Foreign accent conversion through voice morphing

Ricardo Gutierrez-Osuna and Daniel Felps

Department of Computer Science and Engineering
Texas A&M University
{rgutier,dlfelps}@cse.tamu.edu

Abstract

We present a voice morphing strategy that can be used to generate a continuum of accent transformations between a foreign speaker and a native speaker. 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 representing the spectral detail through pulse density modulation and averaging pulses in a pair-wise fashion. The technique is evaluated on parallel recordings from two ARCTIC speakers using objective measures of acoustic quality, speaker identity and foreign accent that have been recently shown to correlate with perceptual results from listening tests.

Index Terms: voice morphing, accent conversion.

1. Introduction

During the last two decades, a few studies have suggested that it would be beneficial for second language (L2) students to be able to listen to their own voices producing native-accented speech [1-3]. The rationale behind this proposal is that removing information that is only related to the teacher's voice quality makes it easier for students to perceive differences between their accented utterances and their ideal accent-free counterparts. As a step towards this goal, we have recently developed a voice-transformation technique that can be used to synthesize native-accented utterances from their foreign-accented counterpart while preserving the speaker's voice quality [4, 5]. Here we propose a morphing technique that generates a continuum of accent-conversions between the learner's productions and those of the teacher.

Morphing accent conversions may serve as a behavioral shaping strategy in computer assisted pronunciation training (CAPT). In behavioral shaping, the teacher asks the student to compare their utterances against their previous efforts rather than against a separate standard [3]. This is accomplished by keeping track of the student's "best" utterances, and using them as a reference. Using a normative reference can be detrimental in the early stages of training, when the student's utterances are very distant from the ideal pronunciation. Instead, by using a "floating" reference (i.e., one that adapts to the performance of the learner), the teacher can provide carefully graded evaluations of the learners' performance and guide them towards the ultimate goal. Likewise, morphing accent conversions during the early stages of learning may be used to produce utterances that have less ambitious prosodic and segmental goals, slowly improving the reference by incorporating the best pronunciation of the learner and higher degrees of morphing towards the teacher's productions.

Morphing techniques have been extensively used for face perception, but are challenging when applied on speech. Whereas facial landmarks are well defined and relatively easy to detect (eyes, mouth, jaw lines, etc.), spectral features in speech (i.e., formant frequencies) are difficult to measure and ill-defined in the case of unvoiced phones. Rather than use formant-tracking techniques, which are notoriously unreliable, a number of methods have been proposed to generate morphs directly from the spectra of two speakers. Slaney et al. [6] generate separate spectrograms for the pitch and broad spectral shape of a sound, and interpolate each channel separately by means of dynamic programming and harmonic alignment, respectively. Pfitzinger [7] also uses dynamic programming to find a frequency warp between two spectra, but in this case the warping is performed on the first-order derivative of the two LP spectral envelopes. Ezzat et al. [8] also use the derivative of the two (log magnitude) spectra but instead employ a technique similar to optical flow to find a correspondence between the two spectra. More recently, Shiga [9] has proposed a method where spectral envelopes are encoded as a distribution of pulses (see Figure 2). In this case, morphing can be performed by pairing individual pulses from the two spectra (according to their order) and then computing the weighted average of each pair. This results in significant time savings as compared to previous methods based on dynamic programming or optical flow.

2. Methods

2.1. Voice morphing through pulse density modulation

Our proposed method for morphing accent conversion is based on the pulse density modulation (PDM) technique of Shiga [9]. The PDM technique employs a delta-sigma modulator to convert a log spectral envelope $x(n)$, where n denotes frequency, into a pulse sequence $y(n) = PDM[x(n)]$ according to the difference equations:

$$e(n) = x(n) - v_c\, y(n-1) \tag{1}$$

$$r(n) = e(n) - r(n-1) \tag{2}$$

$$y(n) = sign\big(r(n)\big) \tag{3}$$

with initial conditions $r(1) = e(1) = x(1)$ and $y(n) = 0$; the term v_c represents the feedback gain of the delta-sigma modulator: $v_c = max(x)$. In turn, the pulse sequence $y(n)$ can be decoded back into a log spectral envelope $\hat{x}(n) = PDM^{-1}[y(n)]$ through the discrete cosine transform (DCT) as:

$$c(n) = DCT[y(n)] \tag{4}$$

$$c(n) = 0 \;\; \forall \;\; n > k \tag{5}$$

$$\hat{x}(n) = DCT^{-1}[c(n)] \times v_c \tag{6}$$

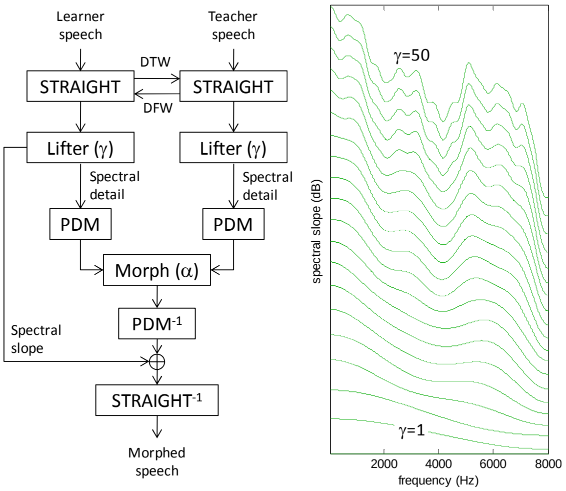

Figure 1. *(a) Block diagram of the morphing accent conversion strategy. (DTW: dynamic time warping; DFW: dynamic frequency warping; PDM: pulse density modulation). (b) Spectral slope $x^L(n)$ as a function of the liftering cutoff $\gamma \in \{1,2,3 \ldots 9,10,12 \ldots 20,25 \ldots 50\}$. The individual spectra have been shifted vertically for visualization purposes.*

which essentially acts as a low-pass filter by truncating the DCT expansion with an appropriate cutoff k ($k = 100$ in our implementation.) Thus, given a pair of spectral envelopes $x_1(n)$ and $x_2(n)$, a morphed spectral envelope can be computed by averaging the position of corresponding pulses in the two spectra:

$$x_m(n) = PDM^{-1}\big[\alpha PDM[x_1(n)] \\ + (1-\alpha)PDM[x_2(n)]\big] \tag{7}$$

where the morphing coefficient α $(0 \le \alpha \le 1)$ can be used to generate a continuum of morphs between the two spectral envelopes $x_1(n)$ and $x_2(n)$.

2.2. Accent conversion through voice morphing

Given parallel recordings from the learner and the teacher, equation (7) produces a morph of *both* the identity and the accent of the two speakers. In accent conversion, however, we seek to morph only the accent while preserving the learner's identity. For this purpose, prior to the PDM encoding in equations (1-3), each spectra $x_i(n)$ is separated into two components, $x_i^L(n)$ carrying the broad spectral features (i.e., spectral slope) and $x_i^H(n)$ carrying the spectral detail (i.e., formant positions). This, again, is performed by liftering in the DCT domain as:

$$x_i^H(n) = DCT^{-1}\big[DCT\big(x(n)\big) \times l(n)\big] \tag{8}$$

$$x_i^L(n) = DCT^{-1}\big[DCT\big(x(n)\big) \times \big(1 - l(n)\big)\big] \tag{9}$$

where $l(n)$ are the liftering coefficients, defined by:

$$l(n) = \begin{cases} n/\gamma & 1 \le n \le \gamma \\ 1 & n > \gamma \end{cases} \tag{10}$$

An accent morph $x_m(n)$ is then produced by combining the learner's broad spectra $x_1^L(n)$ with a morph of the spectral detail of both speakers $x_m^H(n)$:

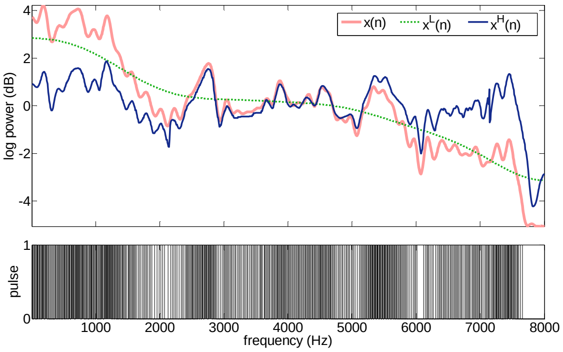

Figure 2. *(a) Decomposition of the spectral envelope $x(n)$ into its global shape $x^L(n)$ and spectral detail $x^H(n)$. (b) Encoding of the spectral detail $x^H(n)$ through pulse density modulation.*

$$x_m(n) = x_1^L(n) + x_m^H(n) \tag{11}$$

$$x_m^H(n) = PDM^{-1}\big[\alpha PDM[x_1^H(n)] \\ + (1-\alpha)PDM[x_2^H(n)]\big] \tag{12}$$

Larger values of the liftering coefficient γ in (10) ensure that increasing spectral detail is preserved in the learner's broad envelope $x_1^L(n)$ and that, likewise, equivalent spectral detail is discarded from the teacher's spectral detail $x_2^H(n)$. The overall accent-conversion process and liftering results for different values of γ are illustrated in Figure 1; morphing results for different values of α are illustrated in Figure 4.

3. Experimental validation

The proposed method was evaluated on two speakers from the ARCTIC corpus [10]: *ksp_indianmale*, who was treated as the foreign-accented learner, and *rms_usmale2*, who was treated as the native-accented teacher. The STRAIGHT vocoder [11] was used to generate smooth spectrograms and resynthesize the resulting voice morphs. Prior to performing the morphing accent conversions, learner utterances were time-aligned at the frame level (80ms windows, 1ms shift) to those of the teacher using dynamic time warping (DTW) and a conventional 39-dimensional feature vector (13 MFCCs, delta and delta-delta features) computed from the STRAIGHT spectrum. To account for differences in vocal tract length, teacher utterances were then frequency warped to those of the target; a global warping function was obtained by applying DTW in the frequency domain [12] to 100 sentences in ARCTIC's "B" set. Finally, utterances were resynthesized using the teacher's pitch contour shifted to the baseline of the learner. As a result of these steps, all subsequent morphs conformed to the timing and pitch dynamics of teacher, but had the global frequency warp and pitch baseline of the learner.

Morphing accent conversions were generated for parameter values $\gamma \in \{1,2,3 \ldots 9,10,12,14 \ldots 20,25,30 \ldots 50\}$ and $\alpha \in \{0, 0.1, 0.2 \ldots 1\}$. One hundred sentences from ARCTIC's "A" set were synthesized for each of these 11×21 combinations, and analyzed in terms of their acoustic quality, speaker identity and foreign accentedness. Three objective measures shown in our earlier work [5] to correlate with listening tests were used for this purpose. Namely, acoustic quality was estimated through the ITU-T recommendation P.563, speaker identity was estimated from a linear discriminant analysis (LDA) of natural utterances from the learner and the teacher, and foreign accent was assessed by the forced-alignment score (log-likelihood) of acoustic models trained on North American speakers using HTK's HVite tool [13]. Due to space constraints, the reader is referred to [5] for additional details on these objective measures.

4. Results

Figure 3 shows the average performance of the morphing accent conversion in terms of the three objective measures. The acoustic quality, shown in Figure 3(a), improves for higher values of the liftering cutoff γ and low values of the morphing parameter α. This result can be explained as follows. As the value of γ increases, additional spectral structure is retained for the learner's broad envelope $x_1^L(n)$. As a result, the spectral detail $x_i^H(n)$ becomes flatter for large γ, which improves the PDM encoding (i.e., for a spectrum with a significant spectral slope most of the pulses will be placed at the lower frequencies). Overall, however, the result in Figure 3(a) shows that the acoustic quality of the morphed accent conversions remains at an estimated mean-opinion-score (MOS) above 4.7, which in our earlier study [5] corresponds to a perceived MOS of 4.1.

Results from the speaker identity score are shown in Figure 3(b) in terms of the ratio:

$$ID = \frac{\sum_u \sum_i \left[d(y_{u,i}, \mu_L)/\sigma_L - d(y_{u,i}, \mu_T)/\sigma_T \right]}{d(\mu_L, \mu_T)/(\sigma_L + \sigma_T)/2} \quad (13)$$

where $d(\cdot)$ is the Euclidean metric, $y_{u,i}$ is the projection of acoustic frame i in utterance u onto the LDA solution for the two speakers, μ_L, μ_T are the average LDA projection for learner and teacher utterances, respectively, and σ_L, σ_T are their standard deviations. Thus, ID values greater than 0 indicate that the morph is closer to the learner than to the teacher, and vice versa. As shown in Figure 3(b), the morphed accent conversions remain closer to the learner except for a small number of parameter combinations (large α and small γ); the dashed line indicates the maximum-likelihood decision boundary between both speakers. These results are to be expected since for large values of α the morph is dominated by the target speaker (the teacher) and for small values of γ only the overall spectral slope of the source speaker (the learner) is preserved.

Results from the accented measure are shown in Figure 3(c) in terms of the HTK forced-alignment score:

$$ACC = \frac{\sum_u \sum_p (s_{u,p} - s_{u,sil})}{N_u N_p} \quad (14)$$

where $s_{u,p}$ is the score (log-likelihood) of phone p on utterance u, N_u is the number of test utterances and N_p is the size of the phone set $(N_u = 100;\ N_p = 39 + sil)$. Subtraction of the silence score $s_{u,sil}$ controls for misalignment errors. As may be expected, large values of the morphing parameter α reduce the foreign accentedness. In addition, the more information about the learner that is preserved in the spectral slope (i.e., by increasing the liftering cutoff γ), the larger the morphing value will have to be in order to achieve a given accent score. Comparison of Figure 3(b) and Figure 3(c) shows that the accent measure improves (i.e., morphs become more native) faster than the identity measure degrades (i.e., morphs become more like the teacher), which suggests that there is a "sweet spot" where foreign accent reduction can be achieved while preserving the identity of the learner. To illustrate the degree of change in quality, accent and identity as a function of model parameters, included with this manuscript submission are original and morphed utterances for ARTIC sentence "*His immaculate appearance was gone*" (a0059) for a progression in morphing parameter α (with fixed $\gamma = 15$) and a progression in liftering cutoff γ (with fixed $\alpha = 0.5$).

5. Discussion

We have presented a method for foreign accent conversion that combines a cepstral decomposition of the spectral envelope and a voice morphing technique through pulse density modulation. Given parallel recordings from a native speaker and a foreign speaker, we decompose the spectral envelope into its overall shape, which captures speaker-dependent cues (i.e. spectral slope), and its spectral detail, which captures linguistic content. The critical step in the morphing process is matching peaks across two spectra. We address this issue by representing the spectral detail as a distribution of a large number of pulses. In this manner, morphing two spectra is equivalent to averaging the position of their pulses in a pair-wise fashion. Results in Figure 4 show that the technique provides a smooth transition between two spectral envelopes without duplicating the number of spectral peaks; the latter would result if one morphed the two spectra $x_i^H(n)$ directly rather than their pulse densities $PDM[x_i^H(n)]$. The overall procedure contains two parameters: a liftering cutoff γ that determines the amount of information to be preserved in the foreign speaker's spectral slope, and a morphing coefficient α that determines the degree of morphing between the spectral detail of both speakers.

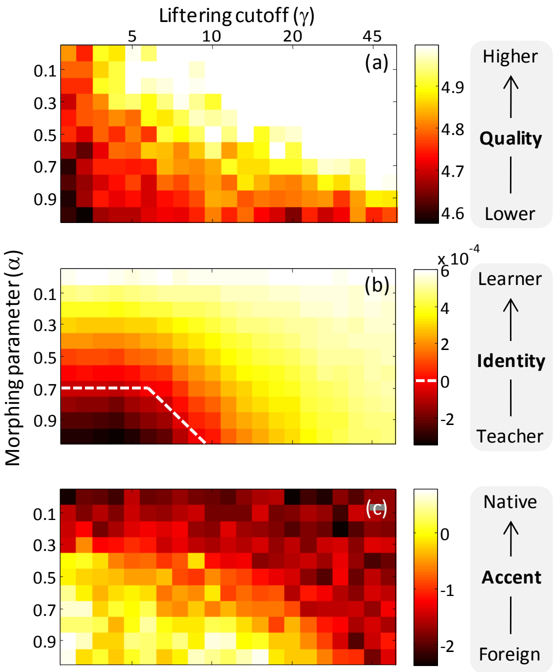

Figure 3. *(a) Quality, (b) identity, and (c) accentedness of the morphing accent conversions as a function of the liftering cutoff γ and morphing coefficient α. Lighter color denotes desirable effects (e.g., high quality, learner identity, and native accent). Dashed line in (b) represents the maximum-likelihood boundary between both speakers, as measured in the LDA subspace.*

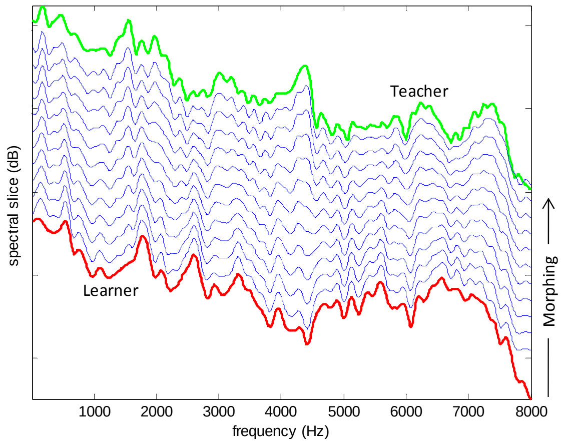

Figure 4. *Evolution of the morphed spectral envelope slice as a function of the morphing coefficient* $\alpha \in \{0, 0.1, 0.2 \dots 1\}$. *A value of* $\gamma = 1$ *was used in this case. The individual spectra have been shifted vertically for visualization purposes.*

The procedure was evaluated on parallel recordings from two speakers in the ARCTIC corpus using three objective measures (acoustic quality, speaker identity and foreign accent) that we have previously shown to correlate with perceptual results from listening tests. The results indicate that there is a trade-off between quality, identity and accent. Higher quality and identity scores are obtained by retaining as much of the learner's spectral information as possible (large γ and small α) at the expense of reducing accent scores. However, our results also show a region in parameter space where significant reductions in accent are obtained while preserving cues to the learner's identity.

Our approach preserves the fundamental frequency and overall vocal tract length of the learner, and assumes that speaker-dependent and linguistic cues in the spectral envelope can be separated through cepstral decomposition (i.e., spectral slope vs. spectral detail, respectively). While F0, vocal tract length and spectral slope are known to be good discriminator among speakers [14], additional acoustic cues from the learner's voice could be captured and preserved before the morphing stage. As an example, jitter and shimmer (cycle-to-cycle variations in fundamental frequency and amplitude, respectively) have been used to characterize various voice qualities [15], as well as fine structure in the speech signal [16]. Other features from the speaker recognition literature (see [17] for a recent review) may also be investigated while considering that our goal is synthesis rather than recognition. Future work may also investigate filtering techniques (i.e., head-related transfer functions) to reduce differences between speakers' perception of self-produced speech and their speech recordings [18], which may become important in computer assisted pronunciation training.

6. Acknowledgements

This work was supported by NSF award 0713205 and the SMART scholarship program at the Department of Defense. We are grateful to Prof. Steve Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during RGO's sabbatical stay at the Center for Speech Technology Research (University of Edinburgh). We are also grateful to Prof. Hideki Kawahara for permission to use the STRAIGHT analysis-synthesis method.

7. References

[1] M. Jilka and G. Möhler, "Intonational Foreign Accent: Speech Technology and Foreign Language Teaching," *ESCA Workshop on Speech Tech. Lang. Learn.*, pp. 115-118, 1998.

[2] A. Sundström, "Automatic prosody modification as a means for foreign language pronunciation training," in *Proceedings of ISCA Workshop on Speech Technology in Language Learning (STILL 98)* Marholmen, Sweden, 1998, pp. 49-52.

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

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

[5] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, in press.

[6] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP-96)*. vol. 2: IEEE Computer Society, 1996.

[7] H. R. Pfitzinger, "Unsupervised speech morphing between utterances of any speakers," in *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Macquarie University, Sydney, 2004, pp. 545–550.

[8] T. Ezzat, E. Meyers, J. Glass, and T. Poggio, "Morphing Spectral Envelopes Using Audio Flow," in *INTERSPEECH*, 2005, pp. 2545-2548.

[9] Y. Shiga, "Pulse Density Representation of Spectrum for Statistical Speech Processing," in *INTERSPEECH*, Brighton, UK, 2009, pp. 1771-1774.

[10] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University Language Technologies Institute 2003.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187-207, 1999.

[12] E. P. Neuburg, "Frequency warping by dynamic programming," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-88)*, 1988, pp. 573-575.

[13] S. J. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," Department of Engineering, Cambridge University 1993.

[14] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 554-568, 1999.

[15] M. Farrus and J. Hernando, "Using jitter and shimmer in speaker verification," *IET Signal Processing*, vol. 3, pp. 247-257, 2009.

[16] C. R. Jankowski, Jr., T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: application to speaker identification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, 1995, pp. 325-328 vol.1.

[17] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.

[18] L. I. Shuster and J. D. Durrant, "Toward a better understanding of the perception of self-produced speech," *Journal of Communication Disorders*, vol. 36, pp. 1-11, 2003.