The contribution of silent pauses to the perception of prosodic boundaries in Korean read speech.

Hyongsil Cho & Daniel Hirst

Laboratoire Parole et Langage CNRS UMR 6057
Université de Provence, Aix en Provence
{hyongsil.cho; daniel.hirst@lpl.univ-aix.fr}

Abstract

This paper discusses the importance of silent pauses in the perception of prosodic boundaries in Korean speech. It is suggested that in speech in general, and in particular in spontaneous speech, silent pauses are neither necessary nor sufficient for the perception of prosodic boundaries. In read speech, however, there is a high correlation between the presence of a pause and the perception of a boundary. An experiment was carried out to determine whether removing the silent boundary from an extract of speech had a significant effect on the perception of boundaries in Korean read speech. Results suggest that while the presence of a silent boundary slightly reinforces the perception of a prosodic boundary, subjects are in general capable of perceiving the boundary without the silent pause.

1. Introduction

The study of prosody relies crucially on judgments of acceptability. The empirical basis for such judgments is, however, notoriously difficult to establish. Yet such evidence is clearly necessary in order to avoid circularity. Even judgments by trained listeners, such as those used in the ToBI paradigm [1], are not above all suspicion. It is clear that before empirical studies from large oral databases can be carried out, it will be necessary to establish automatic or semi-automatic techniques of empirical annotation.

Work in the area of speech synthesis and automatic speech recognition has, in the last few decades, brought to light a number of acoustic cues which can be shown to correlate with judgments of prosodic boundaries. Yet, once again, the problem arises of the validation of these judgments. Asking listeners explictly to indicate prosodic boundaries amounts to eliciting meta-linguistic judgments. As is well known, such meta-linguistic judgments are subject to caution and highly dependent on the listeners' conception of the task at hand and on the theory-dependent training which they have been given. Ideally, we should like to be able to validate acoustic models of prosody by evidence obtained directly from untrained subjects.

One of the most obvious phonetic cues, in all languages, for the presence of a prosodic boundary is the presence of a silent pause. In read speech, the correlation between the presence of a pause and the perception of a boundary is extremely high. In spontaneous speech, however, the situation is less straightforward since while a silent pause can accompany a prosodic boundary it can also be a sign of a hesitation. Since listeners in general do not confuse hesitation pauses with prosodic boundaries, this suggests that silent pauses are not in themselves sufficient cues for these boundaries.

This can easily be demonstrated, by taking a recording of speech and editing the signal, removing any silent pauses. We carried out this operation on 40 continuous passages of read speech taken from the Eurom1 corpus [3] for five different languages (English, French, Italian, Russian and Japanese) using the Praat software [2]. Each passage consists of about five sentences constituting a semantically coherent block. Informal reports of listening to these edited recordings by native speakers convinced us that the removal of silent pauses did not seriously affect the perception of prosodic boundaries. The reported impression was generally one of hurried speech. Only very occasionally did the edited recording give the impression of being unnatural, when for example there was a sudden change of pitch height from the syllable preceding the removed pause to the syllable following it.

These facts suggest, then, that the presence of a silent pause is neither a necessary nor a sufficient condition for the identification of a prosodic boundary. Yet, in the case of read speech, as we mentioned above, the presence of a silent boundary can be taken as an uncontroversial sign of the presence of a prosodic boundary. In the rest of this paper we present preliminary results testing the contribution of the silent pauses themselves to the perception of these boundaries in Korean read speech.

2. Boundaries and silent pauses in Korean

2.1. Prosodic units and boundaries in Korean language

For the hierarchy of prosodic units in Korean, the framework of K-ToBI, based on intonational phonology, has been quite widely adopted assuming a hierarchical phonological structure as illustrated in Figure 1. Here, an Accentual Phrase (AP) is smaller than an Intonational Phrase (IP) and larger than a phonological word (W), a lexical item plus a case marker or postposition. An IP is marked by a boundary tone (%) and final lengthening. An AP is marked by a phrase tone, THLH (T=H if the AP initial segment is aspirated or tense, T=L otherwise), but not by final lengthening [5], [6].

Besides this phonological view, sharing the tradition of articulary and perceptual phonetics, some acoustic phonetic studies [12], [13] assume a more acoustic unit, the prosodic phrase. The prosodic phrase is a unit between two clearly perceptible breaks. It is composed of rhythmic units which themselves each contain one strong stressed syllable together with other weak syllables,. Acoustic features in the perception of such a break are discussed in the next chapter.

Figure 1. The intonational structure of Seoul Korean as described in the K-ToBI framework

2.2. Pauses and the perception of boundaries

In its most common definition, a boundary means a perceptual break produced between two syntactic or semantic units, relying on the speaker's breath. The break is usually accompanied by a sudden change of acoustic features (pitch range, vowel quality and quantity) and/or by the presence of a pause. A pause usually corresponds to an acoustic silence although for some authors it is used as a synonym of the break. In this paper we use the term pause to refer specifically to the presence of an acoustic silence.

For the Korean language, several studies have identified the silent pause, together with pitch change, and segmental lengthening, as the major acoustic cues for the perception and identification of prosodic boundaries [7], [9], [10], [11].

In a study on the interaction of acoustic cues formulated with the ToBI framework [7], the silent pause is seen as providing a dominant cue for the perception of breaks and as the decisive one when the pause is longer than 200 ms. It is then confirmed that long or very long silent intervals takes an significant role with pitch change in the major boundary perception.

As far as the duration of silent pauses is concerned, there seems to be a general consensus that an acoustic break longer than 200ms is generally perceived as a silent pause and that the higher the boundary, the longer this silent interval will be. This viewpoint is applied to the automatic treatment of corpora for speech synthesis where pause duration is taken as a correlate of the boundary degree. [8]

2.3. Syntactic structure and boundaries

Since prosodic units correlate with syntactic units, we also need to take into account the role of morpho-syntactic information in the perception of boundaries. An experimental phonetic study [13] showed that within a sentence, some sequences, such as *predicate* + *subject*, *predicate* + *object* or *predicate* + *adverb*, are more likely to be separated by a major prosodic boundary accompanied by a silent pause than others.

An interesting characteristic of the Korean language is the fact that each part of speech in a sentence is identified by one or more grammatical morphemes, making it relatively easy to identify a syntactic boundary, which is also a (potential) prosodic boundary.

Korean, an agglutinative language, possesses a system of particles and endings as markers of part of speech. A noun, for example, needs to be accompanied by a case-marking particle: subjective, objective, genitive, etc -. A verb, in the same way, needs to be accompanied by endings giving information on the tense, the intention or style of the speaker. There are also verbal endings which let us know if the verb is located at the end of a clause within a sentence or at the end of the whole sentence.

In the sentence

해가 지고 달이 뜬다

[The sun goes down and the moon rises.]

each graphic word is syntactically analyzed as:

- 해가 : Subject [Noun + Subject Particle]
- 지고 : Predicate [Verb + Coordinating clause ending]
- 달이 : Subject [Noun + Subject Particle]
- 뜬다 : Predicate [Verb + Declarative ending]

This structure allows us to segment the sentence as [[해가] [지고]] [[달이] [뜬다]] where a pause will normally be located after "고", a clause ending.

In other words, the end of a sentence, of a clause or of any other part of speech is morphologically marked in Korean by the presence of verbal endings or other types of markers. This idea is used in TTS for prosodic phrase extraction and pause prediction [11].

An important consequence for our study of this characteristic of the Korean language is the fact that it is not always obvious whether morpho-syntactic cues or acoustic cues are responsible when a sentence boundary has been identified by a listener.

3. Testing the contribution of silent pauses to the perception of prosodic boundaries

3.1. Material and subjects

The forty continuous passages from the Eurom1 corpus described in §1 above were freely translated into Korean by the first author. They were then recorded by the same author in an anechoic chamber and digitised. The recordings were annotated manually on the level of word and phoneme using the Praat software[2].

From the forty recorded passages we selected 42 short extracts of about 2 seconds duration each. Half of these consisted of speech material containing a silent pause (the duration of the silent pause was not counted in the duration of the extract), the other half containing no silent pause and no obvious prosodic boundary. The first group of these extracts was then edited, removing the silent pauses. The recordings were subsequently grouped into two blocks, A and B. Block A contained the 21 extracts with no silent pauses, plus 10 extracts originally containing pauses but with the pauses removed together with 11 extracts containing the extracts with the silent pauses not removed. Block B contained the same 21 extracts with no silent pauses, the 10 remaining extracts containing silent pauses and the 11 remaining extracts from

which the pauses had been removed. Each block thus contained all the extracts with no pauses and all the extracts containing pauses - approximately half of which had been edited to remove the acoustic silence.

In each block the order of the stimuli was randomised.

In all of the extracts containing a silent pause, the duration of the pause was considerably more than 200 ms, ranging from 348 ms to 814. The distribution of the pauses by duration is shown in the histogram of Figure 2.

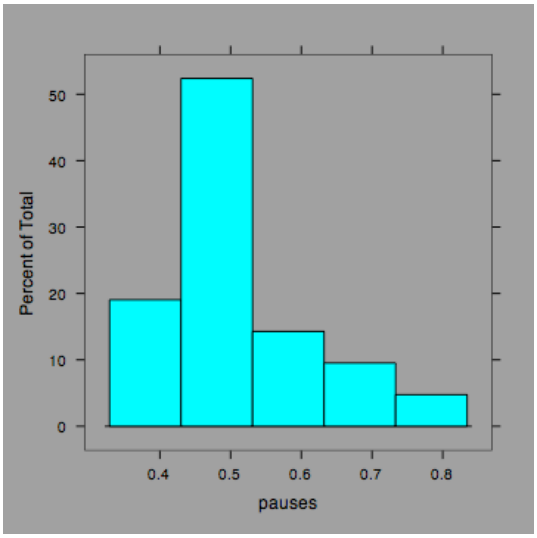

Figure 2. Distribution of pauses in the selected extracts. Duration in seconds.

We decided to present the recordings to native speakers and to elicit judgments on the presence of absence of a prosodic boundary. Ten native speakers of Korean were chosen to take part in this test. None of them reported any hearing impairment.

Owing to the difficulty of obtaining reliable metalinguistic judgements, as discussed in section 1 above, we decided to present listeners with a forced choice asking whether the extract which they heard was taken from one sentence or from two different sentences.

Subjects were played extracts from the following artificial examples

다시 방으로 들어갔다. 나와 엄마는 청소를 했다.

[(we) went back into the bedroom. Mother and I did the housework.]

다시 방으로 들어갔다 나와 엄마는 청소를 했다.

[(we) went back into the bedroom and came out again, (and then) mother did the housework.]

where the choice of prosodic boundary determines the interpretation. They were told that the first extract was part of one single sentence whereas the second extract was taken from two different sentences. We deliberately avoided mentioning boundaries or pauses.

We noted above that in Korean, prosodic boundaries are very often accompanied by specific particles. This makes it difficult to be sure that a listener's identification of a boundary is relying on prosodic rather than on segmental cues.

To ensure that listeners were not using segmental cues we decided to filter the speech extracts to make them unintelligible. After experimentation, we used a low pass filter of 600Hz with the Praat command Filter... (formula) and the value:

```
if x>600 then 0 else self fi;
```

The resulting recordings were unintelligible but the prosody appeared sufficiently audible.

The stimuli were presented in one continuous sound file separated by a silence of 5 seconds. Each stimulus was preceded by a beep and a silence of 500ms. Every fifth stimulus was preceded by two beeps to help subjects to follow the test. Subjects were requested to listen to the complete recording without stopping and to answer "one" or "two" for each stimulus depending on whether they identified the stimulus as part of one sentence or of two. They were then permitted to listen to the recordings a second time and to correct their initial answers if they wished.

4. Results

Since the responses were dichotomous (*one* or *two*) we fitted a logistic regression (linear logit model) to the data with the independent variables **block** (*A* or *B*) and **category** (*pause*, *none* or *removed*). The effect of block was non-significant ($p = 0.43$) but there was a small significant interaction ($p = 0.01$) between **block** and **category**. This was almost certainly due to differences among the subjects. Inspection of individual responses showed a certain variability: in fact one subject identified the boundary for all the stimuli belonging to the category *removed* whereas one other subject did not identify any.

Across subjects, the following histogram shows that most of the stimuli belonging to the category *removed* were in fact identified as containing a boundary.

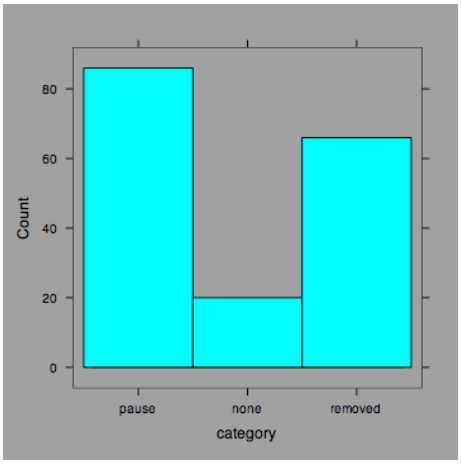

Figure 3: Percentage of stimuli identified as containing a prosodic boundary (i.e. response = "two") by category.

Ignoring the effect of block, the logistic regression of **category** on **response** showed that the difference between categories *none* and *pause* was highly significant ($p < 2e\text{-}16$) and that the difference between *none* and *removed* was also significant but less so ($p = 0.0024$). It turned out, in fact,that this latter difference was almost entirely due to the responses of one subject, who, as mentioned above, did not identify any of the stimuli of the category *removed* as containing a boundary. Without this subject's responses, the significance of

the difference between *none* and *pause* remains at the same value whereas that between *none* and *pause* falls to non-significance (p = 0.0724).

Figure 4 shows the number of times for each subject that a stimulus belonging to the category *removed* was identified as containing a prosodic boundary.

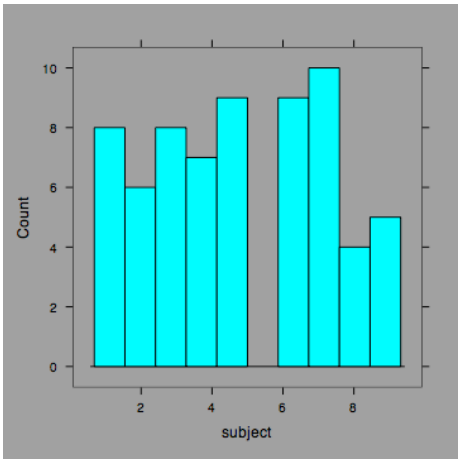

Figure 4. Counts per subject of stimuli of category *removed* identified as containing a prosodic boundary.

5. Conclusions

This preliminary study demonstrates that when silent pauses have been edited out of an utterance, speakers of Korean are generally capable of identifying a prosodic boundary, even when the speech has been filtered to render it unintelligible. For most of the listeners the results with the edited stimuli were identified almost as well as the unedited stimuli, and for one subject just as well. One of the ten subjects, however, clearly relied entirely on the presence of a silent pause and for most subjects the percentage of prosodic boundaries identified was slightly lower for the edited stimuli than for the non-edited ones.

In further work it is planned to extend this experimental procedure to a larger number of subjects and to other languages. It would also be interesting to perform the same task with non-native listeners.

Acknowledgements

We should like to thank Robert Espesser for his help with the statistical analysis.

6. References

[1] Beckman, Mary & Gayle Ayers 1994. *Guidelines for ToBI Labelling*. Unpublished ms. Ohio State University. Version 3. March 1997. [http://ling.ohio-state.edu/Phonetics/etobi_homepage.html].

[2] Boersma, P., & Weenink, D. 2005. Praat: doing phonetics by computer. (Version 4.4) [computer program] retrieved December 22, 2005 from: htttp://www.fon.hum.uva.nl/praat/

[3] Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., & Zeiliger, J. 1995. EUROM: a spoken language resource for the EU. *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech '95*, (Madrid) 1, 867-880.

[4] Fox, John 2002 *An R and S-Plus Companion to Applied Regression*. Sage Publications, London.

[5] Jun, Sun-Ah 1993. *The phonetics and phonology of Korean Prosody*. PhD dissertation, Ohio State University.

[6] Jun, Sun-Ah. 2000. K-ToBI Labelling conventions. http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html (UCLA.)

[7] Kim Hyo Sook, Kim Chung Won, Kim Sun Ju, Kim Seoncheol, Kim Sam Jin, Kwon Chul Hong 2002. Prediction of Break Indices in Korean Read Speech. *Malsori, Jounal of The Korean Society of Phonetic Sciences and Speech Technology*

[8] Kim Sang-hun, Park Jun, Lee young-jik 2001. Corpus-based Korean Text-to-speech Conversion System, *The Journal of the Acoustical Society of Korea*, 4 v.020, n.003

[9] Ko Hyun-ju, Kim Sang-hun, Lee Sook-hyang 2001. Cues to phrase boundary perception in the Korean read speech. *Proceedings of International Conference on Speech Processing*, vol. 2, 853-857, 2001

[10] Lee Chan-Do 1997. A computational study of prosodic structures of Korean for speech recognition and synthesis: Predicting phonological boundaries. *Journal of Korean Information Processing Society*

[11] Lee Jong-Ju, Lee Sook-Hyang 1996 On phonetic characteristics of pause in the Korean read speech. *Proceedings of ICSLP '96*,

[12] Lee Sangho, Oh Yung-Hwan 1998. The modeling of prosodic phrasing and pause duration using CART. *KSCSP* '98

[13] Seong Cheol-Jae 1996. An experimental phonetic study of the correlation between prosodic phrase and syntactic structure. *Journal of The Linguistic Association of Korea* Vol.18