말소리와 음성과학 제1권 제1호 (2009)
pp. 25~31

The Role of Prosody in Dialect Synthesis and Authentication

Yoon, Kyuchul[1)]

ABSTRACT

The purpose of this paper is to examine the viability of synthesizing Masan dialect with Seoul dialect and to examine the role of prosody in the authentication of the synthesized Masan dialect. The synthesis was performed by transferring one or more of the prosodic features of the Masan utterance onto the Seoul utterance. The hypothesis is that, given an utterance composed of the phonemes shared by both dialects, as more prosodic features of the Masan utterance are transferred onto the Seoul utterance, the Seoul utterance will be identified as more authentic Masan utterance. The prosodic features involved were the fundamental frequency contour, the segmental durations, and the intensity contour. The synthesized Masan utterances were evaluated by thirteen native speakers of Masan dialect. The result showed that the fundamental frequency contour and the segmental durations had main effects on the perceptual shift from Seoul to Masan dialect.

Keywords: speech synthesis, dialect, Masan, Seoul, prosody, fundamental frequency, duration, intensity, Praat

1. Introduction

Regional dialects are known to have differences at many levels of linguistics. Regional dialects of Korean are no exception and different dialects of Korean display different characteristics. There are differences at the phonetic level. [1] compared two Korean alveolar fricatives of Seoul and Busan speakers and found that the two fricatives display different subsegmental characteristics. Although each of the fricatives is perceived to be the same phoneme by speakers of both dialects, the phonetic aspects of the sounds are different. Specifically, Busan fricatives showed much shorter frication and aspiration intervals in word-initial and word-medial positions. Differences in the frequency domain can also be found in Korean regional dialects. [2] found that the two alveolar fricatives of Kyungsang and Cholla dialects showed differences in the frequency domain by more than 1kHz.

In addition to segmental differences, it is also well known that regional dialects display differences at the prosodic level. The prosodic aspect may be one of the most outstanding aspects of regional dialects. A native speaker of Korean would have no difficulty in identifying different regional dialects of Korean just by listening to the prosodic pattern of an utterance. Likewise, a sentence composed of phonemes shared by different regional dialects would be produced with different prosodic patterns. One of the questions that one might ask is, given the same set of segments for an utterance, would it be possible to transform one dialect into another? In other words, would it be possible to do the transformation just by manipulating the prosodic aspects of the utterance? If it were possible, would the components of the prosodic aspect be equally important in the creation of another dialect?

The prosody of an utterance is considered to be composed of three features. The three prosodic features are the fundamental frequency (F0) contour, the segmental durations and the intensity contour. One of the difficulties in the study of prosody is that it is not easy to selectively apply one or more of the prosodic features from one utterance to another. One reason for the difficulty is that the segmental durations of the prosody donor and recipient utterances are not the same. However, the technique of selective cloning of prosodic features was developed in [3]. The technique clones the F0 contour or the intensity contour from

1) Yeungnam University, kyoon@ynu.ac.kr
(This research was supported by the Yeungnam University Research Grant in 2008. (208-A-054-024))

Received: January, 28, 2009
Revision received: March, 8, 2009
Accepted: March, 10, 2009

one utterance to the other after the matching segments are rendered the same in their segmental durations. Applying the technique, it would possible to transfer one or more prosodic features from one dialect to another. Assuming the same segmental composition, it would be possible to transform one dialect to the other dialect at least in terms of the prosodic features of the utterances involved.

One advantage of being able to selectively clone the prosodic features of an utterance is that it is possible to examine the effect of each of the prosodic features in the perception of the cloned utterances. For example, in transforming a Seoul utterance into a Masan utterance prosodically, one can examine the effect of each of the prosodic features in itself or in combination. This would enable one to assess the extent and magnitude of the perceptual effects of the prosodic features concerned. It is expected that as more prosodic features of the Masan utterance are transferred onto the Seoul utterance, more Masan listeners will identify it as authentic Masan dialect.

The goals of this paper are to test the viability of prosodically synthesizing Masan utterances from Seoul utterances and to assess the role of each of the prosodic features in the authentication of the synthesized Masan utterances. In order to evaluate the perceptual effect of the prosodic features alone, it is necessary to control the segmental compositions of the natural utterance stimuli in a perception experiment. However, it is not easy to extract the "neutral" segments from a natural utterance that do not display any subsegmental differences in either dialect. Synthesizing such segments from a formant synthesizer is another option, but it would be harder to synthesize such segments for a sentence-sized utterance. An operational assumption that was made for this work was that phonemes shared by both dialects do not have any subsegmental differences. Of course, this is not entirely true, but had to be adopted for the experiment to proceed.

2. Methods

As mentioned above, Korean regional dialects are known to vary with respect to their phonological, morphological, syntactic, and prosodic aspects among others. Thus it was important to create a set of sentences that were "neutral" in all of these aspects. A "neutral" sentence here means that the speaker of either dialect can utter the sentence and perceive it as their own dialect without any noticeable awkwardness. With the help of a native speaker of Masan and a native speaker of Seoul, two sample "neutral" sentences were created as below.

동대구에 볼 일이 없습니다.

Dong.dae.gu.e bol il.i eobs.seub.ni.da

"I have no business in Dongdaegu."

바다에 보물섬이 없다.

Ba.da.e bo.mul.seom.i eobs.da

"There are no treasure islands in the sea."

All the phonemes of the two sentences are shared by the two dialects. With the operational assumption about the subsegmental differences suggested above, each of the two sentences are regarded as authentic by the speakers of both dialects in terms of their segmental compositions. These sentences are also assumed to display the same type of phonological, morphological, and syntactic characteristics that are shared by the two dialects. Thus, when uttered by the speakers of both dialects, these sentences were different only in terms of their prosodic aspects.

One thing to note when creating synthetic stimuli is the quality of the stimuli. The preparation of the experimental stimuli in this work involves manipulating each of the three prosodic features of the natural utterances. The Seoul utterances will serve as the basic building blocks for creating simulated Masan utterances. It would be unfair to mix non-synthetic natural utterances and synthetic utterances in the same listening experiment. Thus, it was necessary that we have two speakers for Masan dialect. Two native speakers of Masan dialect and one native speaker of Seoul participated in the preparation of the experimental stimuli. One Masan speaker, i.e. "the prosody-donor", gave his prosody to the other Masan speaker, i.e. "prosody-recipient". The Masan prosody-donor also gave his prosody to the Seoul speaker. Therefore, for the first sentence, the following stimuli were synthesized. The same procedure was repeated for the second sentence.

#1. Authentic, but synthetic, Masan utterance
#2. Seoul utterance with Masan segmental durations (D)
#3. Seoul utterance with Masan F0 contour (F)
#4. Seoul utterance with Masan intensity contour (I)
#5. Seoul utterance with Masan durations and F0 contour (D+F)
#6. Seoul utterance with Masan durations and intensity contour (D+I)
#7. Seoul utterance with Masan F0 contour and intensity contour (F+I)
#8. Seoul utterance with Masan durations, F0 contour and intensity contour (D+F+I)

The first experimental stimulus (#1) is an authentic Masan

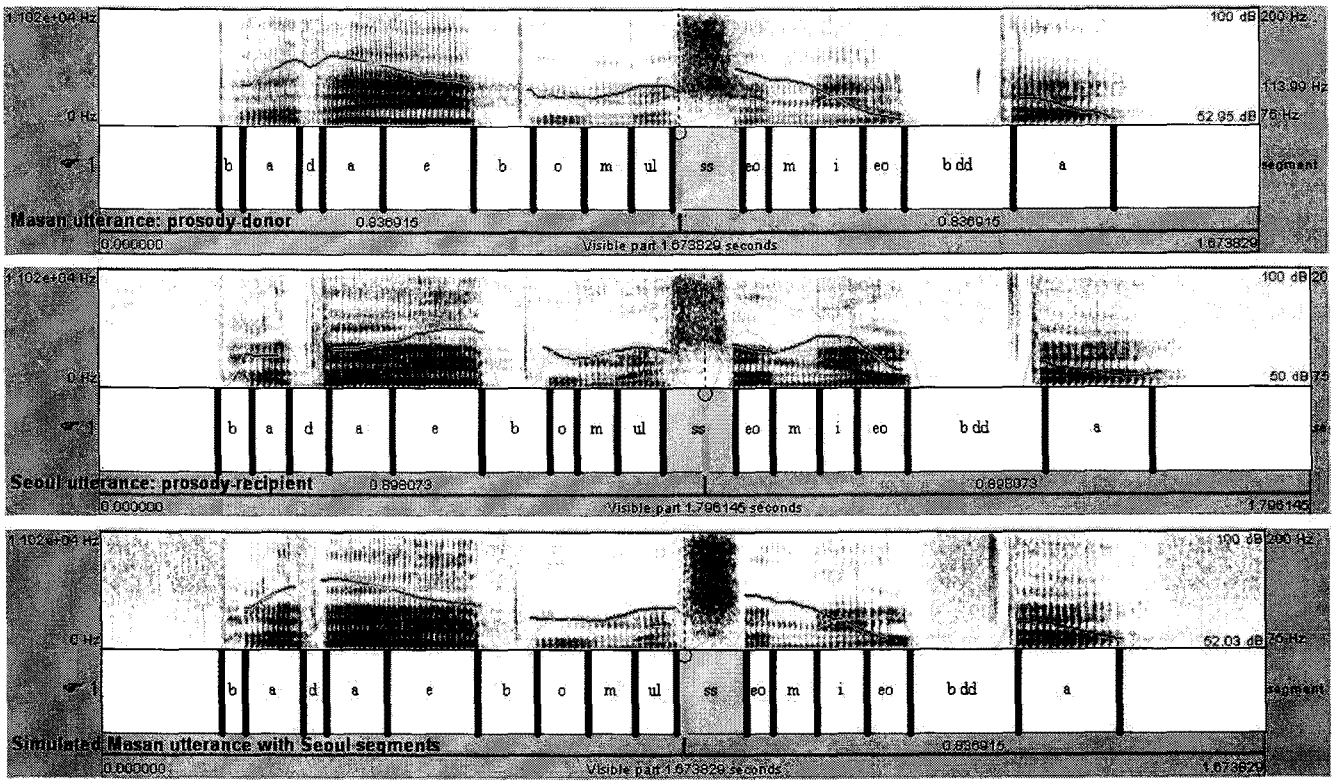

Figure 1. Creation of a synthetic Masan utterance from a Seoul utterance. Upper panel represents the Masan prosody-donor. Middle panel represents the Seoul utterance. Bottom panel represents the synthetic Seoul utterance with Masan prosody.

utterance produced by the Masan prosody-recipient. However, the prosodic features of the stimulus comes from the Masan prosody-donor. The rest of the stimuli are Seoul utterances, but their prosodic compositions differ as indicated above in the parentheses. The stimulus #2 is an authentic Seoul utterance except for its segmental durations. The durational feature comes from the Masan prosody-donor. The stimulus #3 has its F0 contour cloned from the Masan prosody-donor. Two prosodic features come from the Masan donor in the stimuli #5, #6 and #7. For the last stimulus, all the prosodic features of the Masan prosody-donor replace the original Seoul prosody. Thus, the stimulus #8 becomes the "best" synthetic Masan utterance whose segments are from Seoul utterance.

If listeners do not distinguish the stimulus #1 from the stimulus #8, we can safely say that it is possible to synthesize Masan dialect from Seoul dialect just by replacing the prosodic features alone. However, this is not likely to happen because the operational assumption that subsegmental differences do not exist among Korean regional dialects is definitely not true. Then the question is how similar is #8 to #1? We can compare responses for the two stimuli and see if there is any significant difference between the two series of responses.

Two male native speakers of Masan dialect and one male native speaker of Seoul dialect participated in the experiment. All of them wore a head-mounted microphone (Shure SM10A) and the microphone was connected directly to a computer. They produced the two sample sentences and the utterances were sampled at 22kHz. They were asked to say them as naturally as possible in a quiet room. As mentioned above, one Masan speaker served as the prosody-donor. The other Masan speaker and the Seoul speaker received one or more of the prosodic features from the donor. Although the male speakers were in their twenties and thirties and were assumed to have similar pitch ranges, a pitch range modification prior to the synthesis could have employed and produced better synthesis quality.

In order to apply the prosody cloning technique [3], all the utterances were first manually segmented in Praat [5] as shown in <Figure 1>. As <Figure 1> shows, the three prosodic features are different for the two utterances (upper and middle panels). The F0 contour (blue thicker lines) is noticeably different as well as their segmental durations. The intensity contour (yellow thinner lines) is also different. After cloning all the prosodic features of the Masan prosody-donor (upper panel) onto the Seoul utterance (middle panel), the synthetic Seoul utterance (bottom panel) looks very much like its prosody-donor in terms of the component prosodic features.

<Figure 2> shows how each prosodic feature can be cloned onto the target utterance. The top panel is from the Masan

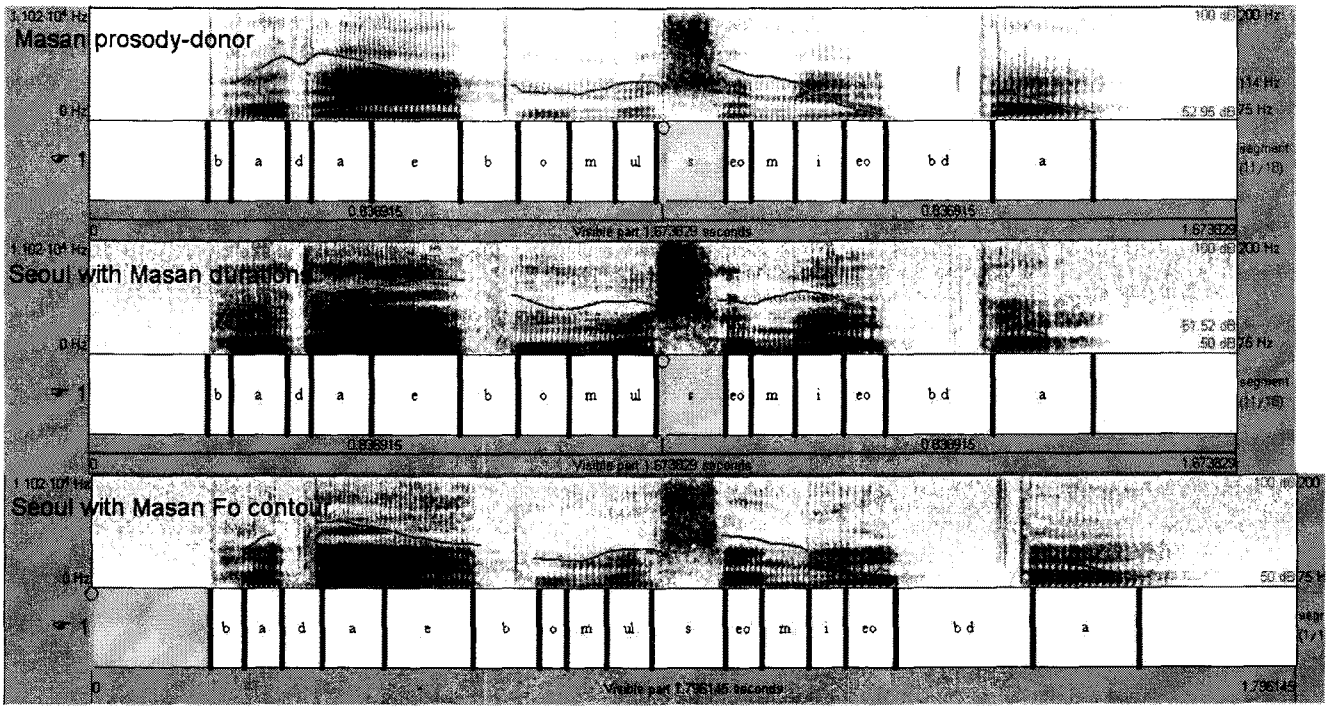

Figure 2. Selective cloning of prosodic features. The Seoul utterance in the middle panel has the segmental durations from the Masan prosody-donor. The Seoul utterance at the bottom panel has the F0 contour from the Masan prosody-donor.

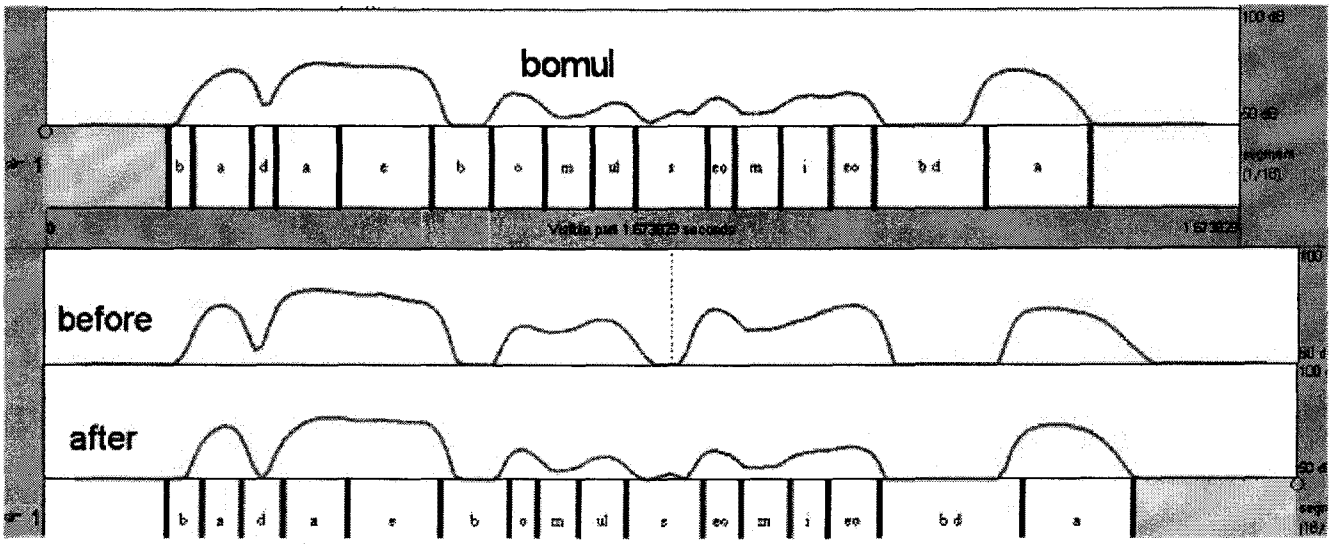

Figure 3. The intensity contour before and after the selective cloning. Note the difference near the word "bomul" before and after the cloning.

prosody-donor. The segmental durations of this utterance were transferred onto the Seoul utterance in the middle panel. The utterance in the middle panel, thus, has all the features as a Seoul utterance except for its segmental durations. The utterance at the bottom has only the F0 contour from the Masan prosody-donor. <Figure 3> shows how the intensity contour was switched. Before the cloning, the intensity contours near the segments "bomul" are very different in terms of the intensity magnitude of the intensity peaks. After the cloning, the intensity contours are nearly the same. Since it was only the intensity contour that has been switched, notice that the overall utterance lengths are different even after the cloning.

Two of the three prosodic features were combined to synthesize the stimuli #5, #6 and #7 in the same manner introduced in [3]. After eight experimental stimuli were synthesized, the same procedure was repeated for the second sentence, yielding another eight stimuli. All the sound files were normalized according to the average intensity of in dB of the

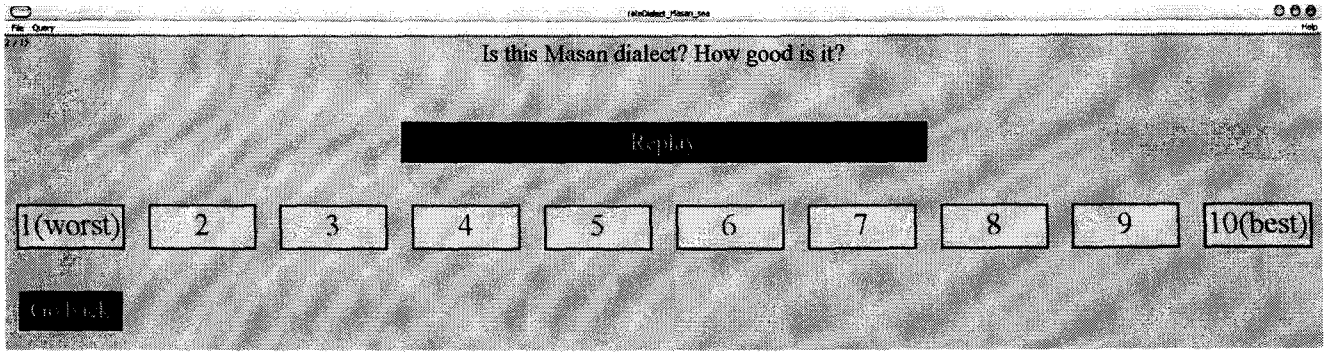

Figure 4. A sample screen from the listening experiment.

files involved.

The sixteen experimental stimuli, eight for the first sentence and the other eight for the second sentence, were presented to thirteen normal-hearing listeners of Masan dialect. They were raised in Masan and Changwon area until they graduated from high schools. The sixteen stimuli were randomized using the *PermuteBalancedNoDoublets* randomization strategy implemented in Praat ExperimentMFC object. For each stimulus played over a headphone, the listeners were asked rate subjectively the goodness (as Masan utterance) of it on a scale of 1 (worst) to 10 (best) and click the corresponding button on the computer screen. The listeners were allowed to replay the stimuli up to ten times before making a decision and allowed to go back to the previous stimulus if a mistake was made.

Before the actual experiment session, a sample session with the synthetic Masan and Seoul utterances was given for the purpose of calibration of the subjects. The synthetic stimuli were two #1 stimuli and one Seoul utterance synthesized in the same way as #1 stimulus. They were told that the two #1 Masan utterances would get a 10 and the other Seoul utterance would get a 1. They were also told that the sixteen stimuli in the actual session could get points in between 1 and 10 depending on their own subjective judgment. A sample computer screen is given in <Figure 4>. The results were statistically analyzed.

3. Results

The histogram of listener responses is given in <Figure 5>. The number of responses for the score points 1, 2, 9 and 10 was 123 out of 182 total responses, which corresponds to 68%. Put differently, it appears that the effects of adding each one of the three prosodic features one at a time were not equal in size. The histogram suggests that the listener responses were somewhat categorical.

The box plot for the listener responses with and without F0

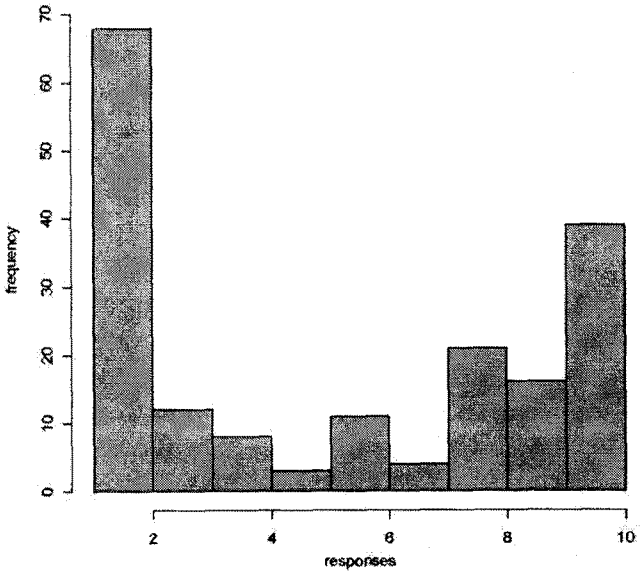

Figure 5. The histogram of listener responses. The horizontal axis represents the score points from 1 to 10.

contour cloning in <Figure 6> shows this pattern. When the Masan F0 contour was copied onto the Seoul utterances, the stimuli was identified as more authentic Masan utterances irrespective of the other prosodic features. For the other two prosodic features, such patterns were not observed. The means for the listener responses in <Figure 7> confirms this pattern. When the synthetic Seoul utterances have the Masan F0 contour, the responses shifted to near 8 score point, but not as high as the responses for the synthetic Masan utterances.

When the listener responses for the two experimental stimuli #1 and #8 were compared using the *t*-test, there was a significant difference in means (*p*-value = 0.0014). Recall that #1 was the authentic synthetic Masan utterance and #8 was the Seoul utterance with all the prosodic features of its matching Masan utterance. The *t*-test result means that listeners responded differently for the two types of stimuli. Since they are both synthetic stimuli, listeners may have noticed some unnaturalness in the synthetic quality or they may have noticed some other

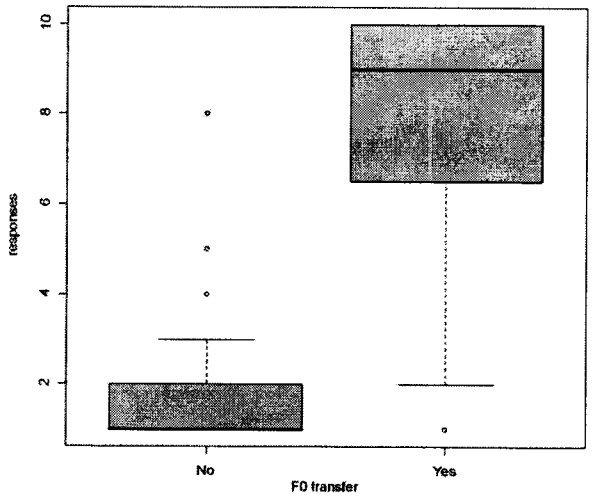

Figure 6. The box plot of responses for the stimuli with (Yes) and without (No) F0 contour cloning.

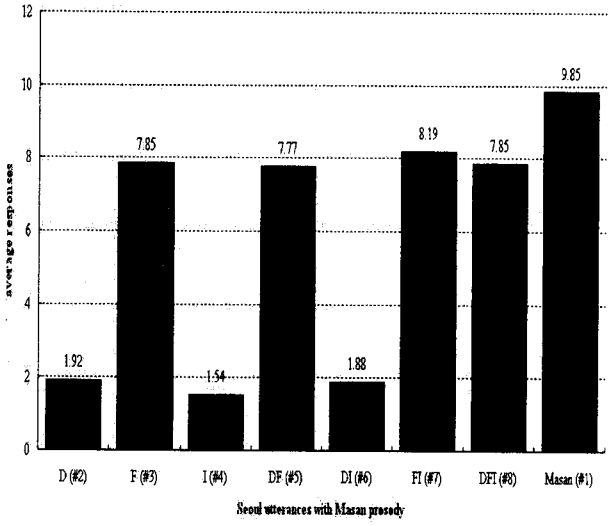

Figure 7. The mean responses for the Seoul utterances. D, F and I represent durations, F0 and intensity respectively. The rightmost column is for the authentic synthetic Masan utterances. The numbers in parentheses indicate the stimulus #.

known and unknown segmental differences. This also suggests that cloning prosodic features alone could not render a Seoul utterance into a Masan utterance even though the stimuli are considered to be composed of the phonemes shared by both dialects. Allophonic differences of the component segments may have affected the responses.

When the responses were analyzed using the three-factor (presence or absence of duration, F0 and intensity transfer) ANOVA analysis with repeated measures, there were main effects of segmental durations ($F(1,12)=11.53$, $p=0.005$) and the F0 contour ($F(1,12)=141.12$, $p=0.001$) but no interactions. The p-value of the F0 contour agrees with the box plot given in <Figure 6>. When a regression analysis was performed on the three factors, we get <Table 1>. As <Table 1> shows, the three factors account for 67.3% of the total responses. Of the three factors, the presence or absence of the F0 contour transfer is responsible for the most of the accountable responses (0.825 out of 1.0), followed by the intensity transfer and the durations transfer.

Table 1. Results of linear regression analysis on the three factors.

Adjusted R square	0.673	
Variables (presence/absence of)	***Beta***	***Significance***
(1) Segmental durations transfer	0.004	0.930
(2) F0 contour transfer	0.825	0.000
(3) Intensity transfer	0.006	0.884

4. Conclusion

One goal of this paper was to test if it was possible to synthesize an authentic Masan utterance from an authentic Seoul utterance by transferring only the prosodic features of the Masan utterance. The experimental result showed that although the stimuli were supposedly composed of phonemes shared by both dialects, listeners favored the Masan utterance whose segments were also from a Masan utterance. The Seoul utterances with all of the Masan prosody did not get as high scores as the authentic synthetic Masan utterances. It appears that listeners were sensitive to the segmental differences of the synthetic stimuli. The voice quality may also have contributed to the result. The switching of the voice source appeared to have effects on the overall authenticity of the synthesized utterances. However, this informal observation needs to be confirmed with an additional experiment.

The other goal of this paper was to examine the role of each of the prosodic features in the authentication of the synthetic Masan utterances. Although it was assumed that transferring more prosodic features of the Masan utterance would have cumulative effects on the listener responses, it was only the F0 contour that had any noticeable differences on the responses. A regression analysis confirmed that the presence or absence of the F0 contour transfer was responsible for most of the accountable responses. However, an ANOVA analysis showed that both the F0 contour and the segmental durations had main effects on the listener responses.

Although it remains to be tested whether the findings from this study can be generalized to the other Kyungsang dialects, the techniques used in this study can readily be used with other dialects of Korean such as Cholla and Chungcheong dialects.

Findings from experiments with the other dialects of Korean may shed light to the nature and hierarchy of the prosodic features in the authentication of Korean regional dialects.

References

[1] Kyung-Hee Lee, "Comparison of acoustic characteristics between Seoul and Busan dialect on fricatives", *Speech Sciences*, 9(3), pp. 223-235, 2002.

[2] Hyun-Gi Kim, Eun-Young Lee, and Ki-Hwan Hong, "Experimental phonetic study of Kyungsang and Cholla dialect using power spectrum and laryngeal fiberscope", *Speech Sciences*, 9(2), pp. 25-47, 2002.

[3] K. Yoon. "Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody". *Journal of the Modern British & American Language & Literature*, 25(4), pp. 197-215, 2007.

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

[5] P. Boersma. "Praat, a system for doing phonetics by computer", *Glot International* 5(9/10), pp. 341-345, 2001.

• **Kyuchul Yoon**
School of English Language and Literature
Yeungnam University
214-1, Dae-dong, Gyeongsan-si
Gyengsangbuk-do (712-749), Korea
Email: kyoon@ynu.ac.kr