FEEDBACK UTTERANCES FOR COMPUTER-ADIED LANGUAGE LEARNING USING ACCENT REDUCTION AND VOICE CONVERSION METHOD

Sixuan Zhao[1], Soo Ngee Koh[1], Soon Ing Yann[1], Kang Kwong Luke[2]

[1] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore
[2] School of Humanities & Social Sciences, Nanyang Technological University, Singapore

ABSTRACT

This paper considers the generation of feedback utterances for speaking skills training of non-native English learners. The proposed feedback is in the form of a combination of the learner's voice and the linguistic gestures, i.e., the prosody or pronunciation, of a native speaker. Both accent reduction method and voice conversion method are employed to generate feedback stimuli. For accent reduction, three speech synthesis methods, namely pitch-synchronous overlap and add (PSOLA), harmonic stochastic model (HSM), and speech transformation and representation by adaptive interpolation of weighted spectrogram (STRAIGHT) are used to reduce the accent of the utterances of English learners. For voice conversion, the teacher's voice is converted to that of the learner and the converted speech is used as a feedback. Objective measurements are employed to assess the nativeness and acoustic quality of the generated stimuli. A feedback scheme which combines the accent reduction and voice conversion methods is also proposed.

Index Terms: CALL, feedback utterances, accent reduction, voice conversion

1. INTRODUCTION

With the advent of advanced speech & language processing technologies, computer-aided language learning (CALL) is playing an increasingly important role in second language (L2) English learning. The traditional approach for speaking skills training in a CALL system requires the learner to repeat a sentence pronounced by a native speaker. However, two observations motivate the development of a more effective system. First, the learning efficiency may be reduced by the dissimilarity between the voice features of the learner and those of the native speaker [1]. Second, based on the linguistic study in [2], the training process could be more effective if the system can generate reference stimuli for the learner by taking into account his English proficiency.

A "golden speaker" as suggested in [1] is defined as the native speaker who can offer the most appropriate feedback to L2 learners. The "golden speaker" possesses voice features which are closest to that of the learner, enabling the learner to focus on pronunciation and prosody issues. One problem of this approach stems from the difficulty in providing a "golden speaker" for every user of the system. A possible solution is to generate stimuli by combining the learner's voice and the teacher's linguistic gestures. It has been shown that the perceived accent of non-native speech can be reduced after correction of prosody or pronunciation, as reported in [3, 4]. Furthermore, a pedagogical study in [5] suggests that prosody-corrected speech of the L2 learners is more effective for the learners compared to pre-recorded native speech. Therefore, as proposed in [3, 6, 7], it could be beneficial for the non-native learners to listen to their own accent-corrected utterances rather than to follow the teacher's utterances. Methods to obtain the learner's own accent-reduced speech are called "accent reduction" or "accent conversion". With the playback of the accent reduced utterances, the learner can then identify the deficiency of his pronunciation and focus on improving his speaking skills.

Some work has been done on accent reduction [4, 6, 8], but each of these studies has certain limitations, e.g., limited experimental corpus [4], use of synthesized speech rather than natural speech [6], or reliance on a large database which is difficult to be obtained [8]. Furthermore, all these papers are based on the same synthesis method (PSOLA) which has some limitations in producing feedback utterances with a desired quality.

As the proposed feedback utterances are to take the form of a combination of the correct prosody or pronunciation of the teacher and the voice features of the learner, an alternative method to achieve the desired feedback is to convert the voice features of the teacher's speech to those of the learner by using a voice conversion (VC) method [9]. Although voice conversion techniques have been in existence for years, such an application has not been proposed and studied in the past.

This paper will focus on accent reduction using different synthesis methods, as well as using the VC technique to generate feedback. While the presented study is related to previous works [4, 6, 8] by using accent reduction to generate feedback utterances for language learners, it considers and compares different synthesis techniques to find the optimal choice. In addition, the VC technique is introduced to generate feedback utterances, which is an application not explored before.

2. ACCENT REDUCTION METHOD

To reduce the accentedness of the learner's speech, three different speech synthesis methods are employed and compared: PSOLA method, HSM method and STRAIGHT method. As stated in [5, 7, 10], prosodic features play the main role of the perceived

accentedness in the non-native speech. Therefore, this paper is primarily focused on the modification of prosodic features. The overall modification scheme is shown in Fig. 1:

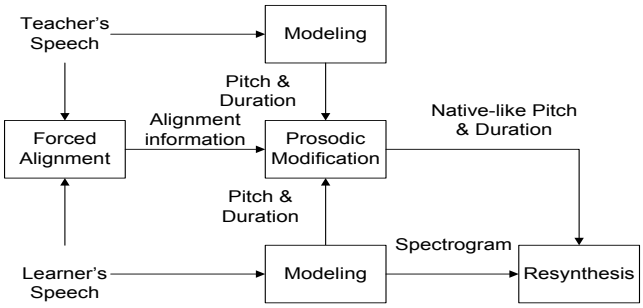

Fig.1: *Scheme of Prosodic Modification*

The first step is to obtain the phonetic time alignment. This process is done by forced alignment using HTK [11] and acoustic models trained on WSJ [12]. The acoustic models can generate an overall segmentation accuracy of 92% with 30 ms tolerance for the TIMIT database. Since accent reduction modifies pitch and duration of continuous utterances, it is not as sensitive to small alignment errors as a concatenative text-to-speech (TTS) system which concatenates individual phoneme units. Hence, the forced alignment results can fulfill the requirement of accent reduction.

With obtained phone-level time alignments, duration modification can then be performed. The time-scale modification ratio α of each phoneme is calculated by dividing the teacher's phoneme duration by that of the learner. The ratio is constrained to the range of [0.25 4]. Each phoneme of the learner is stretched or compressed by the ratio α to approximate the duration of the teacher's corresponding phoneme.

Subsequently, pitch modification is performed. The log pitch contour of each of the learner's phoneme is linearly interpolated to have the same length as that of the teacher, so that the modified phoneme will have the same frames as that of the teacher. Thus, a frame level mapping can be found from each frame of the interpolated learner's phoneme to that of the teacher's phoneme. Suppose the learner's log pitch contour to be replaced is denoted as $P^L(t)$ and the aligned teacher's one in the same phoneme is $\psi(P^T(t))$, with $\overline{P^T(t)}$ and $\overline{P^L(t)}$ as the mean log pitch values, then the pitch-scale modification factor is:

$$\beta = \frac{\psi(P^T(t)) - \overline{P^T(t)} + \overline{P^L(t)}}{P^L(t)} \quad (1)$$

The pitch modification factor is time-varying and calculated in a frame-by-frame basis. The same scaling factors are applied to all the three synthesis methods. Brief introductions of the three synthesis methods are given as follows:

PSOLA can be categorized as a waveform method, which modifies pitch and duration by directly manipulating speech waveform, as stated in [13].

HSM [9] is based on harmonics and noise components:

$$s^{(k)}[n] = \sum_j A_j^{(k)} \cos(jw_0 n + \varphi_j^{(k)}) + \sigma[n] * h_{LPC}^{(k)}[n] \quad (2)$$

where j is the number of harmonics, k is the frame number, A_j is the amplitude, w_0 is fundamental frequency, φ_j is the phase, h_{LPC} is the LPC filter of residuals, and σ is white noise.

STRAIGHT as proposed in [14] is a high-quality speech synthesis method which can be viewed as an advanced version of a phase vocoder:

$$x(t) = \sum_k h_k(t) * s_k(t) \quad (3)$$

where h_k denotes the excitation generated from the new pitch contour which is calculated from the scaling factors; s_k denotes the spectrogram of the learner's utterances after the pitch periodicity is removed; and k is the frame index. The high quality of STRAIGHT basically comes from two aspects -- the elimination of pitch mark detection in the analysis process and the use of group delay all-pass filters adopted in synthesis process for the fine control of pitch and excitation signals, as described in [14].

3. VOICE CONVERSION METHOD

An alternative way to generate desirable feedback utterances stems from voice conversion method. Voice conversion deals with voice features which are correlated to speaker identities, leaving prosody and pronunciation unchanged. Therefore, by performing voice conversion on the teacher's utterances to transform the voice to that of the learner, the output utterances will possess both the teacher's linguistic gestures and the learner's voice.

The voice conversion method used in this paper is mainly based on [15], which proposes high quality voice conversion based on Gaussian mixture models (GMMs). Line spectral frequencies (LSF) generated by HSM parameters as in [9] are used as the feature vector. In this paper, GMMs with 16 mixtures are used and the feature vector of converted speech is given as follows:

$$y_n = F(x_n) = \sum_{i=1}^{M} P(\theta_i \mid x_n)\left[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x_n - \mu_i^x)\right] \quad (4)$$

where x_n is the input feature vector, $P(\theta_i|x_n)$ is the probability of belonging to i-th mixture given x_n, μ_i^x and μ_i^y are means of GMM, Σ_i^{xx} and Σ_i^{yx} are covariance matrices of GMM, M is the number of mixtures, and y_n is the converted feature vector.

4. OBJECTIVE MEASUREMENT OF FEEDBACK UTTERANCES

Experiments are performed on all the generated stimuli in terms of accentedness and acoustic quality. The experimental corpus contains 40 teachers' utterances (as the reference speech for accent reduction or the source speakers for voice conversion) and 183 non-native speakers' utterances. All of those students' utterances are recorded in a quiet lab which is not a sound-proof room to simulate the real usage environment of a CALL system. The transcriptions are selected from the Boston University Radio News Corpus (BURNC). Experimental conditions are listed in Table 1.

Table 1. *Experimental Conditions*

Database	BURNC
Recording Conditions	16 kHz, 16 bit, in a quiet lab which is not a sound-proof room
Transcriptions	20 unique sentences from BURNC
Learners' Utterances	Total of 183 utterances from 10 students in Singapore (Chinese, Indian, Vietnamese and Singaporean)
Teachers' Utterance	Speaker M1B and F2B in BURNC with selected transcriptions

As is known, the mispronunciations of non-native speakers can negatively influence the forced alignment results. Therefore, 17 sentences (out of the 200 original recordings) which contain mispronunciations (deletion, insertion, substitution) are eliminated from the experiment, leaving 183 learners' utterances with only stress- and prosody-related issues for experiment. Objective measurements of accentedness and acoustic quality, with reported human-machine correlation of over 0.8, are performed as proposed in [16]. The posterior score as suggested in [17] replaces the likelihood score in [16] to give a more accurate evaluation.

4.1 Accentedness Measurements

Posterior score generated by HTK is the benchmark used to measure accentedness of an utterance. The output posterior probability score shows the normalized probability that a speech segment is correctly pronounced, i.e., corresponding to the correct acoustic model trained using native speech. Hence, the deviation of the input speech from native speech (the standard norm, i.e., American English used here) which is used to train acoustic models can be measured by the posterior score. The definition of sentence level score is given by:

$$S_{accent} = -mean\{\log \frac{p(o_j \mid \lambda_j)}{p(o_j \mid \lambda_{\max})}; j=1,2,...,n\} \quad (5)$$

where S_{accent} is the sentence level accentedness score , o_j is the j-th observation, λ_j is the correct phoneme label of j-th observation, $\lambda_{\max}$ is the phoneme label which generates o_j with the highest probability, and n is the total number of phonemes in the sentence. A lower score indicates a higher nativeness.

The mean accentedness scores for all the stimuli groups obtained with acoustic models trained on WSJ [12] are shown in Fig. 2. These stimulus groups include original learners' utterances, original teachers' utterances, and feedback utterances generated by either modifying learners' utterances with three synthesis models or converting the voice features of teachers' utterances to those of learners using voice conversion method.

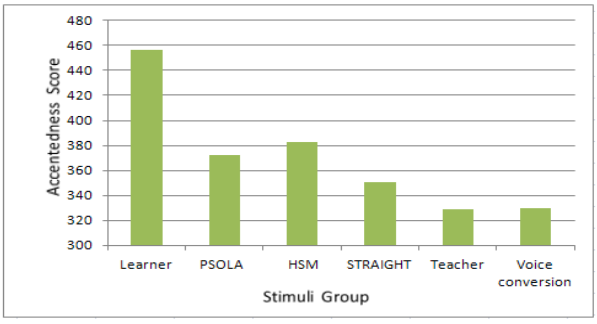

Fig. 2: *Accentedness Scores of Different Stimuli.*

As shown in Fig. 2, the difference in accentedness score between each pair of stimuli is statistically significant (t-tests show $p<0.01$), except for the pair of teachers' speech and speech generated by the voice conversion method. After prosodic modifications using three different speech synthesis models, the mean accentedness score is reduced, showing an improved nativeness. Compared to the other two methods, the STRAIGHT method achieves the lowest score, i.e., the highest nativeness. This may be due to the different working schemes of the three methods: STRAIGHT uses a new pitch contour to reconstruct the excitation for synthesis, whereas the other two models just overlap and add (PSOLA) or interpolate original speech (HSM) to change the intonation contour. Because the excitations from the STRAIGHT method are generated from the new pitch contour using minimum phase filters and not influenced by the original (learner's) pitch contour, the synthesized speech has an intonation contour which is closest to the native one, leading to a higher nativeness. In contrast, PSOLA and HMS modify the original excitations by overlapping and interpolating, the synthesized speech inevitably contains prosodic features of the learner, which increases the accentedness.

The speech generated by voice conversion method shows a low mean accentedness score which is similar to that of the original teachers' speech. This is expected as voice conversion only converts the voice, without changing prosody and pronunciation. Compared to accent reduction which improves prosodic features, stimuli generated by voice conversion also possess correct pronunciation, resulting in higher nativeness.

4.2 Acoustic Quality Measurements

Acoustic quality is assessed by the mean opinion score (MOS) generated by the ITU Standard P.563 [18]. P.563 is a single-ended method originally designed for evaluating telephone speech in terms of the naturalness of vocal tracts and background noises, which are valuable cues for assessing the accent-reduced speech as well. Fig. 3 shows the MOS of different stimuli.

In Fig. 3, t-tests show a significant difference ($p<0.01$) for each pair of stimuli. The acoustic quality of the modified speech is degraded when using accent reduction methods. However, the quality of the modified speech using STRAIGHT is the closest to that of the original learners' speech. Therefore, STRAIGHT can maintain a higher quality than the other two methods. Prosodic modification in this paper uses factors varying from phoneme to phoneme to change the intonation contours. In addition, STRAIGHT uses new generated excitations for speech synthesis and thus avoids the interferences as introduced by overlapping and adding in PSOLA or interpolation in HSM. In addition, the group-delay manipulation used in STRAIGHT, which enables finer pitch and excitation signal control by using phase interpolation, also contributes to the higher acoustic quality.

The MOS of the original learners' speech is lower than the teachers' speech due to the original high acoustic quality of teachers' speech in the BURNC corpus. The speech generated by voice conversion has the lowest MOS. This is due to the spectral distortions introduced by the linear transformation on spectral features during the VC conversion process, which significantly reduces the quality of the converted speech.

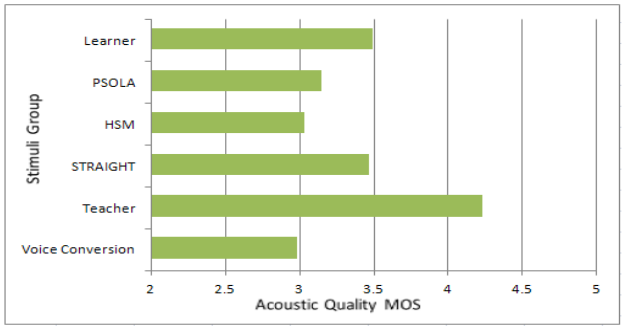

Fig. 3: *Acoustic Quality of Different Stimuli.*

5. DISCUSSIONS

From the experimental results presented in the previous section, it can be found that the STRAIGHT method is the best method for accent reduction because it has the highest acoustic quality and greatest reduction of accentedness. In addition, it seems that feedback utterances generated by voice conversion sound more native like, resulting from the correct pronunciations which are not addressed by prosodic modification. However, the lower quality of speech generated by voice conversion method is still an issue. Furthermore, the speaker identity issue, i.e., the preservation of the learner's voice in feedback utterances, should also be examined. A comparison of spectrograms, which provides useful cues about speaker identity, is shown in Fig. 4:

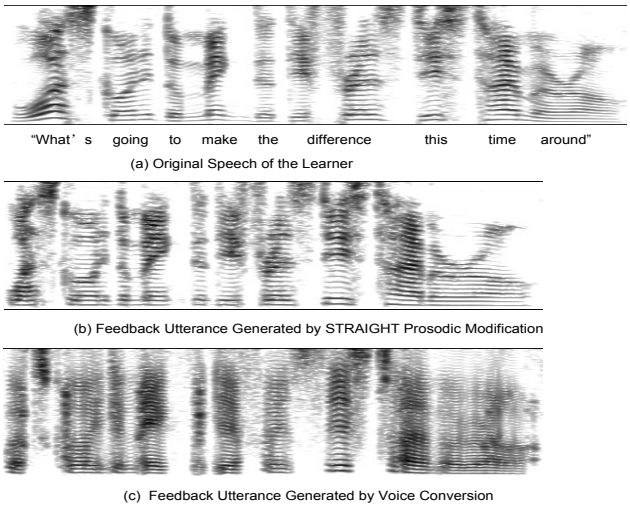

Fig. 4: *Spectrogram Comparison*

The duration of the original learner's speech is longer as the teacher speaks faster than the learner. It is obvious that the spectrogram of the accent-reduced speech using STRAIGHT is similar to that of the original one. As a result, the learner's speaker identity is fully preserved. In contrast, the converted speech shows a smoothed spectrogram that is significantly different from the learner's one. Although some of the differences may result from the different pronunciations of two speakers, the obvious over-smoothness in the spectrogram introduced by GMMs obscures the speaker identity as well. In fact, an informal subjective listening test suggests that the voice of the converted speech is in-between that of the teacher and the learner – the converted speech still keeps some of the teacher's voice features, even though the voice is more similar to that of the learner. What's more, the training corpus (from the learner) required by voice conversion method may be difficult for English beginners who cannot speak fluently.

Therefore, a feedback system which combines two methods may be desirable. At the beginning stage, the STRAIGHT based accent reduction method can be used. As the speaker identity is fully preserved and the prosody is improved, the learner can imitate the generated stimuli to improve their prosody. After a period of training, it will be possible to gather the learner's utterances to train a voice conversion system as the learner will be able to speak more fluently. Thus, voice conversion method can be used to generate feedback stimuli with not only correct prosodic features, but also standard pronunciations. The proposed scheme is shown in Fig. 5:

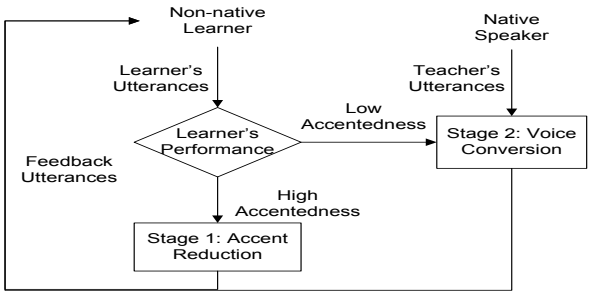

Fig. 5: *Proposed Feedback Scheme*

6. SUMMARY

This paper studies the available speech synthesis methods to determine the most suitable one which should be used for accent reduction purposes. Moreover, the proposal of using voice conversion method provides an alternative way to generate feedback utterances for English learners. Objective measurements show that the STRAIGHT method is the most suitable synthesis method for accent reduction. Feedback stimuli generated by voice conversion method possesses the highest nativeness, but it yields the lowest acoustic quality and a partial loss of the learner's speaker identity. In addition, the training corpus required by voice conversion creates a difficulty for English beginners. Therefore, a multi-stage feedback system is proposed to facilitate the learning process of non-native English learners.

In future, a pedagogical study will be designed to verify the proposed scheme and explore other possibilities to generate more informative feedback utterances.

6. ACKNOWLEDGMENT

The authors would like to acknowledge the Ph.D. grant from the Institute for Media Innovation, Nanyang Technological University, Singapore.

REFERENCE

[1] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors-In search of the golden speaker," *Speech Communication,* vol. 37, no. 3-4, pp. 161-173, 2002.

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

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

[4] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication,* vol. 51, no. 10, pp. 920-932, 2009.

[5] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in Proc. of the 11th Australian International Conference on Speech Science &

Technology, University of Auckland, New Zealand, 2006, pp. 24-29.

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

[7] M. Jilka, and G. Möhler, "Intonational foreign accent: speech technology and foreign language teaching," in Proc. ESCA Workshop on Speech Technology in Language Learning, 1998, pp. 115-118.

[8] Q. Yan, and S. Vaseghi, "Modeling and synthesis of English regional accents with pitch and duration correlates," *Computer Speech & Language,* vol. 24, no. 4, pp. 711-725, 2010.

[9] D. E. Eslava, "Intra-lingual and cross-lingual voice conversion using harmonicplus stochastic models," *PhD Thesis*, 2008.

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

[11] S. Young, G. Evermann, D. Kershaw *et al.*, "The HTK book," 1997.

[12] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments," *Cavendish Laboratory, University of Cambridge*, 2006.

[13] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication,* vol. 9, no. 5-6, pp. 453-467, 1990.

[14] H. Banno, H. Hata, M. Morise *et al.*, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoustical science and technology,* vol. 28, no. 3, pp. 140-146, 2007.

[15] A. B. Kain, "High resolution voice transformation," *PhD Thesis*, 2001.

[16] D. Felps, and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, no. 5, pp. 1030-1040, 2010.

[17] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication,* vol. 51, no. 10, pp. 832-844, 2009.

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