ACCENT REDUCTION FOR COMPUTER-AIDED LANGUAGE LEARNING

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

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2] School of Humanities and Social Sciences, Nanyang Technological University, Singapore
Email: zhao0120@e.ntu.edu.sg esnkoh@ntu.edu.sg kkluke@ntu.edu.sg

ABSTRACT

This paper studies accent reduction techniques which are used to provide English learners with their own converted speech as a reference for speaking skills training. Three kinds of modifications, namely prosodic modification, segmental modification and combined modification are compared and examined using objective measurements. Two different corpora, a prosody abundant corpus and a prosody flat corpus, are used in our study. Modified utterances show a clear reduction of accentedness and an acceptable acoustic quality when compared with the original speech, demonstrating the effectiveness of the proposed accent reduction techniques. Furthermore, differences between experimental results from the two corpora in terms of the reduction of accentedness are observed and explanations for this phenomenon are presented. This paper also discusses other issues in this area for future research.

Index Terms— accent reduction, foreign accent, CALL.

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 English learning. The conventional approach for speaking skills training requires the learner to repeat a sentence after the sentence uttered by a native speaker is played back to him. However, there are two issues with such a system. First, the dissimilarity between the voice features of the learner and those of the native speaker may reduce the learning efficiency as shown in [1]. Second, as illustrated in a linguistic study [2], a system which can provide a learner with reference utterances by considering his English proficiency will be more effective.

It is proposed in [1] that a "golden speaker" can be found to offer the most appropriate feedback to L2 learners. The "golden speaker" possesses the voice with the highest similarity with the learner, thus enabling the learner to concentrate on the pronunciation and prosody issues. Even if a group of "golden speakers" are included in the system, it is difficult to guarantee that a "golden speaker" for every user has been included. A number of papers [3-6] have argued that it is beneficial for language learners to listen to their own accent-corrected voices. Specifically, subjective evaluation results in [3, 6] demonstrate significant reductions of non-native speech accent after modification. Further, a pedagogical study in [5] suggests prosody-corrected speech of the learner is a more effective stimuli for L2 learners compared to pre-recorded native speech. To address this issue, some papers [3, 7] have stated that it can be beneficial for the learners to listen to their own modified voices rather than following the teacher's utterances. Those methods to obtain the learner's own accent-reduced speech as feedbacks for language learning purposes are called "accent reduction" or "accent conversion".

In this paper, for consistency, we will use the term "accent reduction". By "accent reduction", we mean the modification of foreign accents which will bring them in line with "standard norms". "Standard norms" refer to national standards of pronunciation and prosody in the U.S. or Britain. And "accent" here refers to "foreign accent" which can be defined as deviations of segmental (spectral envelope) and prosodic (phoneme duration and pitch contour) features from "standard norms" in foreign learners' speech, and in our case, these will be students in Singapore.

Some related works have been done by other researchers [6-9]. Unfortunately, each of those studies leaves some issues open to questions. In [7, 9], the accent reduction process is performed on synthesized speech rather than natural speech, which is not the best for language learning purpose. The accent reduction among American, Australian and British English reported in [8] is useful, but the problem is the involved training process. As the accent reduction in [8] is based on a thorough study of three regional English accents, it is difficult to generalize this method to other accent pairs where the English corpus from L2 learners is not available. A more detailed study is reported in [6]. It covers evaluation of accent-reduced speech in terms of voice quality and accentedness. One of the problems in that paper is the small experimental corpus (20 sentences from the same pair of speakers). Also, it only experiments on one corpus and thus omits comparison across different corpora.

This paper studies current accent reduction methods to reduce accentedness of non-native speech while retaining the learner's voice. Accent reduction methods are discussed and experimented on two groups of utterances from different corpora. The main focuses of this paper are the studies and comparisons of existing accent reduction techniques as well as the impact of various factors, e.g., reference corpus, nationality, etc.. Some refinements (e.g., selective pitch modification and spectral interpolation) of existing accent reduction method are also introduced to improve the quality of accent-reduced speech.

2. ACCENT REDUCTION TECHNIQUES

The accent reduction scheme in this paper is generally the one proposed in [6], with refinements to increase the efficiency and quality. This scheme involves the well-known source-filter model. As proposed by this model, a speech utterance can be decomposed into excitation, which is mainly responsible for the prosodic features and speaker identity, and vocal tract filter resonance, which represents the linguistic gestures of speech. Although the vocal tract filter also contains to a certain extent the features associated with speaker identity, vocal tract length normalization (VTLN) can be used to restore the learner's voice features as reported in [10]. As a result, accent reduction can be implemented by separating speech signals into two components (excitation and vocal tract resonance) and processing each component to reduce the foreign accents. This paper uses the linear predicative pitch synchronous overlapping and adding (LP-PSOLA) [11] rather than the frequency domain PSOLA (FD-PSOLA) as proposed in [6], because LP-PSOLA possesses a much higher efficiency, which is important to a real-time CALL system.

In the modification process, parallel speech signals from two speakers (the teacher and the learner) are needed. Speech signals from both speakers are first decomposed into pitch-synchronous frames. The 20-order liner predicative coefficients (LPCs) are used to obtain the vocal tract filters. After decomposition, each frame is filtered by the inverse LPCs to obtain excitations. Subsequently, modifications are performed on the excitation to change the prosodic features and the LPCs of the learner are substituted by that of the teacher to obtain the desired vocal tract features. Finally, the modified excitation and vocal tract filters are combined to synthesize the accented-reduced speech. .

.

2.1 Prosodic Modification

In prosodic modification, phoneme duration and pitch contour of the learner's utterance are modified according to the teacher's utterances. Energy in speech on the other hand is mainly associated with spectral envelopes, which are addressed by segmental modification with PSOLA model.

The first step of duration modification is to obtain the phoneme durations of each speaker. Phoneme durations can be acquired by forced alignment based on speech recognition software HTK. Following that, the time-scale modification ratio α of each phoneme is calculated by dividing the teacher's phoneme length by that of the learner. The ratio α is constrained to the range of [0.25 4].

The pitch scale modification is implemented by a selective replacement of pitch contours. Shapes of pitch contour of each pair of phonemes are detected as proposed by [12]with four pitch types: H*L, L*H, H*LH and L*HL. If the shapes of two pitch contours are not the same, then pitch contour of the learner will be substituted by that of the teacher. Otherwise, no modification will be performed. This method only modifies the learner's pitch contours whose shapes are different from that of the teacher to minimize distortions. Supposing the learner's pitch contour to be replaced is considered as $P^L(t)$ and the time 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 pitch values, the modification factor for each frame to substitute the pitch contour is obtained by:

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

with the pitch modification scale limited to [0.5 2].

2.2 Segmental Modification

The general idea of segmental modification is to replace the vocal tract filter of each frame of the learner by that of the teacher. Three steps are involved in this process.

The first step is the alignment between two speakers' vocal tract features. As the number of frames of two speakers can be quite different, it is necessary to align frame pairs from the teacher and the learner. Different from [6] which only uses a simple linear-piecewise alignment process based on phoneme boundaries, dynamic time warping (DTW) is also performed to align the frame pairs from two speakers. The line spectral frequencies (LSF) based alignment is adopted for its high acoustic quality.

The second step is the vocal tract (or spectral envelope) substitution. After the alignment, the LPC filters of each frame of the learner should be replaced by the aligned frame of the teacher. Unlike [6, 7] which combine the modified frames directly, spectral interpolation based on LSF [13] is used to smooth the phoneme boundaries. The interpolation of spectral envelope is illustrated in Fig. 1:

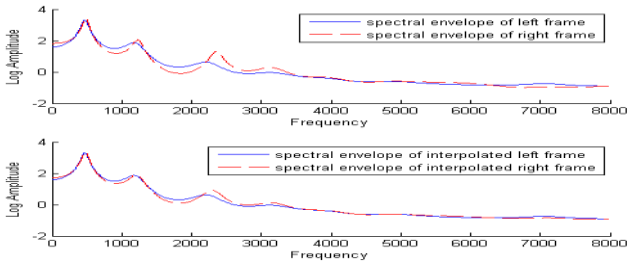

Fig. 1: Interpolation of Spectral Envelopes

As shown in Fig. 1, the spectral envelopes of two neighboring frames at phoneme boundaries are closer to each other after interpolation, thus reducing discontinuities.

The last step is the normalization of vocal tract. Before combining the modified excitations and the vocal tract filters, a linear-piecewise vocal tract length normalization (VTLN) as in [10] is performed to restore the learner's speaker identity. Finally, all of the modified frames are combined in the way as proposed in [11].

3. OBJECTIVE MEASUREMENT OF CONVERTED SPEECH

To test the proposed accent reduction methods, experiments are performed on two corpora in terms of accentedness (the deviation from the standard norm as stated in the previous section) and acoustic quality. The first corpus consists of 200 students' utterances and 40 teachers' utterances (as references for conversion). All the transcriptions are selected from Boston University Radio News Corpus (BURNC), a prosody abundant corpus. 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 details are listed in Table 1.

Table 1. *Experimental Conditions*

Database	BURNC
Recording Conditions	16k Hz, 16 bit, quiet lab which is not a sound-proof room
Transcriptions	20 unique sentences selected from BURNC
Students' Utterances	Total of 200 utterances from 10 students in Singapore (Chinese, Indian, Vietnamese and Singaporean)
Teachers' Utterance	Speaker M1B and F2B in BURNC with selected transcriptions

To study the impact of prosodic and segmental modification on different corpora, the second corpus based on CMU_ARCTIC is also used for experiments. A total of 120 sentences (six times of the experiment in [6]) of one speaker pair (Indian speaker WSP and US speaker RMS) are selected as the student's and the teacher's utterances.

Objective measurements of accentedness and acoustic quality, with reported human-machine correlation over 0.8, are performed as proposed in [14]. The likelihood score in [14] is replaced by the posterior score as suggested by [15] to give a more accurate evaluation result.

3.1 Accentedness Measurement

The accentedness measurement is based on posterior score using speech recognition software HTK. The output posterior probability score indicates the normalized probability that a speech segment corresponds to the correct acoustic model trained by native speech. Hence, the posterior score measures the deviation of the input speech from native speech (the standard norm, which is American English) used to train acoustic models. As pitch, duration and spectral envelope modification correspond to the stress, duration and pronunciation of phonemes which model the acoustic model, the output probability will be affected by those modifications. The sentence level score is defined as:

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

where S_{accent} is the sentence level accentedness score , o_j is the observation of *j*-th phoneme, λ_j is the correct label of *j*-th phoneme, $\lambda_{\max}$ 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 lower accentedness, i.e., higher nativeness, of an utterance.

The mean accentedness scores for all kinds of stimuli over 200 sentences obtained with acoustic model trained on WSJ [16] are shown in Fig. 2, with vertical axis indicating accentedness score and horizontal axis indicating different stimuli groups. The mean accentedness scores are connected by straight lines for comparison:

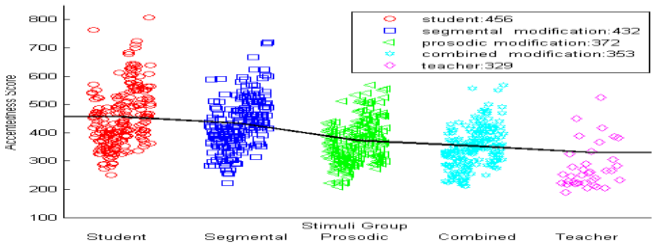

Fig. 2: Accentedness Score of Self-created Corpus

According to Fig. 2, it is obvious that the accentedness score of the utterances with prosodic modification and combined modification are lower than the student score, showing a reduction of accentedness. In contrast, accentedness scores of utterances with segmental modification are not far from those of the original students' utterances. Two-way ANOVA performed on stimuli groups show $p<0.05$ between utterances with segmental modification and original students' utterances and $p<0.001$ for all the other pairs of stimuli.

Experiments based on ARCTIC corpus, however, show a reverse performance of segmental and prosodic modification in Fig. 3:

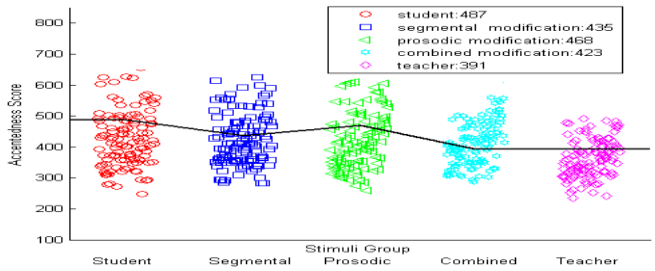

Fig. 3: Accentedness Score of ARCTIC Corpus

In the experiment with ARCTIC corpus, two-way ANOVA show significant differences for all the pairs of stimuli (p<0. 01), except for the pair between original students' utterances and utterances with prosodic modification (p=0.55). Fig. 3 shows a much more significant accent reduction from segmental modification compared with that of prosodic modification, while the reverse observation is true in the experiment with the BURNC based corpus as shown in Fig. 2. Such conflicting observations can also be found in previous papers ([4] [5] versus [6] [14]). It should be noted that multi-corpora experiments with the same conversion method are carried out in our study, which is different from the previous studies that were confined to a single corpus. The different performance of prosodic and segmental modification on the two different corpora shows that the corpus on which the accent reduction is based can affect the performance of accent reduction methods.

3.2 Acoustic Quality MOS

The ITU Standard P.563 [17] is used to assess the acoustic quality and to generate mean opinion score (MOS). As a single-ended method, P.563 is originally designed for evaluating telephone speech in terms of the naturalness of vocal tracts and background noises, which are also valuable cues for assessing the accent-reduced speech. Fig. 4 shows the MOS of the BURNC and ARCTIC based corpus:

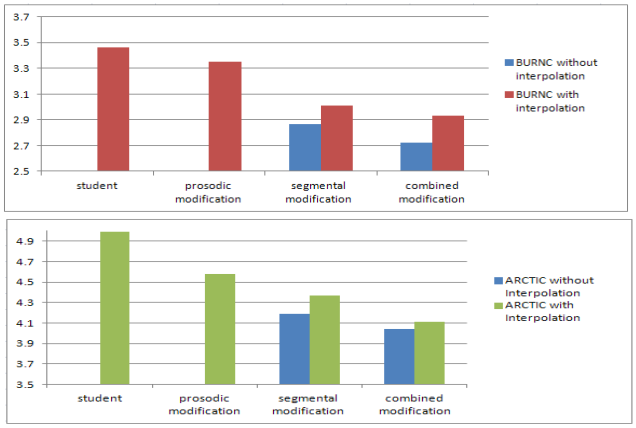

Fig. 4: Acoustic Quality MOS by P.563

The MOS of teachers' utterances are not calculated as accent reductions are only performed on students' utterances. The MOS of the ARCTIC corpus is much higher than that of the BURNC corpusdue to the high quality of the original speech. Accent reduction leads to a degradation of acoustic quality of the modified utterances and MOS score is used to reflect the quality of the accent-reduced utterances. Results show that segmental modification introduces more distortions than prosodic modification, and combined modification introduces the highest level of distortions. These results are also consistent with those reported in previous studies ([6] and [14]) which use the ARCTIC corpus. In addition, the MOS of accent-reduced utterances without interpolation are also calculated for comparison. As shown in Fig. 4, it is obvious that the modified utterances without spectral interpolation have a lower MOS, showing the necessity of spectral interpolation. The differences among each pair of MOS shown in Fig. 4 are statistically significant based on two-way ANOVA (p<0.001).

4. DISCUSSION ON EXPERIMENTAL RESULTS

4.1 Influence of Nationality and Corpus

The evaluation results in the last section demonstrate that the proposed methods can reduce the accentedness of non-native speech. The performance of prosodic modification is much more significant than segmental modification for the BURNC based corpus. This observation is different from the results obtained from the ARCTIC corpus in both of our experiments and those reported in [6].

The first possible explanation is the difference in the nationality and characteristics of English learners. Most of the foreign students in our corpus are Chinese, Singaporean, and Vietnamese (8 out of 10), whose pronunciations are generally acceptable whereas the prosody of their utterances is poor. In contrast, the non-native speaker in the ARCTIC corpus is an Indian whose utterances have prosody close to that of a native speaker but with more issues in pronunciation. In Fig. 5, the reduction of accent scores using prosodic modification are plotted for all the 10 students in the BURNC based corpus. Vertical dash lines show the maximum and minimum of reduced accentedness and boxes show the 80 percent ranges. Mean values are shown by the central red dash line. It is clear that the reduction in accentedness of student 6 and student 7 (the two Indian students) are not obvious compared with all the other 8 students, as indicated by the green horizontal line.

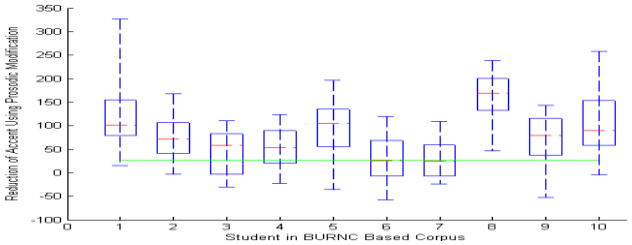

Fig. 5: Accent Reduction Using Prosodic Modification

This observation is also consistent with[4], which reports the use of prosodic modification to help L2 speakers of German who are native Americans with intonation rules quite different from those of German.

The second explanation of the difference stems from the reference corpus. Unlike ARCTIC, BURNC was collected primarily to support generation of prosodic patterns for text-to-speech synthesis systems. Therefore, it possesses abundant prosodic features. In contrast, the prosody is

comparatively flat in the ARCTIC corpus, leading to a weaker impact of prosodic modification as compared with BURNC. The flatness of prosody in ARCTIC has also been cited in [14] to explain the subjective evaluation results.

4.2 Selection of Appropriate Modification Methods

According to the experimental results, there seems to be a tradeoff between accentedness reduction and speech quality: combined modification introduces more distortions but reduces accentedness more significantly than just segmental or prosodic feature modification alone. Though single modification may reduce accentedness to a less extent, it results in a higher acoustic quality. Hence, the preference of language learners (i.e., quality vs. nativeness) should be considered when performing accent reduction

Second, the nationality and characteristics of students are another issue. For students whose prosody is close to that of native speech but with pronunciation issues (e.g., Indian speaker in ARCTIC corpus), segmental modification is more desirable. In contrast, students with unnatural prosody like Singaporean or Chinese speakers in our corpus may prefer prosodic modification. The most effective accent reduction method can be selected by catering to the problematic issues of English learners.

Finally, the reference corpus can also influence accent reduction significantly. A prosody abundant reference corpus may require more efforts in producing the desired prosody, thus prosodic modification is preferred. In contrast, a corpus with less prosody dynamics reduces the difficulties in imitating the native prosody, thus enabling students to focus more on pronunciation practice.

In summary, accent reduction method should take into account the student's preference, nationality as well as the available reference corpus.

5. SUMARRY

This paper studies accent reduction techniques which generate reference speech in the learner's own voice for language learning purposes. Both prosodic and segmental modifications are used to reduce the accentedness of speech. Selective prosodic modification and spectral envelope interpolation are employed to enhance the quality of the accent-reduced speech. Results from different corpora show the influence of reference corpus and nationalities. Such issues therefore should be considered when selecting a suitable accent reduction method.

In future, the effects of other factors on accent reduction (e.g., speaking rate, gender, etc.) will also be studied.

6. ACKNOWLEDGMENT

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

7. REFERENCES

[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] 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.

[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, New Zealand, 2006, pp. 24-29.

[6] 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.

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

[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] K. Yanagisawa, and M. Huckvale, "Accent morphing as a technique to improve the intelligibility of foreign-accented speech," in International Congress of Phonetics Sciences, Saarbrücken, Germany, 2007.

[10] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 676-681.

[11] 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.

[12] C. Kim, and W. Sung, "Implementation of an intonational quality assessment system," in ICSLP-2002, pp. 1225-1228.

[13] D. T. Chappell, and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication,* vol. 36, no. 3-4, pp. 343-373, 2002.

[14] 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.

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

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

[17] 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.