SWAPPING NATIVE AND NON-NATIVE SPEAKERS' PROSODY USING THE PSOLA ALGORITHM

Kyuchul Yoon
English Division
Kyungnam University
449 Wolyong-dong, Masan, Kyungnam, 631-701, Korea
E-mail: kyoon@kyungnam.ac.kr

ABSTRACT

This paper presents a technique of imposing the prosodic features of a native speaker's utterance onto the same sentence uttered by a non-native speaker. Three acoustic aspects of the prosodic features were considered: the fundamental frequency (F0) contour, segmental durations, and the intensity contour. The fundamental frequency contour and the segmental durations of the native speaker's utterance were imposed on the non-native speaker's utterance by using the PSOLA (pitch-synchronous overlap and add) algorithm [1] implemented in Praat [2]. The intensity contour transfer was also done in Praat. The technique of transferring one or more of these prosodic features was elaborated and its implications in the area of language education were discussed.

INTRODUCTION

One of the most critical tasks in the acquisition of a foreign language is the acquisition of the prosodic features of the language. The definition of prosodic features can vary but in this study we see them as the intonation contour, the assignment of phrase breaks, the durations of the speech segments and the intensity contour of an utterance.

In the traditional classroom environment, foreign or second language teachers explicitly taught the prosodic features of the target language. For example, Korean teachers teaching English to Korean students put much emphasis on the word stress, the intonation pattern and etc. In most cases students learn these prosodic features of the foreign language by listening/watching and repeating their teachers or the native English speakers played on a cassette/video tape or CD-ROM/DVD education software.

In most cases, the feedbacks the students can get when they make mistakes are limited. In a classroom environment, teachers give their students specific instructions, making them repeat the target utterance. It is not uncommon that the teacher draws the intonation contour of the target utterance on the chalkboard to help students understand the point s/he is making.

In a self-study environment with CD-ROM/DVD education software, however, students could get less feedback than in the classroom environment. There are software, such as Dr.Speaking(R)[3], that give a visual feedback. The software records what the learner produces following a native speaker's utterance, draws the intonation contour of the learner, and displays the two intonation contours to show how closely the two match.

In either the classroom or the self-study environment, the major form of feedback is visual, which does not seem to be optimal given that the student is learning a spoken form of the target language. Even if the feedback is in audio, it is the voice of other people repeating the same target utterance.

What if the feedback is in the voice of the student, but with the prosodic features of the native

speaker? For a student less talented in learning a foreing language, the new type of audio feedback could give motivation that none of the traditional feedback could have offered. The new audio feedback works as follows. The software equipped with the technique plays the target sentence uttered by a native speaker, records what the language learner repeats, imposes only the prosodic features of the native speaker onto the learner's utterance, and plays back the learner's utterance with the native speaker's prosody, demonstrating that she could "speak" like the native speaker.

This paper presents the technique of imposing some or all of the prosodic features of a native speaker's utterance to the same utterance produced by a non-native language learner. The phrase breaks, segmental durations and the intonation contour were manipulated using the PSOLA (pitch-synchronous overlap-add) algorithm [1] implemented in Praat [2]. The intensity contour was also manipulated in Praat.

METHOD

For this study, a male native speaker of Korean in his late thirties read aloud an English question sentence *"What did you say before that?"*, which was also uttered and recorded by a male native speaker of English [3]. The Korean speaker was a high school graduate and did not get any college education. His level of English proficiency was low.

Given the target sentence, the technique of transferring the prosodic features of a native speaker's utterance to the non-native speaker's utterance proceeds in three steps. For the illustrative purpose, however, a sample phrase *"came in"* was used below. In the first step, the speech segments of the non-native version are aligned to those of the native version (See Figure 1). The segment alignment step is the most important of all because the quality of subsequent manipulations depend on it. The alignment is followed by the stretching or shrinking of the non-native segments with respect to the native segments using the PSOLA algorithm [1] implemented in Praat [2]. As a result, the non-native segments get to have the same durations as the native segments. Asan added benefit, the location of the phrase breaks will be the same for the two versions of the target utterance.

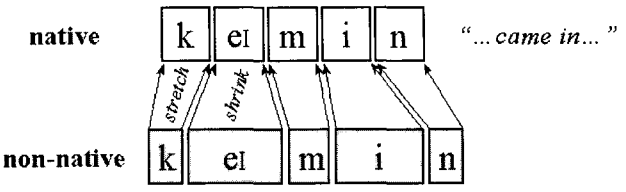

Figure 1. Illustration of step 1: alignment of speech segments. Given an sample phrase "came in", the segments of the non-native version are aligned and manipulated, i.e. stretched or shrinked, with respect to those of the native version.

One thing to note here is that the durational adjustment is performed uniformly, which means that no sub-segmental consideration is made. If, for example, the formant transition at the beginning of the vowel [eI] in the native utterance is longer than that in the non-native utterance before the duration manipulation, the formant transition in the non-native utterance after the manipulation will be much shorter because of the uniform shortening. One way to get around this problem would be to fine-tune the alignment process. If the alignment is done by the sub-segment, e.g. theformant transition versus the steady-state part of a vowel or the gap versus the burst part of a stop, the performance would be improved.

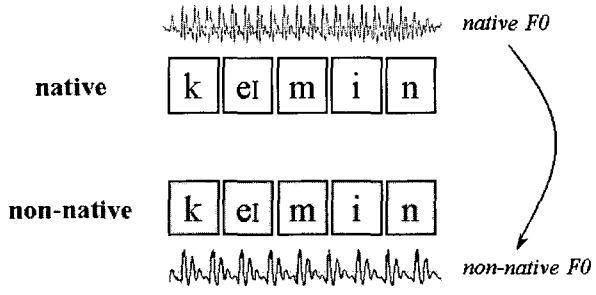

Figure 2. Illustration of step 2: native F0 imposition. After the segmental durations are adjuted, the non-native F0 contour is replaced with the native F0 contour.

In the second step, the fundamental frequency (F0)

contour of the native version is imposed on the non-native version (See Figure 2). This is done in Praat by replacing the F0 contour of the duration-treated non-native version with that of the native version. This is possible, because in the previous step, the durations of the native and non-native version were adjusted. This step is based on the assumption that the duration manipulation of step 1 is perfect. Therefore, it is possible that the relative position of the F0 peak in the vowels of the non-native utterance is slightly different.

In the third and last step, the intensity contour of the native version is imposed on the non-native version (See Figure 3). In Praat, this is done by mathematically "neutralizing" the intensity contour of the non-native version and importing the intensitycontour of the native version. As pointed out in the second step of F0 manipulation, this step also depends on how well the segments were aligned in the previous step. If you go from step 1 through 3, you will have replaced all the prosodic features of the non-native utterance with those of the native utterance. If you stop after the second step, you will have replaced the durations and the F0 contour only.

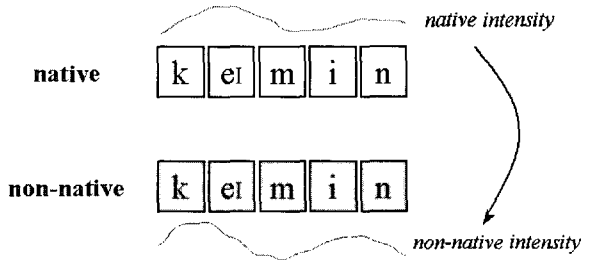

Figure 3. Illustration of step 3: native intensity contour imposition. After the adjustments of the segmental durations and the F0 contour, the non-native intensity contour is replaced with the native intensity contour.

Selective imposition of prosodic features. It is also possible to selectively impose some of the prosodic features of the native speaker's utterance. One can think of a number of ways to do so. Out of the three prosodic features, i.e. the segmental durations, F0 contours and intensity contour, only one from the native utterance can be imposed onto the non-native utterance. Sometimes, two prosodic features can be applied. Depending on the purpose of the work, different sets of features can be manipulated.

If the only prosodic feature that needs to be copied from the native utterance is its segmental durations, then step 1 of the Method section is sufficient. However, if it is either the F0 contour or the intensity contour, an additional step is necessary. Recall that in step 1 the segmental durations of the non-native utterance were adjusted with respect to those of the native utterance (Figure 1). Let us call it the normal order. In the additional step, the reverse order of step 1 needs to be done, i.e. the durations of the native utterance need to be adjusted with respect to those of the non-native utterance. For example, the [k] segment of the native utterance in Figure 1 will have to be shortened.

During the process, additional/excess frames are added/deleted to/from the original sound signal [1], which also affects the F0/intensity contour of the native utterance. After the additional step, the new version of either the F0 contour or the intensitycontour of the native utterance can be imposed on the non-native utterance. As the procedure involves an additional step that modifes the original native utterance, the resulting non-native utterance cannot said to contain the original F0/intensity contour in the strict sense.

If two of the prosodic features need to be copied, either the normal or reverse durational manipulation can be combined with either the F0 or the intensity contour manipulation. For example, if it is the segmental durations and the F0 contour of the native utterance that need to be imposed on the non-native utterance, the normal durational manipulation as shown in Figure 1 can be applied with a subsequent F0 contour replacement. If it is the segmental durations and the intensity contour of the native utterance, the subsequent replacement can be done with the intensity contour of the native utterance.

The impositon of the F0 and intensity contour of the native utterance on the non-native utterance can start with the reverse durational manipulation,

followed by the transfer of the F0 and intensity contour of the native utterance onto the non-native utterance.

RESULTS

The spectrographic comparison of the native and non-native utterance before and after the application of the technique is shown in Figures 4 and 5. The technique of imposing all the prosodic features was employed for the target utterance *What did you say before that?*. As seen in Figure 4, the non-native utterance is different from its native counterpart in every aspect of the prosodic features. Although both speakers were male, the native speaker was generally higher in its F0 contour.

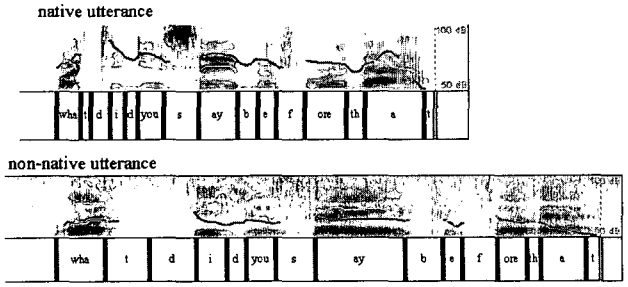

Figure 4. Spectrographic comparison of the native and non-native utterance before the application of the technique. The target sentence was "What did you say before that?". *The blue thicker line represents the F0 contour and the yellow thinner line represents the intensity contour.*

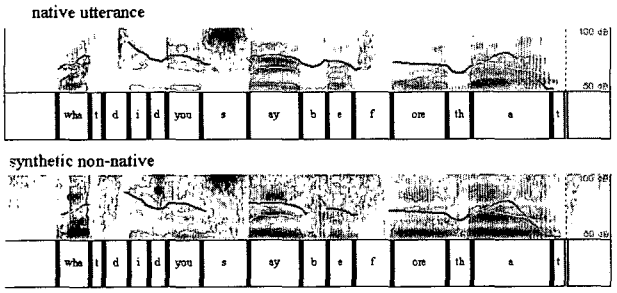

Figure 5. Spectrographic comparison of the native and non-native utterance after the application of the technique. The blue thicker line represents the F0 contour and the yellow thinner line represents the intensity contour.

However, after the application of the technique (See Figure 5), the two utterances became almost identical in all aspects of the prosodic features. The durations of the matching segments are the same, although, as pointed out earlier, the precision dependson the accuracy of the segment alignment process.

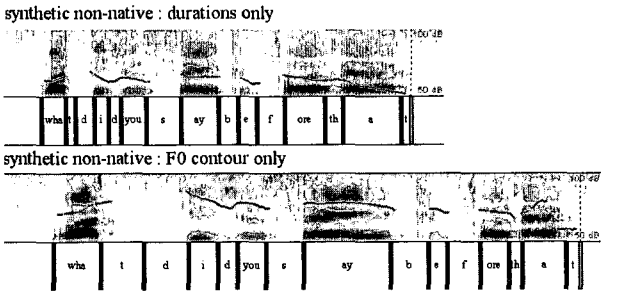

Figure 6. Spectrographic comparison of the non-native utterances after impositon of either the durations or the F0 contour of the native utterance. The blue thicker line and the yellow thinner line represent the F0 contour and the intensity contour respectively.

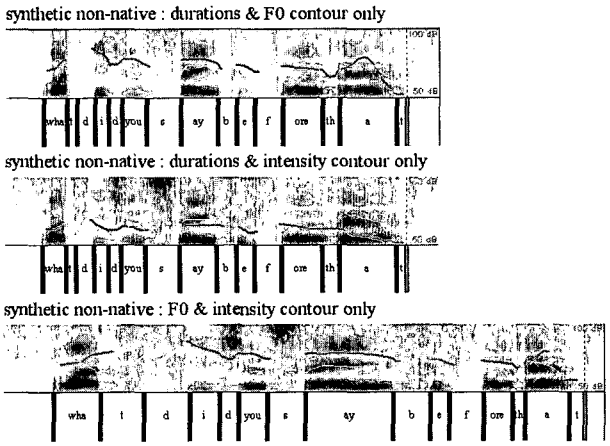

Figure 7. Spectrographic comparison of the non-native utterances after impositon of two of the prosodic features of the native utterance. The blue thicker line represents the F0 contour and the yellow thinner line represents the intensity contour.

The F0 and intensity contour seem almost identical, although slight sub-segmental variations may be present. Differences in segmental quality are observed, for example, in the [s] segment. This may be partly due to the difference in the formant characteristic of the [s] segment of the non-native

speaker and partly due to the weakness of the PSOLA algorithm itself [1].

Figures 6 and 7 show the result of manipulating only one of the prosodic features, i.e. either the durations or the F0 contour, or two of the prosodic features, i.e. the durations and F0 contour, the durations and intensity contour, or the F0 and intensity contour. The durations only panel in Figure 6 shows the native speaker's durations but we can see the typical flat intonation contour at the end of the utterance which was observed in its longer original version (Lower panel in Figure 4). In the F0 contour only panel of Figure 6, we can see that the F0 contour, although "extended" by the PSOLA algorithm, closely resembles the native speaker's (Upper panel in Figure 5).

The results show that it is possible to manipulate some or all of the prosodic features involved in this study. The manipulation started with the adjustment of the segmental durations followed by the swapping of either/both the F0 or/and intensity contour. Despite some degradations in the quality of the synthesized utterances and discrepancies in the sub-segmental alignment, the technique appears to be potentially useful for various purposes in many areas.

CONCLUSIONS

This paper presented a technique of imposing the prosodic features of a native speaker's utterance onto the same sentence uttered by a non-native speaker. The technique of imposing some of the prosodic features selectively was also presented. The spectrographic comparison of the synthetic utterances shows that this technique can be a useful tool for various purposes.

In the second language education, the technique can be used to give an audio feedback to learners. By having the learners listen to their own voice with the prosodic features of the target native speaker, we could motivate them in a different perspective.It could be integrated as an additional audio feedback to exiting language education software which gives a visual feedback in the form of the F0 contour matching. For this to work, the segmental alignment should be done automatically with the help of an accurate automatic speech recognition technology.

As shown above, selective application of prosodic features can give different levels of audio feedbacks. For example, the learner could be given back her utterance with either the native speaker's F0 contour alone or the segmental durations alone. This could make the learners more aware of the prosodic feature manipulated. By having the learners pay more attention to a particular aspect of the prosodic features of the target language, learners may be able to acquire the prosody of the target language with increased efficiency.

This technique can also be used for correcting pronunciation of patients with a vocal disorder. Given a target utterance made by a normal speaker, the patient could be motivated by listening to her pronunciation with all the normal prosodic features of the target speaker. Since the essence of this technique is swapping prosodic features between speakers, it could be used in relevant performance or perception experiments.

ACKNOWLEDGEMENTS

The author would like to thank professors Joong-sun Sohn at the Department of English Education of Daegu National University of Education, No-Ju Kim at the Department of English Language and Literature of Kyungpook National University and members of the Circle of Experimental Phonetics for their help and comments. Special thanks to Holger Miterer for the Praat scriptlet on intensity copying.

REFERENCES

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

[2] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International,* 5, 9/10, 341-345 (2005)

[3] Peter Ladefoged, *A course in phonetics (Thomson Wadsworth, 2006), CD-ROM*