Automatic Annotation for Korean Speech Corpus Analysis

Jiwon Yun (Cornell University), Hyun Kyung Hwang (International Christian University), Seongyeon Ko (Queens College, CUNY)

Introduction

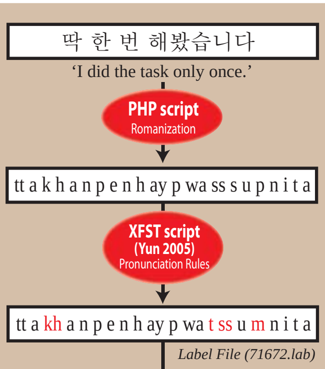
We conducted automatic annotation for a speech corpus in Korean and evaluated its performance in comparison with manual annotation.

Automatic Annotation

30 hours of read speech (24,300 sentences) produced by a single speaker from a speech corpus (ETRI 2006) was annotated in the automated procedure.

Step 1: Labeling

Unlike languages like English whose text-to-speech principles are not highly regular, Korean pronunciation is mostly predictable from the orthography according to a relatively small number of rules. We first romanized the transcripts of the corpus, then used a finite state machine (Yun 2005) to convert them into phonetic symbols by applying the pronunciation rules of Korean.



딱 한 번 해봤습니다

‘I did the task only once.’

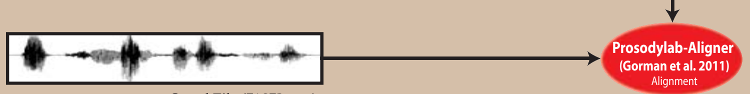
PHP script
Romanization

tt a k h a n p e n h ay p wa ss s u p n i t a

XFST script
(Yun 2005)
Pronunciation Rules

tt a kh a n p e n h ay p wa t ss u m n i t a

Label File (71672.lab)

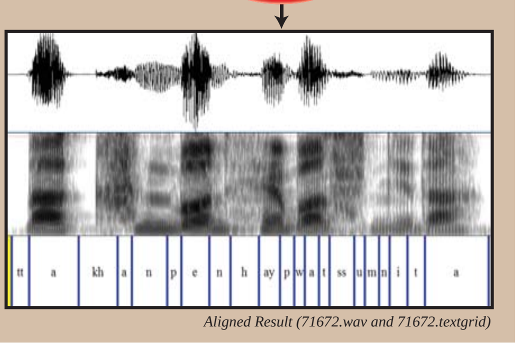


Sound File (71672.wav)

Prosodylab-Aligner
(Gorman et al. 2011)
Alignment

Step 2: Alignment

To align the phonetic symbols to the sound files, we used Prosodylab-Aligner (Gorman et al. 2011). This tool is applicable to any language in the world since its mechanisms are independent of language-specific features.



Aligned Result (71672.wav and 71672.textgrid)

Comparison with Manual Annotation

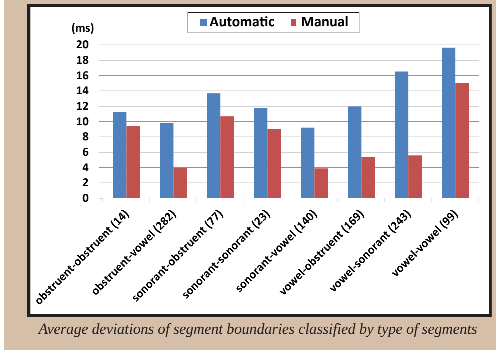
Manual annotation was carried out by three human transcribers for part of the corpus (25 sentences; 1306 segments) to evaluate the performance of the automatic system.

Labeling

1245 segments out of the 1306 were assigned the same labels by the automatic system and the human labelers, yielding **a 95% rate of agreement**. Among the remaining 61 segments, only 23 were unanimously labeled by all of the human transcribers. This suggests that **only 2% of the labels can be regarded as definite errors by the automatic system**. Most discrepancies were due to optional phonological processes or lexical processes that are not predicted by a purely rule-based system.

Alignment

The time differences of boundaries were measured for identically labeled segments. **The average deviation for the automatic procedure was 16 ms, whereas the manual procedure was 6 ms.** This is comparable to results from previous studies such as Wesenick and Kipp (1996) for read German speech (automatic: 18 ms, manual: 10 ms) and Pitt et al. (2005) for spontaneous American English speech (manual only: 16 ms).



Average deviations of segment boundaries classified by type of segments

Both humans and the automatic system yield the greatest deviations for vowel-vowel transitions, while the shortest for consonant-vowel transitions. This suggests that **the degree of difficulty in identifying boundaries is parallel for humans and the automatic system**, corroborating the finding reported in Wesenick and Kipp (1996).

Conclusion

The automatic annotation system for Korean used in this study is fairly reliable, compared to the results of human transcribers or other automatic systems in previous studies. Thus we expect that this automatic system will be easily applicable to phonetics research after a relatively small amount of hand-correcting, which will significantly reduce the amount of time and effort spent on annotation tasks.

References

Electronics and Telecommunications Research Institute (ETRI). 2006. Database of conversational sentences for speech synthesis. http://slrdb.etri.re.kr/DBSearch/Voice.asp.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. Proceedings of Acoustics Week in Canada, Quebec City. http://prosodylab.org/tools/aligner/

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication 45.89-95.

Wesenick, Maria-Barbara & Andreas Kipp. 1996. Estimating the quality of phonetic transcriptions and segmentations of speech signals. Paper presented to the ICSLP 96 (Fourth International Conference on Spoken Language).

Yun, Jiwon. 2005. Finite State Model for predicting Korean pronunciation. Unpublished manuscript, Cornell University.

Acknowledgement

We would like to thank Ho-young Lee and Hyosung Hwang for their help in obtaining the corpus. We are also grateful to Kyle Gorman for help in the use of the Prosodylab-Aligner and useful feedback on our work, and Masayuki Gibson for helpful comments.

International workshop on Japanese and Korean corpus linguistics, National Institute for Japanese Language and Linguistics, October 11, 2012

correspondence: jy249@cornell.edu