ACCENT CONVERSION THROUGH CROSS-SPEAKER ARTICULATORY SYNTHESIS

Sandesh Aryal and Ricardo Gutierrez-Osuna
Department of Computer Science and Engineering, Texas A&M University
[sandesh,rgutier]@cse.tamu.edu

ABSTRACT

Accent conversion (AC) seeks to transform second-language (L2) utterances to appear as if produced with a native (L1) accent. In the acoustic domain, AC is difficult due to the complex interaction between linguistic content and voice quality. Alternatively, AC can be performed in the articulatory domain by building a mapping from L2 articulators to L2 acoustics, and then driving the model with L1 articulators. However, collecting articulatory data for each L2 learner is impractical. Here we propose an approach that avoids this expensive step. Our method builds a cross-speaker forward mapping (CSFM) to generate L2 acoustic observations directly from L1 articulatory trajectories. We evaluated the CSFM against a baseline articulatory synthesizer trained with L2 articulators. Subjective listening tests show that both methods perform comparably in terms of accent reduction and ability to preserve the voice quality of the L2 speaker, with only a small impact in acoustic quality.

Index terms– Data-driven articulatory synthesis, accent conversion, voice conversion

1. INTRODUCTION

Studies in computer assisted pronunciation training (CAPT) for second language (L2) learners have suggested that practice is more effective if the voice to imitate is similar to the learner's [1, 2]. In practical settings, however, it is not possible to find a matching voice for each potential learner. For this reason, various techniques have been proposed to transform utterances from L2 speakers to sound as if they had been produced with a native accent [3, 4]. In previous work [3], we have suggested that such techniques may be used to provide the "golden speaker" for each learner: their own voice but with a native accent.

Accent conversion can be performed in the acoustic [3-5] and articulatory domains [6]. Acoustic methods are appealing since audio recordings are easy to obtain. Given a parallel recording from the L2 learner and a native teacher (L1), these methods attempt to extract the linguistic content from the L1 utterance and combine it with the voice quality carrier from the L2 utterance. However, linguistic content and voice quality interact in complex ways in the acoustic domain, so the resulting blended utterance can often be perceived as having the identity of a third speaker [3]. These methods also require accurate time alignment between L1 and L2 utterances. In contrast, articulatory methods make linguistic gestures readily available via the position and trajectory of the measured articulators [6]. Briefly, these methods build an L2 articulatory synthesizer (i.e., a forward mapping from L2 articulators to L2 acoustics), and then drive the synthesizer with L1 articulators; see Fig. 1a. However, these methods are impractical for most CAPT settings since they require access to articulatory recordings from each L2 learner.

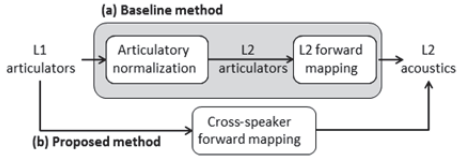

Fig. 1. (a) Conventional approach for articulatory-based accent conversion. (b) Proposed cross-speaker forward mapping.

In this paper, we propose an articulatory method for accent conversion that does not require L2 articulatory recordings. Given an acoustic-only corpus for the L2 learner and a joint acoustic-articulatory corpus from a single L1 teacher, the method builds a cross-speaker forward mapping (CSFM) to predict L2 acoustics directly from L1 articulators; see Fig. 1b. For this purpose, the method performs vocal tract length normalization (VTLN) to reduce differences in the L1 and L2 acoustic spaces that are due to vocal tract physiology [7], and then matches L1 and L2 frames based on their acoustic similarity. In a final step, the method learns a mapping from L1 articulators (those associated to each L1 frame) to L2 acoustics (those of the matching L2 frame). We compare the proposed method against a baseline technique (see Fig. 1a) in terms of perceived accent, voice quality and synthesis quality.

Our study differs from most of the prior research in accent conversion [3-5], which has focused on acoustic rather than articulatory modifications. To our knowledge, the only prior work on articulatory accent conversion is by Felps et al. [6]. The method proposed here overcomes two major limitations of that earlier work. First, Felps et al. used unit-selection synthesis, so the accent conversion performance was limited by the size of the L1 and L2 acoustic-articulatory corpus and the availability of native-like units in the L2 corpus. In contrast, our method uses a probabilistic model to build the forward mapping. More importantly, the method of Felps et al. required access to L2 articulators, whereas our proposed method bypasses this step by predicting L2 acoustics directly from L1 articulators.

2. RELATED WORK

In the above cited work, Felps et al. [6] used unit-selection synthesis to replace mispronounced diphones in an L2 utterance with those from an L2 corpus with similar articulators as those in a reference L1 utterance. Articulatory similarity was measured by means of z-score normalized Maeda parameters [8]. The method was able to preserve voice identity since mispronounced units were replaced with other L2 units, but only achieved a modest 20% reduction in accent. The authors concluded that the unit-selection synthesizer lacked flexibility due to the small size of the L2 articulatory database. This suggests that physics-based articulatory

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

synthesizers [9-11] may be preferable given their added flexibility. As an example, a recent study by Toutios and Maeda [12] reported "quite natural and intelligible" VCV words from vocal tract area functions, estimated from midsagittal articulatory positions from electromagnetic articulography (EMA). However, the synthesized speech did not have the voice quality of the speaker from which the articulatory data had been collected.

Statistical articulatory synthesizers provide a practical tradeoff between unit-selection and physics-based models. Along these lines, Toda et al. [13] used Gaussian mixture models (GMM) to estimate acoustic parameters (MFCCs) from articulatory parameters (seven EMA positions, pitch and loudness). To test the capabilities of the model, the authors manipulated EMA positions to simulate the effect of speaking with the mouth wide open: the synthesized utterance had the expected loss of high frequency components in fricatives. Similarly, Ling et al. [14] showed the feasibility of modifying vowels by manipulating articulatory parameters in a HMM synthesizer [15]. Increasing the tongue-height parameters led to a clear shift in vowel perception from [ɛ] to [ɪ]. Likewise, decreasing tongue-height led to a shift from [ɛ] to [æ]. In more recent work [16], the authors were also able to synthesize vowels unknown to the synthesis model during training.

A major drawback of articulatory-based modification is the expensive process of recording articulators. In recent work [17], we tested whether articulatory positions predicted from acoustics (i.e., via articulatory inversion) could be used for speech synthesis. For this purpose, we used the GMM-based forward mapping of Toda et al. [13, 18]. Replacing measured Maeda parameters with the inverted articulators (i.e., predicted from acoustics) reduced synthesis quality. However, re-training the GMM on inverted articulators produced synthesized speech of higher quality (a reduction in Mel Cepstral distortion of 12%) than that obtained with the measured articulators. Encouraged by this result, in this paper we seek to achieve articulatory accent conversion without the need to measure articulatory data from the L2 speaker.

3. METHODS

Both articulatory methods considered in this study, the baseline method in Fig. 2a and the cross-speaker forward mapping in Fig. 2b, are based on the probabilistic articulatory synthesizer described in [17]. We will review this earlier model first, and then describe how we adapt it for the purposes of accent conversion.

3.1 Probabilistic articulatory synthesis

Consider a speech corpus containing articulatory feature vectors $\boldsymbol{x_t}$, as measured via EMA; and the corresponding acoustic feature vectors $\boldsymbol{Y_t} = [\boldsymbol{y_t}, \boldsymbol{\Delta y_t}]$ (Mel Frequency Cepstral Coefficients, $MFCCs$ and their delta value), at frame t. Our articulatory synthesizer follows the GMM-based forward mapping with global variance of Toda et al. [13, 18]. In a first step, we model the joint distribution of articulatory and acoustic feature vectors $\boldsymbol{Z_t} = [\boldsymbol{x_t}, \boldsymbol{Y_t}]$ with a Gaussian mixture model (GMM) as:

$$P(\boldsymbol{Z_t}|\lambda^{(z)}) = \sum_{m=1}^{M} \alpha_m \, \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \qquad (1)$$

where $\boldsymbol{\lambda}^{(z)} = \left\{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right\}$ are the weight, mean and covariance of the individual $(m = 1,2,..,M)$ Gaussian mixture components.

In a second step, we model the global variance (GV) of predicted acoustics to account for over-smoothing effects of the GMM. Consider the within-sentence variance of the d^{th} acoustic feature $y_t(d)$, given by $v(d) = E[(y_t(d) - E[y_t(d)])^2]$. The GV of this feature can be written as $\boldsymbol{v(y)} = [\, v(1), v(2) \ldots v(D)]$, where D is the dimension of $\boldsymbol{y}_t$ and $\boldsymbol{y}$ is the sequence $[\boldsymbol{y}_1, \boldsymbol{y}_2 \ldots]$ of acoustic vectors in an utterance. We model the distribution of GVs for all the utterances in the training set, $P(\boldsymbol{v(y)}|\boldsymbol{\lambda}^{(v)})$ with a single Gaussian $\mathcal{N}(\boldsymbol{v(y)}; \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)})$.

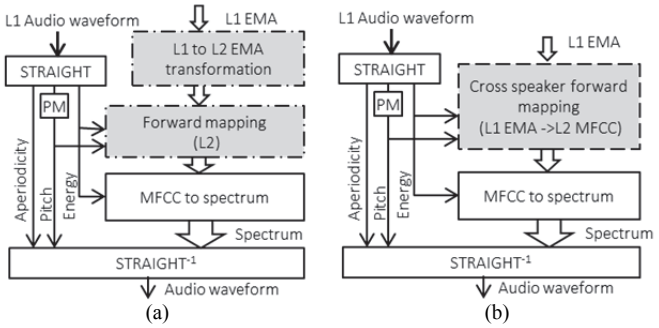

Fig. 2. (a) Baseline articulatory accent conversion system. (b) The proposed articulatory accent conversion with a cross speaker forward mapping. (PM: pitch modification, see [18]).

At synthesis time, given the trained models $[\boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}]$ and a new sequence of articulatory vectors $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \ldots \boldsymbol{x_T}]$, we obtain the maximum-likelihood acoustic (static only) trajectory $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}|\boldsymbol{\lambda}) = \underset{\boldsymbol{y}}{\text{argmax}} \; P(\boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)})^{\frac{1}{2T}} . \; P(\boldsymbol{v(y)}|\boldsymbol{\lambda}^{(v)}) \qquad (2)$$

where $\boldsymbol{Y} = [\boldsymbol{y_1}, \boldsymbol{\Delta y_1}, \boldsymbol{y_2}, \boldsymbol{\Delta y_2}, \ldots \boldsymbol{y_t}, \boldsymbol{\Delta y_t}]$ is the time sequence of acoustic vectors (both static and the dynamic) and $\boldsymbol{v(y)}$is the variance of static acoustic feature vectors. Following [18] we solve eq. (2) via Expectation Maximization.

3.2 Articulatory normalization-synthesis (baseline method)

Given a parallel corpus[1] of acoustic-articulatory recordings for both speakers, $\{\boldsymbol{Z^{(L1)}}, \boldsymbol{Z^{(L2)}}\} = \left([\boldsymbol{x^{(L1)}}, \boldsymbol{Y^{(L1)}}], [\boldsymbol{x^{(L2)}}, \boldsymbol{Y^{(L2)}}]\right)$, we generate accent conversions as follows. First, we build a forward mapping for the L2 speaker $f_2: \boldsymbol{x^{(L2)}} \rightarrow \boldsymbol{y^{(L2)}}$ via eq. (2). Then, we drive the model with articulatory trajectories from the L1 speaker. To account for differences in vocal tract and EMA pellet placement, we transform L1 articulators into the equivalent positions for the L2 speaker using a pellet-specific second-order polynomial transformation[2]. Namely, given the coordinates (a_x, a_y) for pellet $'A'$ in the L1 frame, we estimate the equivalent position $(\hat{a}_x, \hat{a}_y)$ of that pellet in L2 as:

$$[\hat{a}_x, \hat{a}_y] = [1 \quad a_x \quad a_y \quad a_x \times a_y \quad a_x^2 \quad a_y^2] \times \boldsymbol{V} \qquad (3)$$

where $\boldsymbol{V}$ is a pellet-specific 6×2 matrix, optimized to minimize the mean square error between the estimated and the actual L2 pellet positions in the parallel corpus $(\boldsymbol{x^{(L2)}}, \boldsymbol{x^{(L1)}})$

The complete synthesis pathway is summarized in Fig. 2(a). L1 articulators are mapped into the L2 articulatory space via the normalization step in eq. (3) and then converted into MFCCs via the forward mapping in eq. (2). To generate a pitch contour, we extract the L1 pitch trajectory with STRAIGHT [19], and normalize it to match the pitch range of L2; see [18]. In a final step, the predicted MFCCs are combined with the normalized pitch, L1 energy and L1 aperiodicity to resynthesize an accent-converted

[1] Parallel utterances are aligned at the frame level, i.e., by minimizing Mel Cepstral distortion via dynamic time warping.

[2] In our experiments, a second-order model worked significantly better than a linear transformation and comparably to higher-order models

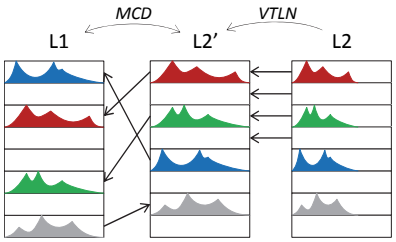

Fig. 3. Pairing L1 and L2 frames based on normalized acoustic similarity; MCD: Mel Cepstral distortion.

utterance via STRAIGHT [19]; refer to [17] for additional details.

3.3 Cross-speaker forward mapping (proposed method)

The above baseline method is impractical for CAPT because it requires measuring articulatory trajectories for each L2 learner. The method proposed in this section overcomes this issue by mapping L1 articulatory trajectories directly into L2 acoustics. As shown in Fig. 2(b), this amounts to replacing the cross-speaker articulatory normalization $g_{12}: \boldsymbol{x}^{(L1)} \rightarrow \boldsymbol{x}^{(L2)}$ and the L2 forward mapping $f_2: \boldsymbol{x}^{(L2)} \rightarrow \boldsymbol{y}^{(L2)}$ with a single cross-speaker forward mapping (CSFM) $f_{12}: \boldsymbol{x}^{(L1)} \rightarrow \boldsymbol{y}^{(L2)}$.

Generating a CSFM requires a lookup table of L1 articulatory and L2 acoustic pairs $\{\boldsymbol{x}^{(L1)}, \boldsymbol{y}^{(L2)}\}$ from which to train the GMM in eqs. (1) and (2). One may be tempted to obtain these pairs by time-aligning L1 and L2 parallel utterances in the acoustic domain (i.e., via dynamic time warping). Unfortunately, doing so would cause the CSFM to capture the non-native gestures of L2. To avoid this issue, here we propose a method that generates pairs of L1-L2 frames based on their normalized acoustic similarity. The overall process is illustrated in Fig. 3. Namely, we apply vocal tract length normalization (VTLN) to the L2 acoustic vectors $\boldsymbol{y}^{(L2)}$ to account for physiological differences in the vocal tract of both speakers. Following Panchapagesan and Alwan [7], we perform VTLN via a linear transform between the MFCCs of both speakers:

$$\boldsymbol{W} = \arg\min \left\| \boldsymbol{y}^{(L1)} - \boldsymbol{W}.\boldsymbol{y}^{(L2)} \right\|^2 \qquad (4)$$

where $\boldsymbol{y}^{(L1)}$ and $\boldsymbol{y}^{(L2)}$ are the acoustic vectors for L1 and L2, respectively, and $\boldsymbol{W}$ is the VTLN transform[3]. Next, for each vector $\boldsymbol{y}^{(L1)}$ we find its closest vector $\boldsymbol{y}^{(L2)^*}$ to minimize Mel-cepstral distortion (MCD) between them as:

$$\boldsymbol{y}^{(L2)^*} = \arg\min_{\forall \boldsymbol{y}_2} \left\| \boldsymbol{y}^{(L1)} - \boldsymbol{W}.\boldsymbol{y}^{(L2)} \right\|^2 \qquad (5)$$

We repeat the same overall process to associate each frame $\boldsymbol{y}^{(L2)}$ with the closest L1 frame $\boldsymbol{y}^{(L1)^*}$; this ensures that the CSFM covers the entire acoustic space for both speakers. This matching procedure leads to a lookup table of linguistically similar cross-speaker acoustic pairs $\{\boldsymbol{y}^{(L1)}, \boldsymbol{y}^{(L2)^*}\}$. In a final step, we discard pairs whose acoustic distance $\left\| \boldsymbol{y}^{(L1)} - W\boldsymbol{y}^{(L2)^*} \right\|^2$ is greater than a preselected threshold. The final cross-speaker articulatory-acoustic lookup table $\{\boldsymbol{x}^{(L1)}, \boldsymbol{y}^{(L2)^*}\}$ is obtained by replacing each $\boldsymbol{y}^{(L1)}$ with its corresponding articulatory vector $\boldsymbol{x}^{(L1)}$. It is this table that we use to train the CSFM $f_{12}: \boldsymbol{x}^{(L1)} \rightarrow \boldsymbol{y}^{(L2)}$ as described in §3.1.

[3] Note that generating this transform requires a lookup table of acoustic vectors $\{\boldsymbol{y}^{(L1)}, \boldsymbol{y}^{(L2)}\}$, which we obtain by time aligning the L1 and L2 parallel utterances. As in footnote [2], using a linear function $\boldsymbol{W}$ prevents the VTLN transform from capturing the non-native spectral cues in $\boldsymbol{y}^{(L2)}$.

4. EXPERIMENTAL SETUP

We tested the CSFM and the baseline method on a corpus of parallel recordings from a native speaker of American English and a Spanish non-native speaker. Both subjects recorded 344 phonetically-balanced sentences from the Glasgow Herald corpus [6].We reserved 50 sentences for testing purposes, and used the remaining 294 sentences to train the models. For each sentence, we extracted 12 articulatory features (2D positions for the upper lip, lower lip, lower jaw, tongue tip, tongue body and tongue dorsum) and 26 acoustic features (25 MFCCs, and pitch) at 200Hz. MFCCs and pitch were extracted from the STRAIGHT analysis. Since velum position was not available, we used the text transcription to generate a binary feature that represented nasality. In summary, the articulatory feature $\boldsymbol{x}$ consisted of 2D positions for six EMA pellets, log-pitch, nasality and loudness ($MFCC_0$), and the acoustic feature $\boldsymbol{y}$ consisted of $MFCC_{1-24}$ along with their delta values.

For each of the fifty test sentences, we generated five different transformations[4]:

- **CSFM**: the proposed method, as shown in Fig. 2b
- **Base**: the baseline articulatory synthesis method in Fig. 2a
- **L1**: original L1 utterances, resynthesized[5] from MFCCs
- **L2**: original L2 utterances, resynthesized from MFCCs
- **L1n**: original L1 utterances, normalized[6] to the vocal tract length and pitch range of L2

We evaluated the CSFM against the other four conditions through three listening tests: (1) an accent perception test, to measure whether CSFM generates native-like utterances, (2) a speaker identity test, to measure whether CSFM retains the identity of the L2 speaker, and (3) an acoustic quality test, to measure distortions introduced during resynthesis. Participants[7] were recruited from Mechanical Turk. Following our prior work [5, 6, 17], participants were required to reside in the US and were also screened through a qualification test that required identifying regional American accents. In the quality and accent evaluation, participants were also required to transcribe each utterance to ensure they paid enough attention to the recordings. Incomplete or grossly incorrect responses were excluded from the study.

5. RESULTS

5.1 Accent perception

Sixteen participants listened to 36 pairs[8] of utterances (CSFM-Base, CSFM-L2 and Base-L2) and were asked to (1) select the most native-like utterance of each pair, and (2) rate how confident they were in their selection in a 7-point Likert scale. Presentation order was randomized for conditions within each pair and for pairs of conditions. Participants' choice and confidence ratings were combined into a *preference score* rating from +7 (extremely confident that the first utterance in the pair is more native like) to

[4] Samples may be found in http://psi.cse.tamu.edu/samples/csfm.html

[5] Following [17], we resynthesized the original L1 and L2 utterances from their MFCCs to account for quality losses in the articulatory synthesis that are due to the MFCC compression step in Fig. 2.

[6] This fifth condition was a re-synthesis of L1 utterances in the guise of the L2 speaker, and was included to test whether a simple speaker normalization could be used to achieve accent conversions

[7] 60 participants were recruited for the experiments (20 per test); a few were removed from the final results due to their incomplete/incorrect answers.

[8] Of the 50 test sentences we randomly selected 12 for the experiments; this was done to keep the duration of each listening test below 30 min.

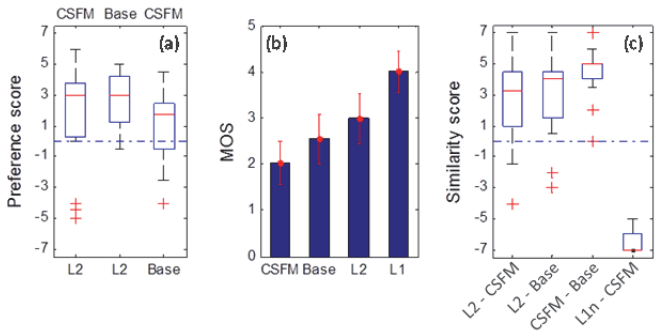

Fig. 4. (a) Median preference scores for native accentedness ("*Which utterance in the pair is more native-like*?"). (b) Mean opinion scores (MOS) for acoustic quality. (c) Median scores for voice similarity ("*Are the two utterances from the same speaker*?")

−7 (extremely confident that the second utterance is more native like). We then tested the statistical significance of this preference score using *Wilcoxon **signed-rank** test.*

As shown in Fig. 4a, participants found both CSFM and Base more native-like than L2 (CSFM-L2: median=3, Z=-2.03, p=0.04; Base-L2: median=3, Z=-3.9, p<<0.001). Moreover, differences in preference scores for (CSFM-L2) and (Base-L2) were not statistically different, as given by *Wilcoxon **rank-sum** test* (Z= -1.277, p=0.20). These results suggest that both articulatory synthesizers can achieve comparable reductions in non-native accent. Direct comparison (CSFM-base) further confirmed the conclusion (median=1.75, Z=-1.6, p=0.11), *an important finding because it suggests that accent conversion is possible without articulatory measurements for each L2 learner*.

5.2 Synthesis quality evaluation

Fifteen participants listened to the 12 test sentences under four conditions (L2, L1, Base, and CSFM) presented in random order, and rated them using a 5-point Mean Opinion Score. As shown in Fig. 4b, the CSFM condition was rated as having 'poor' quality (2.03), lower than the baseline method, which was rated between 'poor' and 'fair' (2.55); the difference was statistically significant (t=7.53, p<<0.001, df=14). Interestingly, L2 was rated as 'fair' (2.99) whereas L1 was rated 'good' (4.0), despite both conditions being equivalent in terms of recording and processing. This result suggests there is an interaction between perceived accent and quality, a finding that is consistent with our previous studies [3].

5.3 Voice identity perception

In a final test, we evaluated whether the accent-conversion transforms were able to retain the voice identity of the L2 speaker. Following [20], we presented participants with a pair of linguistically-different sentences from two experimental conditions, and then asked them to determine (1) if they were from the same speaker, and (2) how confident they were in their assessment, on a 7-point Likert scale. As before, the response and confidence levels were then combined into a *voice similarity score* (VSS) ranging from −7 (extremely confident they are different speakers) to +7 (extremely confident they are the same speaker).

Fourteen listeners rated 48 pairs of utterances, 12 pairs each from (CSFM-L2), (Base-L2), (Base-CSFM) and (CSFM-L1n) randomly interleaved. Presentation order in the pairs was also randomized within subjects. Fig. 4c shows the median voice similarity between experimental conditions. Direct comparison shows that utterances from both articulatory synthesis methods (CSFM-Base) are perceived as being from the same speaker (median=5, Z= -7.9, p<<0.001). Utterances from both methods were also perceived as being from the same speaker as L2 (CSFM-L2: median=3, Z= -3.18, p<0.001; ANS-L2: median=5, Z= -5.78, p<0.001). A *Wilcoxon **rank-sum*** test showed no difference between these two methods in terms of their voice similarity with L2 (Z=-1.13, p=0.2602). To summarize, participants were "quite confident" that both articulatory synthesizers were of the same speaker, and that this speaker was L2. In contrast, participants were "extremely confident" that L1n was different from CSFM, which shows that a simple guise cannot achieve accent conversion.

6. DISCUSSION

We have presented a data-driven articulatory synthesis method for accent conversion that does not require access to articulatory recordings from the L2 speaker. The method performs a cross-speaker forward mapping (CSFM) to predict L2 acoustics directly from L1 articulators. We compared the CSFM against a baseline method that requires L2 articulators; both methods performed comparably in terms of accent conversion accuracy and ability to preserve the voice quality of the L2 speaker.

The CSFM received a lower rating of synthesis quality than the baseline method. Given that both methods use a GMM for the forward mapping, differences in synthesis quality must be attributed to other factors. In particular, these differences may be explained by the speaker normalization step: the baseline method performs normalization in the articulatory domain via eq. (3), whereas the proposed method uses VTLN in the acoustic domain [7]. This explanation is consistent with the fact that inter-speaker differences are larger (and more complex) in the acoustic domain than in the articulatory domain [21].

Both methods received relatively low ratings of acoustic quality, a result that we attribute to the quality of the L2 corpus. To enable comparison against the baseline method (which requires L2 articulators), we used an L2 corpus that had been collected via EMA, and this interfered with proper articulation. Note, however, that CSFM does not require articulatory recording, so improved quality may be obtained by training the model on an L2 corpus recorded in ideal acoustic conditions (i.e., an audio booth).

At present, our approach uses the L1 aperiodicity and does not consider speaker individuality cues that may be present in the L2 aperiodicity [19]. 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 them; see [22]. Additional work is needed to validate the method on multiple L2 speakers from different L1 backgrounds. By avoiding the expensive process of collecting articulators from each learner, the proposed method makes articulatory-based accent conversion a viable option for CAPT. Using the articulatory manipulation methods described in Toda et al. [13] and Ling et al. [14], our approach may also be used to illustrate (to the learner) the effect that subtle articulatory differences have on acoustics observations, in this way targeting problem areas for each L2 learner.

7. ACKNOWLEDGMENTS

This work is supported by NSF award 0713205. We are grateful to Prof. Steve Renals and SICSA for their support during RGO's sabbatical stay at CSTR (University of Edinburgh).

8. REFERENCES

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

[2] Nagano, K. and Ozawa, K., "English speech training using voice conversion," in *ICSLP*, Kobe, Japan, 1990, pp. 1169-1172.

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

[4] Huckvale, M. and Yanagisawa, K., "Spoken language conversion with accent morphing," in *Proc. ISCA Speech Synthesis Workshop, Bonn, Germany*, 2007, pp. 64-70.

[5] Aryal, S., Felps, D., and Gutierrez-Osuna, R., "Foreign accent conversion through voice morphing," in *Interspeech*, 2013, pp. 3077-3081.

[6] Felps, D., Geng, C., and Gutierrez-Osuna, R., "Foreign Accent Conversion Through Concatenative Synthesis in the Articulatory Domain," *IEEE Transactions on Audio Speech and Language Processing,* vol. 20, pp. 2301-2312, Oct 2012.

[7] Panchapagesan, S. and Alwan, A., "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer speech & language,* vol. 23, pp. 42-64, 2009.

[8] Maeda, S., "An articulatory model of the tongue based on a statistical analysis," *The Journal of the Acoustical Society of America,* vol. 65, p. S22, 1979.

[9] Hanson, H. M., McGowan, R. S., Stevens, K. N., and Beaudoin, R. E., "Development of rules for controlling the HLsyn speech synthesizer," in *ICASSP*, 1999, pp. 85-88.

[10] Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America,* vol. 59, pp. 1208-1221, 1976.

[11] Birkholz, P., Jackel, D., and Kroger, B. J., "Construction And Control Of A Three-Dimensional Vocal Tract Model," in *ICASSP*, Toulouse, France, 2006, pp. 873-876.

[12] Toutios, A. and Maeda, S., "Articulatory VCV Synthesis from EMA Data," in *Interspeech*, Portland, Oregon, 2012.

[13] Toda, T., Black, A. W., and Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication,* vol. 50, pp. 215-227, 2008.

[14] Ling, Z. H., Richmond, K., Yamagishi, J., and Wang, R. H., "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing* vol. 17, pp. 1171-1185, 2009.

[15] Zen, H., Tokuda, K., and Kitamura, T., "An introduction of trajectory model into HMM-based speech synthesis," *ISCA SSW5,* pp. 191-196, 2004.

[16] Ling, Z.-H., Richmond, K., and Yamagishi, J., "Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis," in *Interspeech*, 2012, pp. 991-994.

[17] Aryal, S. and Gutierrez-Osuna, R., "Articulatory inversion and synthesis: towards articulatory-based modification of speech," in *ICASSP*, 2013, pp. 7952-7956.

[18] Toda, T., Black, A. W., and Tokuda, K., "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 2222-2235, 2007.

[19] Kawahara, H., "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *ICASSP*, 1997, pp. 1303-1306.

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

[21] Arora, R. and Livescu, K., "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, Vancouver, BC, Canada, 2013, pp. 7135 - 7139.

[22] Silén, H., Helander, E., and Gabbouj, M., "Prediction of Voice Aperiodicity Based on Spectral Representations in HMM Speech Synthesis," in *Interspeech*, 2011, pp. 105-108.