# The role of prosody in dialect authentication
# Simulating Masan dialect with Seoul speech segments

Kyuchul Yoon
Division of English, Kyungnam University
E-mail: kyoon@kyungnam.ac.kr

## Abstract

The purpose of this paper is to examine the viability of simulating one dialect with the speech segments of another dialect through prosody cloning. The hypothesis is that, among Korean regional dialects, it is not the segmental differences but the prosodic differences that play a major role in authentic dialect perception. This work intends to support the hypothesis by simulating Masan dialect with the speech segments from Seoul dialect. The dialect simulation was performed by transplanting the prosodic features of Masan utterances unto the same utterances produced by a Seoul speaker. Thus, the simulated Masan utterances were composed of Seoul speech segments but their prosody came from the original Masan utterances. The prosodic features involved were the fundamental frequency contour, the segmental durations, and the intensity contour. The simulated Masan utterances were evaluated by four native Masan speakers and the role of prosody in dialect authentication and speech synthesis was discussed.

## I. Introduction

The development of theoretical linguistics and the technology of natural language processing has been changing the way we live. One such change has to do with the speech synthesis technology. We are more likely to come across various types of speech synthesis systems in our daily lives than ever before in human history. Although the overall quality of the synthesizers varies greatly on subjective and objective grounds, significant amounts of resources are being used today in the development of multi-lingual speech synthesizers. Just as synthesizers covering multiple languages are useful, those that cover multiple regional dialects of a single language should be equally useful. A general-purpose speech synthesizer that speaks the "standard" variety of the dialects of a particular language can be a good start. The next goal may be the development of multi-dialectal speech synthesizers, perhaps customizable to the dialectal needs of the users.

The time and cost of building such multi-dialectal speech synthesizers may depend on the system architecture. No matter what system architecture one may choose to adopt, one thing to note is that it would not be desirable to build a separate synthesizer for each regional dialect of the language. This is neither practical nor efficient in terms of both the system design and use of resources. With respect to linguistic knowledge, this approach is not recommended as well. As will be mentioned below, studies of Korean regional dialects have shown that they display differences at the levels of phonetics, phonology, morphology, syntax, prosody, etc. Examination of these differences will help design a multi-dialectal speech synthesizer for Korean. For example, the grapheme-to-phoneme module of a concatenative speech synthesizer can be equipped with sets of dialect-specific lexical and postlexical phonological rewrite rules so that appropriate concatenation units may be selected. The

selection of vocabulary should be sensitive to each dialect. With regards to the syntactic and prosodic features, the synthesizer could be equipped with modules that can selectively apply the syntactic and prosodic features of a particular dialect. The users can then select which dialect to use at run-time.

When the two regional dialects display different phonological behavior, the differences can be mapped through a set of phonological rewrite rules, which deal with categorical phonemic changes. If all the phonological behavior involved categorical phonemic changes, dialect simulation through prosody transfer would be a relatively simple task. However, Korean regional dialects are known to be different at the phonetic level. That is, the phonemes that are perceived to be the same by listeners of different dialects show segmental differences that do not involve categorical phonemic change. For example, Lee [1] compared Seoul and Busan dialects with respect to two fricatives and found that the two fricatives display different subsegmental characteristics. Specifically, Busan fricatives showed much shorter frication and aspiration intervals in word-initial and word-medial positions. Since the difference would not involve any categorical phonemic change, the grapheme-to-phoneme module of a synthesizer may not be able to capture the phenomenon. This type of durational difference could be captured in a module dealing with segmental durations. However, adjustments are necessary so that the module can control subsegmental intervals rather than the duration of the whole segment.

A more challenging phonetic aspect of dialect simulation involves subsegmental differences that cannot be captured by any of the conventional concatenative synthesizer modules. One such difficulty can be found when the same phoneme display differences in frequency domain. For example, the lower frequency cutoff of the two Korean fricatives of Kyungsang and Cholla dialects showed frequency differences of more than 1,000 Hz [2]. Unlike formant synthesizers, this type of discrepancy is a serious challenge for concatenative synthesizers.

It is the influence of this phonetic aspect that the current work intended to test. When simulating a target dialect with the speech segments of another dialect, would non-phonemic or phonetic, segmental differences interfere with the perception of the target dialect? For example, utterances of Masan dialect could be synthesized with segments of Seoul dialect. Assuming that these utterances have the phonological, morphological, syntactic, and prosodic features of a typical Masan dialect, would Masan listeners perceive them as authentic Masan utterances? The answer to this question has implications for concatenative synthesizers. If the listeners of the target dialect approves the simulated synthetic utterances as authentic, this means that no re-recording of the target dialect segments would be necessary to build multi-dialect concatenative synthesizers. In other words, the concatenation units of one dialect could be used for another dialect. If the answer is contrary to our assumption, then it would mean that different concatenation units would be required to build a multi-dialect concatenative synthesizer.

As suggested above, this work is based on the assumption that there is a concatenative synthesizer that is equipped with advanced modules that can deal with the phonological, morphological, syntactic, and prosodic differences of the dialects involved. As such a system is not available to the best of author's knowledge, we will simulate such a system by manually creating synthetic utterances with the help of a Praat script introduced in [3]. The script allows one to transfer the prosodic features of one utterance to another utterance. The prosodic features that are transferred are the fundamental frequency contour, segmental durations, and intensity contour. The script uses the PSOLA algorithm [4] implemented in Praat [5]. Therefore each of the simulated target utterances will have all the prosodic aspects of the target dialect except for the fact that the component speech segments are from a different dialect. In terms of concatenative speech synthesizers, the simulated synthetic utterances can be seen as the output from a multi-dialect synthesizer with a near-perfect prosody module. The prosody module can be said to be near-perfect because the prosodic features of the natural utterances were involved in the synthesis process.

In addition to the segmental differences among different Korean regional dialects, there seems to be

differences in voice quality of the utterances. Qualitative examination of the utterances from Masan and Seoul dialects showed differences in voice quality of the embedded vowels. When the first two harmonics of the vowels were examined, the vowels of Masan dialect showed characteristics of a creaky voice compared to those of Seoul dialect. The vowels of Seoul dialect were breathy. For the current work, the voice quality difference was not employed in the simulation of the Masan dialect.

It is the aim of this study to examine the degree to which known and unknown segmental differences of Korean regional dialects, i.e. Masan and Seoul dialects, have in the dialect perception and to test the viability of simulating Masan dialect with the speech segments of Seoul dialect. The hypothesis of this study is that, at the level of phonetics involving non-phonemic differences, it is not the segmental differences, but the prosodic differences among Korean regional dialects that play a major role in authentic dialect perception. If the hypothesis is supported by this work, a multi-dialect concatenative synthesizer could be built whose concatenation units are uni-dialectal. If not, more resources, e.g. different concatenation units for different dialects, would be required to build such a multi-dialectal concatenative synthesizer.

## II. Methods

Since Korean regional dialects are known to vary with respect to their phonological, morphological, syntactic, and prosodic aspects, 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. With the help of a native speaker of Masan and a native speaker of Seoul, a set of sample "neutral" sentences were created as given in <Table 1>.

**교수님 가시는 길이 구미로 곧장 갑니까?**
(ㄱ,ㅅ,ㄴ,ㅁ,ㄹ,ㅉ,ㅇ,ㄲ)
**동대구에 볼 일이 없습니다.**
(ㄷ,ㅇ,ㄱ,ㅂ,ㄹ,ㅆ,ㅁ,ㄴ)
**바다에 보물섬이 없다.**
(ㅂ,ㄷ,ㅁ,ㄹ,ㅆ,ㅉ)
**바람이 불어서 먼지가 많다.**
(ㅂ,ㄹ,ㅁ,ㅅ,ㄴ,ㅈ,ㄱ,ㅌ)
**서울에 사는 삼촌이 왔다.**
(ㅅ,ㄹ,ㄴ,ㅁ,ㅊ,ㅉ)
**쌀 사고 난 후에 와라.**
(ㅆ,ㄹ,ㅅ,ㄱ,ㄴ,ㅎ)
**싸기는 해 보여도, 비싸기는 마찬가지다.**
(ㅆ,ㄱ,ㄴ,ㅎ,ㅂ,ㄷ,ㅁ,ㅊ,ㅈ)

<Table 1>. Sample sentences used in the experiment. Segments in parentheses indicate the phones that appear in the sentences.

The sentences in <Table 1> contain all the phones of the two dialects except for /ㅋ, ㅛ, ㅞ, ㅉ/. These sentences were assumed to contain the phonological, morphological, and syntactic components that are shared by the two dialects. That is, when uttered by the speakers of the two dialects, these sentences were different only in terms of their phonetic and prosodic aspects. The prosodic aspects of the Seoul utterances were replaced with those of the Masan utterances. The Seoul utterances ended up having all the prosodic aspects of the Masan dialect except for their non-phonemic or phonetic segmental composition.

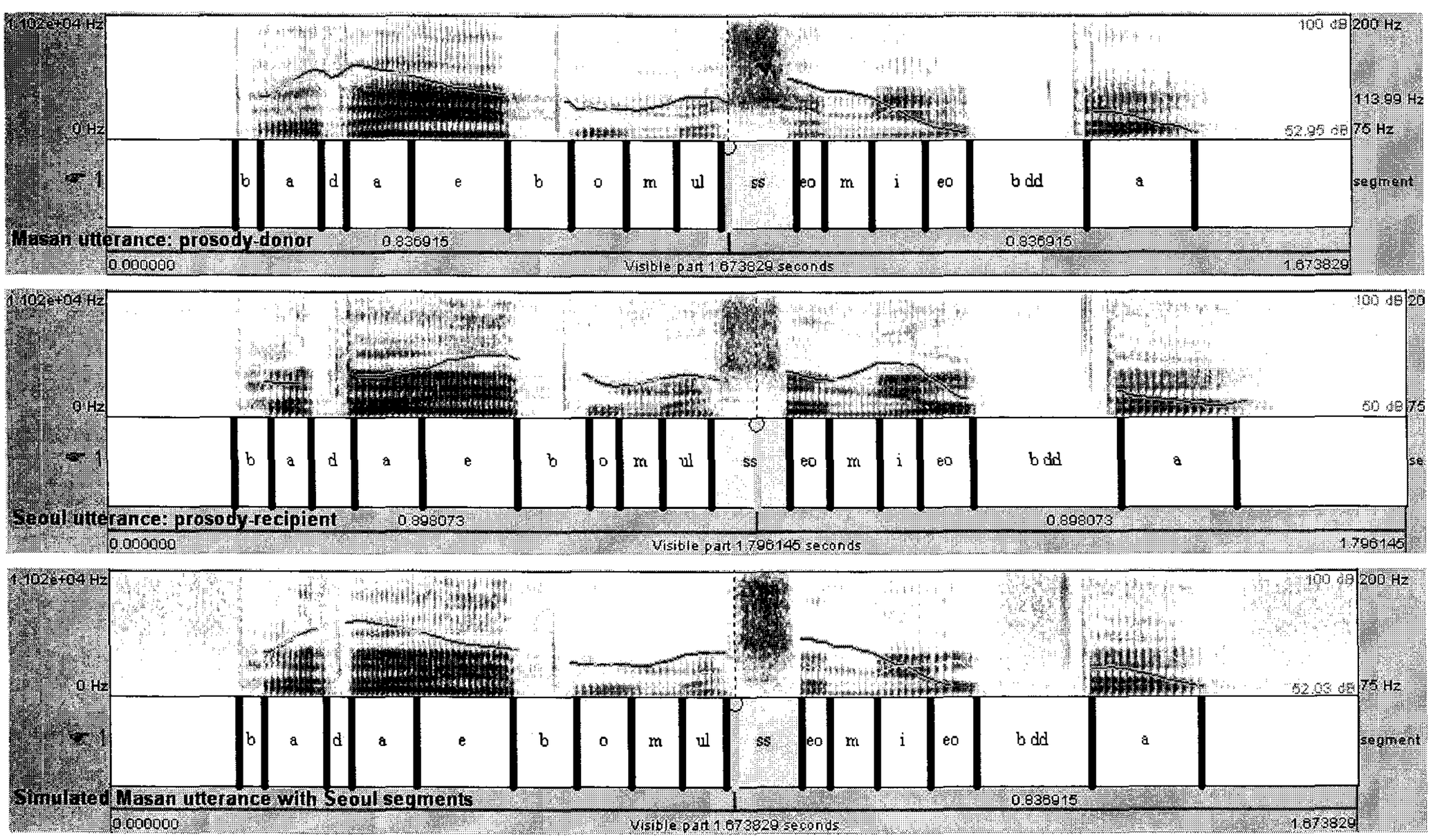
Two male speakers of Masan dialect and two male speakers of Seoul dialect participated in the experiment. All of them produced the seven sample sentences, which were sampled at 22kHz. They were asked to say them as natural as possible in a quiet room. In order to test the simulated Masan utterances, one speaker from the Masan dialect served as the "prosody donor". The other Masan speaker and one Seoul speaker served as the "prosody recipient". The utterances from the other Masan speaker acted as the control group. The control Masan utterances and the Seoul utterances "received" the authentic Masan prosody from the "prosody donor" by the technique introduced in [3]. After the application of the technique, two sets of test stimuli were prepared. One set, i.e. the control group, contained the segments of the Masan prosody recipient and the prosody of the Masan prosody donor and the other contained the segments of the Seoul speaker and the prosody of the Masan prosody donor.

The first control set here represents the imaginary output from a multi-dialect speech synthesizer whose concatenation units are many, one

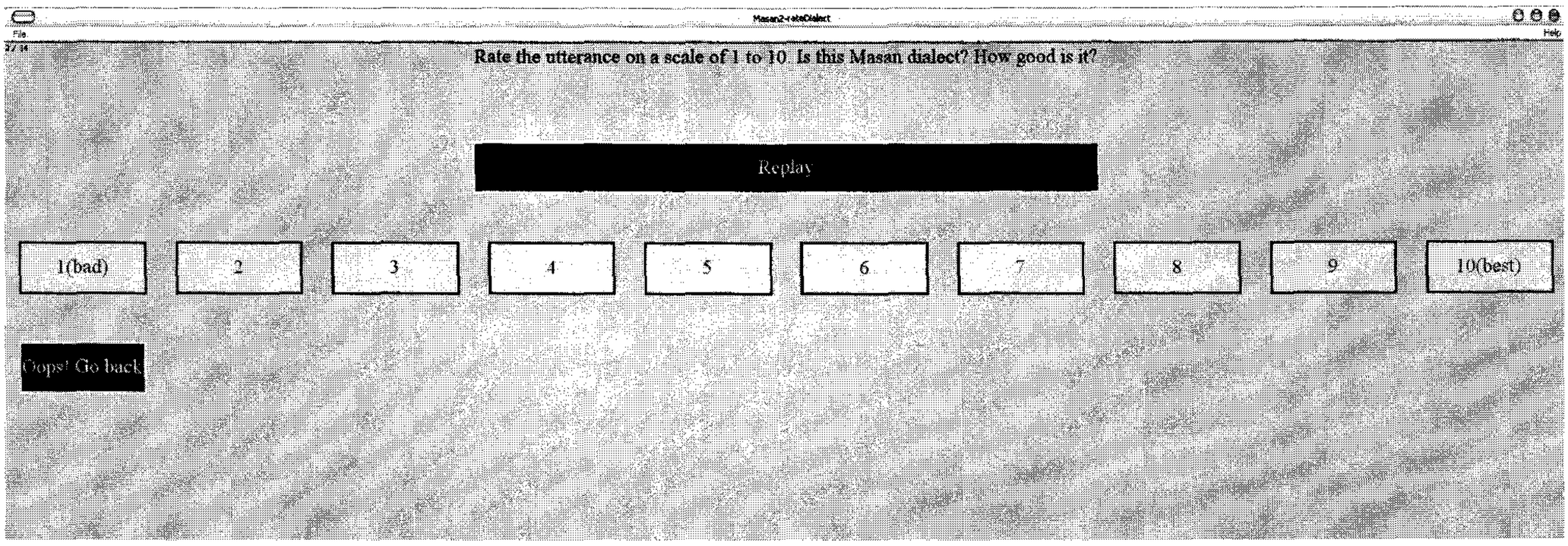
of which is from a Masan speaker. The second test set represents the imaginary output from a multi-dialect speech synthesizer whose concatenation units are single, i.e. from a Seoul speaker. Of course the prosody module of either system is assumed to have assigned authentic Masan prosody to the output.

In order to apply the prosody-swapping technique, the utterances from the two dialects were first manually segmented in Praat as shown in <Figure 1>. As shown in <Figure 1>, the three aspects of the prosodic features manipulated in this study are different in both utterances, i.e. the middle and bottom panels. Note the different segmental durations, the fundamental frequency (F0) contour (blue thicker line), and the intensity contour (yellow thinner line) between the upper and middle panels. After the application, however, the two utterances, i.e. the top and bottom panels, are the same in the three prosodic features (See <Figure 2>). The simulated Masan utterance in the bottom panel has the same segmental durations, the F0 contour, and the intensity contours those of the Masan prosody donor in the top panel.

The control set composed of seven Masan utterances and the test set composed of seven simulated Masan utterances were presented to four normal-hearing listeners of Masan dialect. Three of them were female. The fourteen stimuli were randomized using the *PermuteBalancedNoDoublets* randomization strategy implemented in Praat ExperimentMFC object. For each stimulus played over a headphone, the listeners were asked to rate the utterances on a scale of 1 (bad) to 10 (best) and choose the corresponding button on the computer screen by clicking on it. The listener was allowed to replay the stimulus up to ten times before making a decision and to go back to the previous stimulus if a mistake was made. A sample computer screen is given in <Figure 2>. The responses were analyzed qualitatively for each listener.




<Figure 1> Creation of a simulated Masan utterance with Seoul segments. Upper panel represents the Masan prosody donor. Middle panel represents the Seoul prosody recipient. Bottom panel represents the simulated Masan utterance with Seoul segments.




<Figure 2> A sample screen from the listening experiment.

## III. Results

The results of the listening experiment are given in <Table 2>. The upper table shows the responses for the control group. Recall that the stimuli in the control group have the Masan segments and the Masan prosody, but from different Masan speakers. The lower table shows the responses for the test stimuli, which have the Seoul segments but their prosody comes from the same Masan speaker that was used to create the control stimuli.

Since there were only four listeners in the experiment, no statistical analyses were performed. From the qualitative point of view, listeners generally gave better scores for the control group, i.e. on a scale of 1 (bad) to 10 (best), 7.07 points for the control group and 6.43 points for the test group. This means that the listeners generally preferred the control stimuli, which may suggest that the segmental properties of the Seoul segments in the test stimuli prevented the Masan listeners from perceiving the simulated Masan stimuli as being authentic Masan utterances.

However, if we compare the responses one by one, we get a different picture. Although listeners generally identified the control stimuli as being more authentic Masan dialect than the test stimuli, three out of seven test stimuli were identified as being more authentic Masan utterance than those in the control group. The second, fourth, and fifth test stimuli have better scores than the corresponding control stimuli, i.e. 8, 4, and 8.5 points average as compared to 7, 2.75, and 8.25 points. There were three test stimuli that Listener 2 liked better than their corresponding control stimuli. Listener 3 had one and Listener 4 had four such cases. These cases suggest that swapping the prosodic features can be a viable option in the creation of Masan utterances with the speech segments of Seoul dialect.

As suggested earlier, Masan utterances may have a voice quality different from that of Seoul utterances. After the listening experiment, the four Masan listeners were asked about what they liked and did not like about the test stimuli. Some of them mentioned the inappropriateness of some of the vocabularies used in the test sentences. They were not able to point out any one particular feature of the test utterances. When the voice source of a Masan utterance was transferred to its corresponding test utterance, the overall quality of the test utterance appeared to get better. The voice source was splitted from the control sound file by using the LPC synthesis technique implemented in Praat. A future experiment involving the swapping of the voice source as well as the three prosodic features used in the current work is expected to reveal the influence of the voice source in the perception of Korean regional dialects.

| Test Group | Listener 1 | Listener 2 | Listener 3 | Listener 4 | Average |
|---|---|---|---|---|---|
| 싸기는 해 보여도 비싸기는 마찬가지다 | 1 | *6* | 9 | 1 | 4.25 |
| 동대구에 볼 일이 없습니다. | 5 | 8 | 10 | *9* | *8* |
| 교수님 가시는 길이 구미로 곧장 갑니까? | 10 | 7 | 10 | *10* | 9.25 |
| 쌀 사고 난 후에 와라 | 1 | *8* | 4 | *3* | *4* |
| 바다에 보물섬이 없다 | 10 | *10* | *10* | 4 | *8.5* |
| 서울에 사는 삼촌이 왔다 | 8 | 7 | 4 | 6 | 6.25 |
| 바람이 불어서 먼지가 많다 | 5 | 7 | 1 | *6* | 4.75 |
| **Average** | 5.71 | 7.57 | 6.86 | 5.57 | 6.43 |

<Table 2> Responses from the listening experiment. Upper table is for the control group and lower table is for the test group. Listeners gave a score to each stimulus on a scale of 1 (bad) to 10 (best).

## Ⅳ. Conclusion

This paper examined the viability of simulating Masan dialect with the speech segments of Seoul dialect through prosody transfer. The hypothesis was that, at the level of phonetics, it is not the segmental differences, but the prosodic differences among the two Korean regional dialects that play a major role in authentic dialect perception. Masan dialect was simulated by transferring the prosodic features of the Masan utterances to the Seoul utterances. The prosodic features involved were the fundamental frequency contour, the segmental durations, and the intensity contour. The simulated Masan utterances composed of Seoul segments were evaluated by four native listeners of Masan dialect and some of the simulated Masan utterances were rated better than their corresponding control Masan utterances. Although further experiments are necessary to fully assess the plausibility of simulating Masan dialect with Seoul segments, the results of the experiment suggest that it may be viable to create a multi-dialect concatenation-based speech synthesizer whose concatenation units are from a single dialect.

## References


[1] Kyung-Hee Lee, "Comparison of acoustic characteristics between Seoul and Busan dialect on fricatives", Speech Sciences, Vol.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, Vol.9/2, pp.25-47, 2002.

[3] Kyuchul Yoon, "Swapping native and non-native speakers' prosody using PSOLA algorithm", Proceedings of the Korean Society of Phonetic Sciences and Speech Technology, Spring Conference, pp.77-81, 2006.

[4] E. Moulines and F. Charpentier, "Pitch synchronouswaveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, 9:n 5-6, 1990.

[5] P. Boersma, "Praat, a system for doing phonetics by computer", Glot International, Vol.5, 9/10, pp.341-345, 2005.