Korean MULTEXT: A Korean Prosody Corpus

[1]*Sunhee Kim,* [2]*Daniel Hirst,* [2]*Hyongsil Cho,* [3]*Ho-Young Lee,* [3]*Minhwa Chung*

[1]Center for Humanities and Information, Seoul National University, Korea
[2]Laboratoire Parole et Langage, CNRS/Aix-Marseille Université, France
[3]Department of Linguistics, Seoul National University, Korea
{sunhkim, hylee, mchung}@snu.ac.kr, {daniel.hirst, hyongsil.cho}@lpl-aix.fr

Abstract

This paper describes the contents of the Korean prosody corpus (Korean MULTEXT), which is a Korean version of the speech database Eurom1. The corpus consists of about 2 hours of read speech, transcribed primarily in orthography (in Korean alphabet and in a Romanized transcription), in IPA and in SAMPA. Furthermore, it includes the original F0 values, stylized F0 values extracted using Momel, and hand-corrected F0 values. The prosodic events are annotated in two ways. They are annotated with the automatic annotation algorithm, INTSINT, and also labeled manually into prosodic units with two tones on the hand-corrected pitch targets. It is found that the resulting tone patterns from the proposed Momel-based two tone labeling correspond to those defined in K-ToBI.

1. Introduction

Despite the prevailing understanding that prosodic information plays a crucial role in speech synthesis and speech recognition, its application to those areas in the Korean language has been quite limited due to the lack of prosody corpora with proper annotations of relevant information. This paper describes the contents of the Korean prosody corpus (Korean MULTEXT), which is a Korean version of the speech database Eurom1[1]. MULTEXT[2] is a prosody corpus of a given language with annotations of prosodic parameters, developed to support speech synthesis and speech recognition technologies.

The corpus consists of about two hours of read speech. They are transcribed in various symbols: the Korean alphabet, a Romanized transcription, IPA and SAMPA. For the prosodic annotation, an automatic algorithm of pitch stylization and a prosodic annotation system, Momel and INTSINT are used[3, 4, 5]. By using Momel[3, 4], the pitch targets for the original and stylized F0 values are extracted. Then the stylized curves are manually corrected, so that the F0 values of the hand-corrected pitch targets are provided along with those extracted by Momel in the corpus. The prosodic events are annotated in two ways. First, they are annotated with the automatic annotation algorithm, INTSINT[4, 5]. Second, they were manually labeled into prosodic units with two tone labels (H and L) on the pitch targets obtained with Momel. The resulting tone patterns are compared to those defined in K-ToBI[11], which is known to be a standard intonation model of Korean.

The development and description of the present corpus is meant to contribute to the study of Korean prosody as well as to the development of Korean speech systems with prosody information.

2. Korean prosody corpus

The corpus consists of about two hours of Korean read speech. The original English version of the Eurom1 text was translated into Korean. It is composed of 40 passages (168 sentences). The texts in Korean alphabet are Romanized and also transcribed in SAMPA and IPA.

Speech was recorded in an anechoic room using a Shure SM-58 microphone with a digital recorder, Marantz PMD 670, with 16000 Hz of sampling frequency on 16 bits.

10 Seoul speakers (5 male and 5 female) took part in the recording session. They were all Korean native speakers in their twenties, either undergraduate or graduate students of Seoul National University. Each speaker read all 40 passages.

3. Prosody annotation

3.1. Phoneme labels

The annotation of phonemes was carried using Praat[6] Textgrid files. Each sound file was segmented into sentences, words and phonemes. An automatic segmentation of phonemes, transcribed with SAMPA, was carried out using Mbrolign. Romanized words and sentences were aligned respectively on each tier of the same Textgrid file containing phoneme labels. Figure 1 shows an example of segmentation into sentences, words and phonemes.

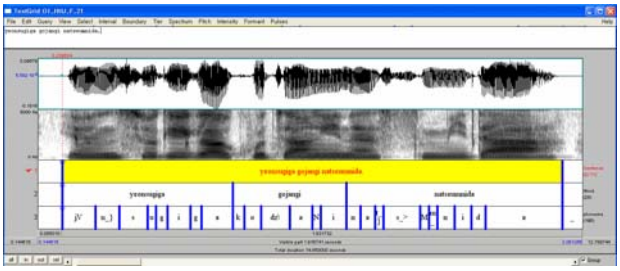

Figure 1: *An example of segmentation with phonemes, words and sentences on a Textgrid file of Praat*

3.2. F0 curves and F0 values

The MOMEL (MOdeling MELody) algorithm proposes a method of automatic stylization of F0 as a sequence of target points by means of a quadratic spline function [3, 4]. Given that F0 variations are considered as the superposition of two components, a microprosodic component, corresponding to local variations of pitch caused by the phonetic nature of the speech segments, and a macroprosodic component corresponding to the overall pitch pattern of the utterance, the Momel algorithm enables to represent the macroprosodic

component as a sequence of pitch targets. A stylized F0 curve is obtained by connecting these pitch targets, which correspond to the linguistically significant points. The F0 values of pitch targets, corresponding to those extracted by Momel, are provided along with the original F0 values.

3.2.1. F0: original vs. predicted vs. hand-corrected

Using an earlier version of the Momel algorithm (Momel1), the original and predicted F0 values of each target point were extracted. With Praat, the correction of the MOMEL pitch targets was carried out manually by listening to each sentence of the passage, so that the hand-corrected F0s were obtained. During the correction procedure, pitch target(s) could be deleted or moved, or new pitch target(s) could be added. The most frequently observed errors and their corrections at this stage are as follows:

- Excessive pitch targets were derived before and after pauses. → Delete these.
- Rising-falling or falling-rising curves were represented as one pitch target, in particular for male speakers with a limited range of pitch. → Move the pitch target according to the contour and add a pitch target.
- Level tones are represented as one pitch target. → Move the pitch target according to the contour and add a pitch target.

3.2.2. F0 obtained using an upgraded Momel (Momel2)

In order to reduce the first type of systematic errors among the above mentioned ones, which are generally found before and after pauses, [7] proposed an upgraded version of Momel (Momel2). Figure 2 is an example of a passage (taken from the French version of the Eurom1 corpus) where a rising pitch before a pause is completely missed by the algorithm, whereas Figure 3 shows the result of using the upgraded version of Momel on the same passage, showing a final concave rise extrapolated with the closest target point that would produce such a rise.

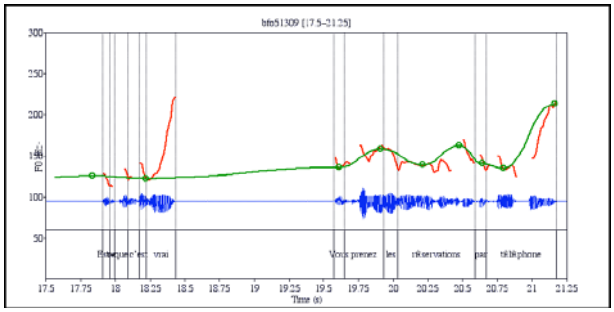

Figure 2: *Raw (red) and modeled (green) fundamental frequency for the extract "Est-ce que c'est vrai? Vous prenez les réservations par téléphone?" (Is it true? You take bookings by phone?) using the original version of Momel*

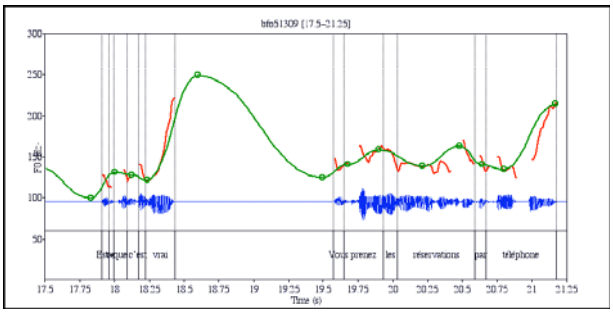

Figure 3: *Raw (red) and modeled (green) fundamental frequency for the extract "Est-ce que c'est vrai? Vous prenez les réservations par téléphone?" (Is it true? You take bookings by phone?) using the new version of Momel.*

Thus, the F0 values derived by the application of the upgraded version of Momel (Momel2) are provided along with three above mentioned F0s: original, predicted by Momel1 and hand-corrected F0s. As in Figure 4, the pitch targets of each curve are represented in PitchTier files of Praat.

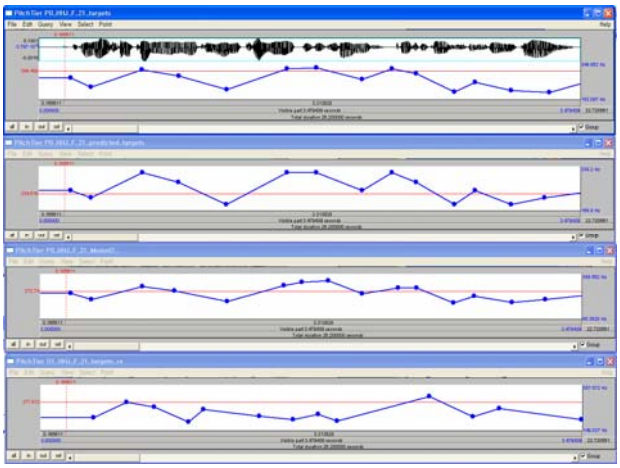

Figure 4: *From the top, the curves of pitch targets of original F0s, predicted F0s by Momel1, predicted F0s by Momel 2 and hand-corrected F0s.*

3.3. Prosody labels

3.3.1. INTSINT labeling

INTSINT[4, 5] is a prosodic annotation system which was intended as a first approximation to a prosodic equivalent of the IPA, reducing target points to "phonological-like" symbols. The system represents target points as values either globally defined relative to the speaker's pitch range: Top (T), Mid (M) and Bottom (B), or locally defined relative to the previous target-point. Relative target-points can be classified as Higher (H), Same (S) or Lower (L) with respect to the previous target. A further category of locally defined target points relative to the previous target-point consists of smaller pitch changes such as Upstepped (U) and Downstepped (D). The transcription in INTSINT symbols are easily convertible to and from a sequence of target points by means of a Perl script[8]. The implementation for a French text-to-speech

system is reported in [9, 10]. The prosodic events without distinction of prosodic units are annotated using INTSINT because the system does not account for the prosodic units.

Tone labeling was carried out on the hand-corrected target points. Figure 5 shows an example of the target points converted into INTSINT symbols on Praat.

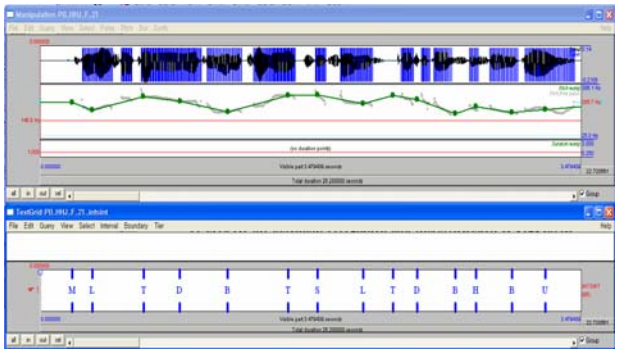

Figure 5: *An example of the target points converted into INTSINT symbols*

3.3.2. Momel- based two-tone labeling

Like earlier research on Korean prosody [11, 12], we assume that Korean prosody consists of two hierarchical prosodic units: AP (Accentual phrase) and IP (Intonation Phrase), where the AP is a sub-unit of an IP. In this corpus, AP and IP boundaries were annotated by three trained linguists by adding a tier under the phoneme labels of the Textgrid file as in Figure 1. When each passage was divided into AP and IP, the tone of each pitch target was labeled either H or L depending on the relative height of the targets within the prosodic unit.

As in Figure 6, when there are three targets in an AP, the highest target should be labeled as H and the other two targets that precede or follow the highest target should be labeled L, so as to derive an AP with a /LHL/ pattern. In this way, the tone patterns are "extracted" based on the values of the targets, while, in K-ToBI, the possible tone patterns are defined a priori.

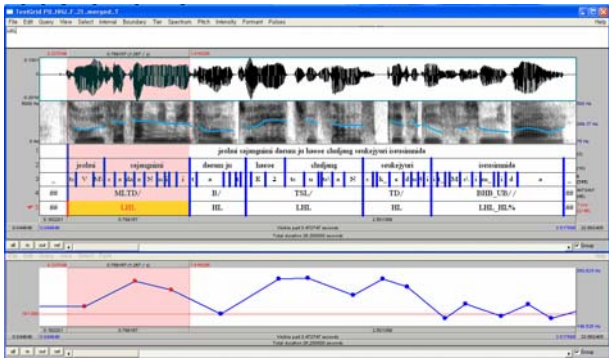

Figure 6: *Tone labeling of an AP*

In order to label IP boundary tones, it is necessary to mark the beginning of the IP-final syllable, and the rest of the tones in the IP are annotated as for the AP.

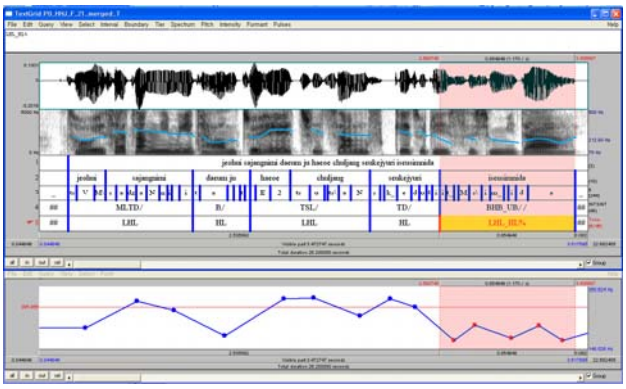

Figure 7: *Tone labeling of an IP*

The study of the tone patterns extracted from the corpus by the proposed Momel-based two-tone labeling compared with those presented in K-ToBI[11] were performed in [13]. The extracted AP patterns are shown in comparison with K-TOBI as in Table 1.

Table 1: *AP tone patterns extracted from the corpus by the proposed Momel-based two-tone labeling in comparison with those defined in K-ToBI*

K-ToBI	Momel-based two-tone labeling	
	AP	Frequency %
H Ha	H	6.3
	HH	5.15
L La	L	2.14
	LL	0.93
L Ha	LH	44.3
L+H Ha	LHH	5.67
L L+Ha	LLH	0.29
H La	HL	9.6
H+H La	HHL	0.81
H L+La	HLL	0.29
H+H L+La	HHLL	0.06
H L+Ha	HLH	3.7
H+H L+Ha	HHLH	0.69
L+H La	LHL	11.8
L+H L+La	LHLL	0.12
	LHHL	0.12
L+H L+Ha	LHLH	7.98
	LHHLH	0.06
Total		100

Table 1 shows that, in addition to all 14 AP tone patterns defined in K-ToBI, /LHLL/ and /LHHLH/ also occur with a very low frequency percentage such as 0.12 and 0.06. And it also shows that the rising tone patterns have higher frequency percentage (62.69%) than the rest, which might support the K-ToBI analysis that the basic phonological AP tone pattern is a rising one (THLH).

Table 2 presents the IP boundary tone patterns extracted from the corpus by the proposed Momel-based two-tone labeling in comparison with those presented in K-ToBI[6].

Table 2: *IP tone patterns extracted from the corpus by the proposed Momel-based two-tone labeling in comparison with those presented in K-ToBI*

K-ToBI	Momel-based two-tone labeling			
	sentence-medial	Frequency %	sentence-final	Frequency %
H%	H%	36.13	H%	15.48
HL%	HL%	46.07	HL%	58.93
L%	L%	14.4	L%	23.21
LH%	LH%	1.57	LH%	0.89
LHL%	LHL%	1.83	LHL%	1.49
HLH%				
LHLH%				
HLHL%				
LHLHL%				
Total				100

Table 2 shows that only 5 IP boundary tones of the 9 defined in K-TBI appear in the corpus. This may be due to the different properties of the corpora: the Korean MULTEXT corpus consists of only read speech whereas the K-TOBI includes different styles of speech such as news broadcasting, movies and dramas.

In Table 1 and Table 2, it is shown that the resulting tone patterns from the proposed Momel-based two tone labeling correspond to those defined in K-Tobi. That is, the proposed method can be used to label the tones once the prosodic boundaries are marked either automatically or manually.

3.4. Summary

The corpus consists of four parts: Speech, Text, F0 curves and F0 values and tone labeling. The structure of the corpus in detail is as follows:

- Speech: raw speech files by speaker
- Scripts: Text
  - Korean alphabet
  - Romanization
  - IPA
  - SAMPA
- F0 curves and F0 values
  - Original
  - Predicted (Stylized)
    - Momel1
    - Momel2
  - Hand-corrected
- Tone labeling
  - INTSINT
  - Momel-based two-tone labeling including prosodic units

4. Conclusions

This paper describes the contents of the Korean prosody corpus, Korean MULTEXT, which is a Korean version of the speech database Eurom1. The prosodic annotation of the corpus is mostly performed with the help of Momel-INISINT. The F0 curves with their values were derived using two Momel versions and corrected manually, so that four types of F0 curves with their values are included in the corpus. Next, the tones of the derived target points are annotated in two ways. First, they are annotated using the INTSINT system. Second, assuming that Korean prosody consists of two hierarchical prosodic units, AP and IP, AP and IP boundaries are marked manually and the tones within AP and IP are annotated with two-tone labels (H and L) based on the hand-corrected pitch targets. The resulting AP and IP tone patterns using the proposed Momel-based two-tone labeling method are similar to those defined in K-ToBI.

A study of the automatic generation of AP/IP boundary detection and tone labeling based on the results of this paper is in progress.

5. Acknowledgements

This work was carried out with the support of the Korean-French Science and Technology Amicable Relationship (STAR) project funded by Egide and the Korean Foundation for International Cooperation of Science and Technology.

6. References

[1] Chan, D. et al (1995) EUROM: a spoken language resource for the EU. *Eurospeech '95*, Madrid. pp.867-880

[2] Campione, E. & Veronis, J. (1998) A multilingual prosodic database. *ICSLP98.* Sydney. pp.3163-3166.

[3] Hirst, D. & Espesser, R. (1993) Automatic modeling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix 15*, pp.71-85.

[4] Hirst, D., Di Cristo, A. & Espesser, R (2000) Levels of representation and levels of analysis for intonation. in M. Horne (ed.) *Prosody: Theory and Experiment.* Kluwer Academic Publishers, Dordrecht. pp.51-87.

[5] Hirst, D.J. & Di Cristo, A. (eds) (1998) *Intonation Systems. A survey of Twenty Languages.* Cambridge, Cambridge University Press.

[6] Boersma, P. & Weenink, D. (since 1995) Praat: doing phonetics by computer [Computer program]. Downloadable from http://www.praat.org/

[7] Hirst, D., Cho, H., Kim, S. & H. Yu. (2007) Evaluating two versions of the Momel pitch modeling algorithm on a corpus of read speech in Korean. *Interspeech 2007*. Antwerp. pp.1649 – 1652.

[8] Hirst, D. (2001) Automatic analysis of prosody for multilingual speech corpora. in E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (eds) *Improvements in Speech Synthesis.* London, John Wiley. pp.320-327.

[9] Courtois, F., Di Cristo, Ph., Lagrue, B. & Véronis, J. (1997) Un modèle stochastique des contours intonatifs en français pour la synthèse à partir des textes. *Acte du 4ème Congrès Français d'Acoustique. 1997* pp.373-376.

[10] Di Cristo, A. & Véronis, J. (1997). A metrical model of rhythm and intonation for French text-to-speech. *ESCA Workshop on Intonation: Theory, Models and Applications*. Athens. pp. 83-86.

[11] Jun, S.-A. (2000) *K-ToBI (Korean ToBI) labeling conventions: Version 3.1*. UCLA Working Papers in Phonetics 99. pp.149-173.

[12] Lee, H. Y. (1990) The Structure of Korean Prosody. PhD thesis. University of London.

[13] Kim, S., Yu, H., Hong, H. & H. Y. Lee (2007) A Study of Korean Intonation Using Momel, *Malsori, Journal of The Korean Society of Phonetic Sciences and Speech Technology. 63.* pp.85-100.