Towards an Automatic Foreign Accent Reduction Tool

Kwansun Cho & John G. Harris

Department of Electrical and Computer Engineering
University of Florida, U.S.A.
{kscho; harris}@cnel.ufl.edu

Abstract

An automatic tool to reduce foreign-accent is described and evaluated. An unaccented speech utterance was used to improve three prosodic features of a corresponding foreign-accented utterance. The duration, pitch and intensity of the foreign-accented speech utterance were modified using DTW (Dynamic Time Warping), WSOLA (Waveform Similarity Overlap Add), and other automatic speech processing algorithms. The modified speech utterance was then evaluated to determine the perceived foreign accent compared to the original. Fifteen native speakers of American English took part in the perceptual test to rate the degree of foreign-accent in Korean-accented American English. The results show that the modified Korean-accented utterances were perceived to have a lower degree of foreign-accent than the original Korean-accented utterances.

1. Introduction

People who learn a second language (L2) as an adult usually speak with a foreign accent. A foreign accent is defined as a change in articulation or intonation patterns of a non-native speaker due to his or her first language (L1) background [6]. It is natural for a non-native speaker to have a foreign accent in L2 speech unless he or she begins learning the L2 at a very early age [1]. Unfortunately, there are several problems with foreign-accented speech. Firstly, there is a decrease in intelligibility, perhaps an occasional word lost for slight accents and possibly a complete loss of understanding for severe accents. Secondly, a foreign accent can have a negative effect on the foreign individuals since it may deny them full acceptance of into society. Therefore, educational tools to help reduce foreign accents and even automatic accent removal algorithms are very desirable.

In this paper, an automatic foreign accent reduction tool is designed and implemented, as a preliminary attempt, based on a prosody transplantation technique [11]. The technique transplants three chosen prosodic features from a native speech utterance unto a foreign-accented speech utterance.

This paper is organized as follows. In Section 2, the overall architecture of an automatic foreign-accent reduction tool is described. In Section 3, we explain the signal processing algorithms used for modifying prosodic features. In Section 4, the perceptual experiment and results are presented. Section 5 discusses the challenges of implementing a practical automatic tool to reduce foreign accent and provides several suggestions as a future work.

2. Automatic foreign accent reduction tool

The overall architecture of an automatic foreign accent reduction tool is illustrated in Figure 1. A foreign accented speech utterance (referred to as "U1") is presented as an input to the tool. A corresponding native speech utterance (referred to as "U2") having the same textual context as U1 is also required. The durations of U1 and U2 are expected to be different since speech is time-dependent and two speech utterances are spoken by two different speakers. Therefore, time alignment needs to be performed with U1 and U2 to obtain the timing relationship between the two speech utterances. The duration of U1 is modified according to the timing relationship information. After the duration of U1 is matched with that of U2, the automatic tool analyzes U1 and U2 in terms of prosodic features. In this paper, the values of the pitch (or F0) and intensity are analyzed. At the last step, the tool transplants the values of two prosodic features from U2 unto U1. The output (referred to as "U1*") is expected to maintain the voice characteristics of U1 except the three features. As explained, the values of the three features of U1* are the same as those of U2. It is hypothesized that the final output, U1* is less-accented than U1.

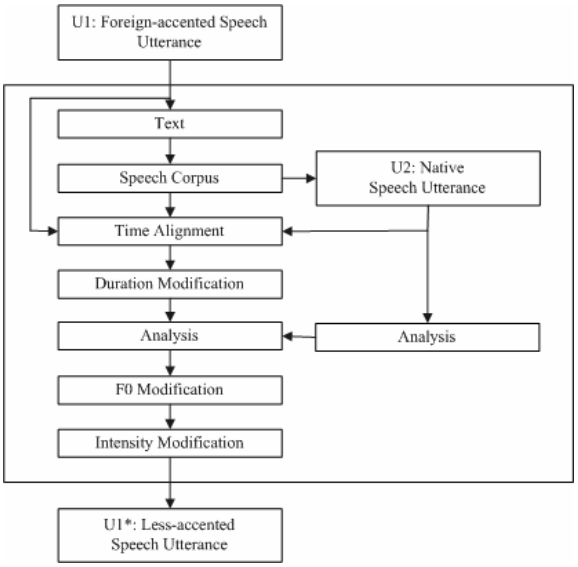

Figure 1: *Architecture of an automatic foreign accent reduction tool*.

There are a few immediate applications of this system, for instance where a famous speaker is needed to narrate a movie in a non-native language. It would not be difficult to record both the famous but accented dialog along with a version from a native speaker. On the other hand, a practical, real-time accent reduction system, implemented on a cell phone for example, would only have access to the foreign accented speech. A sophisticated speech recognition program would

then be required to find the corresponding native language utterances in a database.

3. Modification Techniques

3.1. Duration

3.1.1. DTW (Dynamic Time Warping)

The DTW algorithm [4][9][10] is performed to calculate the temporal relationship between two speech utterances, U1 and U2. Each utterance is divided into Hanning-windowed frames of 30 ms length at 15 ms intervals. The STFT (Short-Time Fourier Transform) features from each frame are extracted. A matrix is constructed by elements, $d(i, j), i = 1,..., I, j = 1,..., J$, and where I and J are the number of frames of U1 and U2, respectively. $d(i, j)$ is computed by the cosine distance between the magnitudes of the STFT. The minimum distance $D(i, j)$ is computed with the following equation.

$$D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \quad (1)$$

As shown in Figure 2, the DTW algorithm obtains the optimal path having the minimum distance through the matrix while reducing the amount of computation. The optimal path in Figure 2 indicates the best matching pairs between the frames of U1 and U2.

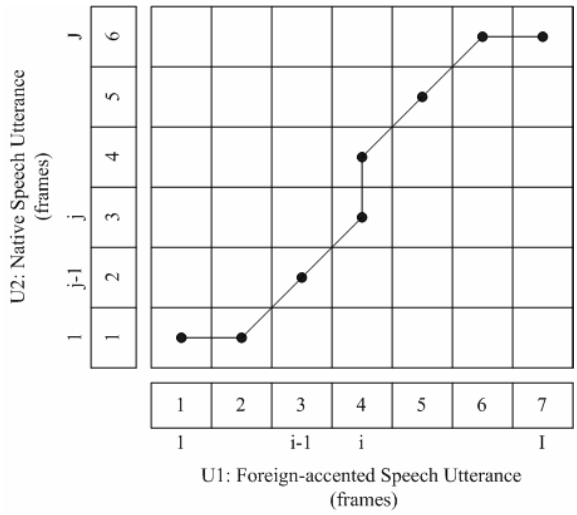

Figure 2: *Time alignment of U1 and U2.*

3.1.2. WSOLA (Waveform Similarity Overlap Add)

The duration of each frame of U1 is modified according to the path indicating the optimal alignment between U1 and U2 using WSOLA [3][7]. WSOLA forms an output speech by overlapping and adding the windowed segments excised from an input speech like the other OLA (Overlap and Add) algorithms do [2]. However, a major difference between WSOLA and the others is that WSOLA considers the concept of waveform similarity to construct the output speech that maintains a natural continuation of the input speech. Figure 3 illustrates the time-scaling procedures of a basic version of WSOLA. Sa, Ss and Δ are defined as an analysis instant, a synthesis instant and a tolerance respectively. Since Sa is longer than Ss in this example, the output speech will be faster than the input speech. At the beginning of the procedures, segment (A) is cut from the input speech around Sa. Segment (A) is repositioned at Ss. Segment (a) is equal to segment (A). A segment (b) that will overlap and add with segment (a) in a synchronized way is needed to be found from the input speech. The candidates of segment (B) are located near the next analysis instant, [Δ-(Sa+Sa), (Sa+Sa)+Δ]. As shown in Figure 3, the WSOLA algorithm will choose the segment among the candidates that best resemble segment (A*). In this paper, the cross-correlation between segment (A*) and one of candidates of segment (B) is used as a waveform similarity measure. The best position for segment (B) is selected by one of candidates of segment (B) having a maximal similarity measure. Segment (B) is excised from the input speech around the best position and is repositioned as segment (b). After overlapping and adding segment (b) with segment (a), the processing is repeated in the same manner until the end of the input speech.

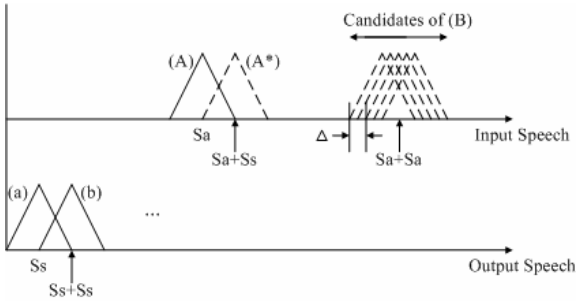

Figure 3: *Illustration of WSOLA.*

3.2. F0 (fundamental frequency)

3.2.1. LPC (Linear Predictive Coding) residual autocorrelation method

A commonly used method for F0 estimation is based on detecting the largest peak value of autocorrelation of a frame [4]. In this paper, the 15th order LPC coefficients of each frame of U1 or U2 are predicted. LPC residual signals are also computed by the inverse filter of each frame of U1 or U2. F0 is the value of the highest peak of the autocorrelation of a LPC residual signal. The F0 value of unvoiced regions is forced into a value, i.e. "0", according to the fixed threshold to classify them since the F0 values of unvoiced regions are essentially random. Before the obtained F0 values are input to F0 modification, some incorrect pitch values are corrected by a smoothing technique. Figure 4 shows an example of F0 contours of U1 and U2.

3.2.2. PSOLA (Pitch Synchronous Overlap and Add)

The PSOLA algorithm [9] directly changes an F0 value of a windowed frame in the time domain. The algorithm consists of three steps. First, the speech utterance is divided into separate frames according to each pitch mark in the original speech utterance. Each frame generally keeps two to four

pitch periods. The second step causes the smaller signals to be altered by repeating or leaving frames depending on the pitch scaling factor of the synthesized speech utterance. (The second step mainly performs F0 modification). In the final step, the PSOLA algorithm recombines the remaining smaller signals by overlapping and adding. The spectrum of the reconstructed speech utterance remains the same after changing F0 values for the last step. Therefore, only the F0 values of the original speech utterance are modified while keeping the other vocal qualities the same. In Figure 4, two F0 contours are indicated. After detecting F0 values of U1 and U2, the F0 contour of U1 is modified with that of U2. The arrows in Figure 4 illustrate the directions of F0 modification of U1.

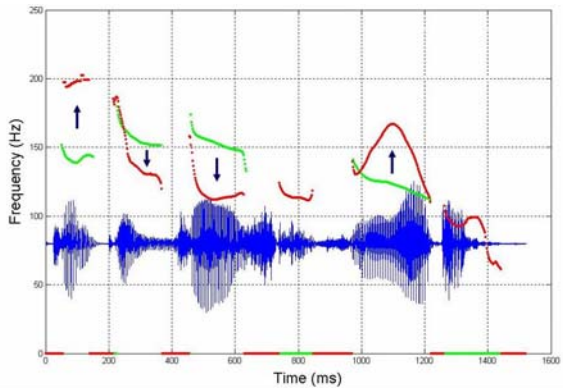

Figure 4: *F0 Detection / Modification*.

3.3. Intensity

3.3.1. Multiplication of RMS (Root Mean Square) energy

Intensity modification is relatively simple compared to duration or F0 modification. The intensity of each frame of U1 is modified using the multiplication of RMS energy. For a vector *X*, RMS energy of *X*, *RMS*(*X*), is defined as follows.

$$RMS(X) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \qquad (2)$$

where $X = \{x_1, x_2, ..., x_n\}$. After calculating the RMS energy of each frame of U1 and U2, the intensity of U1 is modified by multiplying the frames of U1 and the RMS energy ratio of U2 to U1.

4. Subjective Evaluation

In order to investigate if the modification of the three prosodic features results in a change in the perceived foreign-accent and to evaluate the sound quality of the speech utterances modified by the signal processing algorithms, a perceptual test to rate the degree of foreign-accent of a speech utterance has been designed and performed.

4.1. Methods

4.1.1. Talkers

Ten male native speakers of Korean and ten male native speakers of American English having no history of speech or hearing impairment were recruited from the University of Florida for the recording of stimuli. All Korean subjects reported that they had begun learning English in school at the age of 13 years or later.

4.1.2. Stimuli

Four sets of stimuli have been designed for the perceptual test: Korean utterance with American intonation (K*), American utterance with Korean intonation (A*), Korean utterance with Korean intonation (K) and American utterance with American intonation (A). K* and A* were modified stimuli. K and A were original stimuli. In the first two sets of stimuli (K* and A*) the three prosodic features of the Korean speakers and American speakers were switched with one another using the techniques explained in section 3.

4.1.3. Listeners

Fifteen native speakers of American English having normal hearing took part in the perceptual test. All participants were above the age of 18 years on the day of testing.

4.1.4. Procedures

Each talker was asked to read a given list of sentences through a head-mounted microphone with normal speed and intonation in a sound-treated room to record stimuli. The stimuli were digitized at 22.05 KHz with 16 bit resolution using CoolEdit.

In the perceptual test, each listener was presented with one test sentence at a time at a comfortable listening level in a quiet room through headphones. Each listener was asked to rate the degree of foreign-accent using a 7-point rating scale. On the scale, a value of 1 means there is “no foreign-accent” while a value of 7 means there is “strong foreign-accent.” When a button was pushed, the next sentence was presented. Each stimulus was given three times in random order during the perceptual test.

4.2. Results

The mean inter-listener correlation was found to be 0.86 and ranged from 0.65 to 0.94. The ratings obtained for each of the four sets of stimuli were averaged and are shown in Figure 5. The mean foreign accent ratings are 1.24, 2.10, 5.08 and 4.75 for A, A*, K and K* respectively. As expected, the modified Korean-accented stimuli (K*) were perceived to have 8.59% lower degree of foreign-accent on average than the original Korean-accented utterances (K). The sound examples can be heard at http://plaza.ufl.edu/ckstone/ksc_sp2006_files.htm.

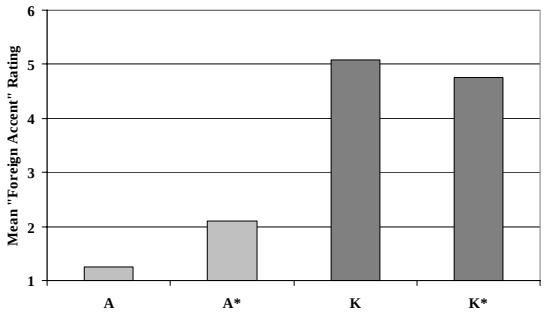

Figure 5: Results of the perceptual test.

5. Discussion

The concept of prosody transplantation has been applied to automatic foreign accent reduction, possibly leading to exciting tools for new applications. However, retaining a reasonable quality for the modified foreign accented speech utterances is very challenging since original foreign accented speech utterances are quite different from the native speech utterance although they have the same textual context due to our L1 background. It has been known that common mistakes often made by native speakers of Korean in English include deleting (or inserting) certain vowels or consonants unnecessarily. These mistakes could be a main reason for the quality degradation after prosody modification since applying the DTW algorithm to the different realizations between two speech utterances is problematic. Therefore, we need a special time alignment algorithm to compensate the unavoidable mismatches between the speech utterances of L1 and L2 speakers and research on new methods for time alignment.

Also, we have only have attempted to implement simple basic algorithms for pitch modification. In order to achieve a development of effective automatic foreign accent reduction algorithms, further research is needed on more reliable prosody modification techniques without serious artifacts and the prosodic features that play a greater role in prosody modification are needed to be done as future work.

Finally, foreign-accented speech is characterized by the prosodic differences discussed in this paper as well as phonemic differences which have been ignored. Often foreign speakers do not (or cannot) produce the correct phonemes which is best exemplified by the classic “r” vs. “l” problems for Japanese speakers of English. In order for a foreign accent reduction system to correct the phonemic differences, a sophisticated speech recognition engine would be required to locate the mispronounced phonemes and splice in alternate phonemes with the correct prosody and coarticulation.

6. Conclusions

An automatic foreign accent reduction tool has been designed and implemented using the well-known signal processing algorithms and techniques. The results of the perceptual experiment in this paper demonstrate that it is possible to reduce the degree of foreign-accented speech utterances by modifying selected prosodic features.

7. Acknowledgements

This project was supported by the Korea Association of Information Management. We would like to thank Dr. Rahul Shrivastav in Communication Sciences and Disorders at the University of Florida for helpful comments.

8. References

[1] Flege, J., 1988. Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America* 84(1), 70-79.

[2] Moulines, E.; Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9(5), 453-467.

[3] Verhelst, W.; Roelands, M., 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of ICASSP-93. IEEE*. 2, 554-557.

[4] Rabiner L.; Juang B., 1993. Fundamentals of speech recognition. New Jersey: Prentice Hall, 204-208, 100-122.

[5] Moulines, E; Verhelst, W., 1995. Time-domain and frequency-domain techniques for prosodic modification of speech. *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal (eds.). Elsevier Science, 519-555.

[6] Arslan, L; Hansen, J, 1997. A study of temporal features and frequency characteristics in American English foreign accent. *Journal of the Acoustical Society of America* 102(1), 28-40.

[7] Verhelst, W., 1997. Automatic postsynchronization of speech utterances. In *Proceedings of European Conference on Speech Communication and Technology*, 899-902.

[8] Verhelst, W., 2000. Overlap-add methods for time-scaling of speech. *Speech Communication* 30(4), 207-221.

[9] Huang, X.; Acero, A.; Hon, H., 2001. Spoken language processing: a guide to theory, algorithm, and system development. New Jersey: Prentice Hall, 383-384, 820-823.

[10] Ellis, D., 2003. http://www.ee.columbia.edu/~dpwe/

[11] Verhelst, W.; Brouckxon, H., 2003. Rejection phenomena in inter-signal voice transplantations. *IEEE Workshop on Applications of Signal Processing to Aduio and Acoutics.* New York: New Paltz.