at least 2 h of active speech are needed to synthesize intelligible speech, a number that is rarely (if ever) achieved with articulatory corpora.

[3]1: Not at all, 3: Somewhat, 5: Quite a bit, 7: Extremely.

[4]Four of 50 test sentences for the L2 speaker had missing EMA data and were removed from the analysis.

[5]Table II also shows a strong positive correlation between intelligibility and glides, an unexpected result because it suggests that lowering the proportion of glides in an utterance reduces its intelligibility. A closer look at the phonetic composition of our 46 test utterances, however, shows that the proportion of glides is negatively correlated with the proportion of stops ($\rho = -0.32$, $p = 0.015$). This provides a more plausible explanation: as the proportion of glides decreases, so does the proportion of stops increase, in turn lowering the intelligibility of the utterance.

[6]Native relative to a monolingual speaker of General American English.

Al Bawab, Z., Bhiksha, R., and Stern, R. M. (**2008**). "Analysis-by-synthesis features for speech recognition," in *Proceedings of ICASSP*, Las Vegas, Nevada, pp. 4185–4188.

Arora, R., and Livescu, K. (**2013**). "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *Proceedings of ICASSP*, pp. 7135–7139.

Aryal, S., Felps, D., and Gutierrez-Osuna, R. (**2013**). "Foreign accent conversion through voice morphing," in *Proceedings of INTERSPEECH*, Lyon, France, pp. 3077–3081.

Aryal, S., and Gutierrez-Osuna, R. (**2013**). "Articulatory inversion and synthesis: Towards articulatory-based modification of speech," in *Proceedings of ICASSP*, pp. 7952–7956.

Aryal, S., and Gutierrez-Osuna, R. (**2014a**). "Accent conversion through cross-speaker articulatory synthesis," in *Proceedings of ICASSP*, Florence, Italy, pp. 7744–7748.

Aryal, S., and Gutierrez-Osuna, R. (**2014b**). "Can voice conversion be used to reduce non-native accents?," in *Proceedings of ICASSP*, Florence, Italy, pp. 7929–7933.

Browman, C. P., and Goldstein, L. (**1990**). "Gestural specification using dynamically-defined articulatory structures," J. Phonetics **18**, 299–320.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psych. Gen. **134**, 222–241.

Felps, D., Aryal, S., and Gutierrez-Osuna, R. (**2014**). "Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis," in *Proceedings of ICASSP*, Florence, Italy, pp. 3051–3055.

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

Felps, D., Geng, C., Berger, M., Richmond, K., and Gutierrez-Osuna, R. (**2010**). "Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database," in *Proceedings of INTERSPEECH*, Makuhari, Japan, pp. 1990–1993.

Felps, D., Geng, C., and Gutierrez-Osuna, R. (**2012**). "Foreign accent conversion through concatenative synthesis in the articulatory domain," IEEE Trans. Audio Speech Lang. Process. **20**, 2301–2312.

Felps, D., and Gutierrez-Osuna, R. (**2010**). "Developing objective measures of foreign-accent conversion," IEEE Trans. Audio, Speech Lang. Process. **18**, 1030–1040.

Geng, C., and Mooshammer, C. (**2009**). "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," J. Acoust. Soc. Am. **125**, 3278–3288.

Ghosh, P., and Narayanan, S. (**2011**). "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," J. Acoust. Soc. Am. **130**, EL251–EL257.

Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Bio. Cybernetics **72**, 43–53.

Hashi, M., Westbury, J. R., and Honda, K. (**1998**). "Vowel posture normalization," J. Acoust. Soc. Am. **104**, 2426–2437.

Huckvale, M., and Yanagisawa, K. (**2007**). "Spoken language conversion with accent morphing," in *Proceedings of ISCA Speech Synthesis Workshop*, Bonn, Germany, pp. 64–70.

Kaburagi, T., and Honda, M. (**1998**). "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proceedings of ICSLP*, Sydney, Australia, pp. 433–436.

Kawahara, H. (**1997**). "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proceedings of ICASSP*, Munich, Germany, pp. 1303–1306.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M. (**2007**). "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am. **121**, 723–742.

Lavner, Y., Gath, I., and Rosenhouse, J. (**2000**). "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**, 9–26.

Ling, Z. H., Richmond, K., Yamagishi, J., and Wang, R. H. (**2009**). "Integrating articulatory features into HMM-based parametric speech synthesis," IEEE Trans. Audio Speech Lang. Process. **17**, 1171–1185.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer Academic, Amsterdam), pp. 131–149.

McGowan, R. S. (**1994**). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Commun. **14**, 19–48.

Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070–1082.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (**2011**). "Speech inversion: Benefits of tract variables over pellet trajectories," in *Proceedings of ICASSP*, pp. 5188–5191.

Pruthi, T., and Espy-Wilson, C. Y. (**2004**). "Acoustic parameters for automatic detection of nasal manner," Speech Commun. **43**, 225–239.

Qin, C., and Carreira-Perpinán, M. A. (**2009**). "Adaptation of a predictive model of tongue shapes," in *Proceedings of INTERSPEECH*, Brighton, UK, pp. 772–775.

Qin, C., Carreira-Perpinán, M. A., Richmond, K., Wrench, A., and Renals, S. (**2008**). "Predicting tongue shapes from a few landmark locations," in *Proceedings of INTERSPEECH*, Brisbane, Australia, pp. 2306–2309.

Rosini, L.-G. (**1997**). *English with an Accent: Language, Ideology, and Discrimination in the United States* (Routledge, London), p. 286.

Rubin, P., Baer, T., and Mermelstein, P. (**1981**). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 321–328.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecologic. Psych. **1**, 333–382.

Sidaras, S. K., Alexander, J. E., and Nygaard, L. C. (**2009**). "Perceptual learning of systematic variation in Spanish-accented speech," J. Acoust. Soc. Am. **125**, 3306–3316.

Stevens, K. N., Kasowski, S., and Fant, C. G. M. (**1953**). "An electrical analog of the vocal tract," J. Acoust. Soc. Am. **25**, 734–742.

Sundermann, D., Ney, H., and Hoge, H. (**2003**). "VTLN-based cross-language voice conversion," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, pp. 676–681.

Toda, T., Black, A. W., and Tokuda, K. (**2007**). "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio Speech Lang. Process. **15**, 2222–2235.

Toda, T., Black, A. W., and Tokuda, K. (**2008**). "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Commun. **50**, 215–227.

Toutios, A., and Maeda, S. (**2012**). "Articulatory VCV Synthesis from EMA Data," in *Proceedings of INTERSPEECH*, Portland, Oregon, pp. 2566–2569.

Toutios, A., and Narayanan, S. (**2013**). "Articulatory Synthesis of French Connected Speech from EMA Data," in *Proceedings of INTERSPEECH*, Lyon, France, pp. 2738–2742.

Turk, O., and Arslan, L. M. (**2006**). "Robust processing techniques for voice conversion," Computer Speech Lang. **20**, 441–467.

Wade, T., Jongman, A., and Sereno, J. (**2007**). "Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds," Phonetica **64**, 122–144.

Westbury, J. R. (**1994**). *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0* (Waisman Center on Mental Retardation & Human Development, University of Wisconsin, Madison, WI), p. 135.

Wrench, A. A. (**2000**). "A multichannel articulatory database and its application for automatic speech recognition," in *Proceedings of 5th Seminar on Speech Production: Models and Data*, pp. 305–308.

Yan, Q., Vaseghi, S., Rentzos, D., and Ho, C.-H. (**2007**). "Analysis and synthesis of formant spaces of British, Australian, and American accents," IEEE Trans. Audio, Speech Lang. Process. **15**, 676–689.

You, H., Alwan, A., Kazemzadeh, A., and Narayanan, S. (**2005**). "Pronunciation variations of Spanish-accented English spoken by young children," in *Proceedings of INTERSPEECH*, Lisbon, Portugal, pp. 749–752.