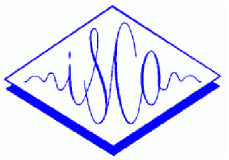
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2012
Sixth International Conference
Shanghai, China
May 22-25, 2012

A Forced-alignment-based Study of Declarative Sentence-ending "ta" in Korean

Tae-Jin Yoon[1], *Yoonjung Kang*[2]

[1]Department of Linguistics and Languages, McMaster University, Canada
[2]Department of Linguistics, University of Toronto, Canada

tjyoon@mcmaster.ca, kang@utsc.utoronto.ca

Abstract

The phonetic characteristics of declarative sentence-ending 'ta' was examined based on the speech of one male Korean speaker in his 20s drawn from a large-scale speech corpus. An analysis using a Unicode-based phone alignment system was compared to an analysis based on manually corrected alignment and the two methods produced largely comparable results. The declarative-ending 'da,' which coincides with prosodic intonational boundaries in Korean, was marked by higher F1 as well as longer duration and lower F0 than medial 'ta', but the effect of utterance boundary on voice quality is inconclusive. The study implies that more accurate duration modeling to the baseline alignment system will enable us to investigate phonetic and prosodic substances of spoken Korean in the large-scale corpus.
Index Terms: Forced alignment, declarative-ending, F0, F1, F2, H1-H2, vowel duration, voice quality.

1. Introduction

The current paper has two goals. One is to test the validity of a Unicode-based automatic phone alignment system developed for Korean. The other is to investigate the phonetic characteristics of the declarative sentence-ending 'ta' of Korean. The declarative utterance boundary in Korean is reported to be marked by final lengthening [1] and a falling or low F0 contour [2, 3]. Unlike English, in which utterance boundaries are often marked by non-modal phonation, such as glottalization [4, 5], it is not known how the utterance boundary affects voice quality. Besides, segmental properties are affected by the prosodic structure [6] in English, a language with impoverished morphological markers. The declarative-ending 'ta' (e.g. 'ka**da**' meaning "go") was chosen as an object of case study because it is the most frequently observed marker signaling the end of sentence. The phoneme sequence 'ta' can occur in sentence-medial position as a morpheme (eg. '**ta**' meaning "all") or as part of a morpheme (e.g. 'ba**ta**' meaning "sea"). We compared the phonetic characteristics of the declarative-ending 'ta' ('final ta' henceforth) and other instances of 'ta' sequences ('medial ta', henceforth) and also compared the results obtained from the automatic phone alignment with the results obtained from hand-corrected alignment.

The data used for this study was drawn from "A Speech Corpus of Reading-Style Standard Korean", created and distributed by the National Institute of the Korean Language (NIKL)[1] in 2007. The corpus contains 930 sentence types from 19 different well-known short stories and essays by 120 speakers, totaling 88,800 files. Speakers from the Seoul and Gyeonggi area whose parents are also from the Seoul and Gyeonggi area are selected. Eighty speakers, in their 20s (20 male, 20 female), 30s (20 male), or 40s (20 female), read all of the 930 sentence types. Speakers in their 50s or older (20 male, 20 female) read 404 sentence types. In total, there are 8,622 phrases that consist of 779,300 characters in the form of syllabaries of (C(C))V((C)C), where V can be a monophthong or a diphthong. The sheer size of the data allows researchers to investigate phonetic and phonological variation both within and across speakers. The current result is based on data from one male speaker in his 20s.

We developed a baseline phone alignment system based on Unicode (a standard for the encoding of text) and the Hidden Markov Model (HMM). The HMM-based baseline system needs as its required components acoustic models for each phone based on a pronunciation lexicon. One challenge in building a baseline system for languages such as Korean is the lack of an easily available pronunciation dictionary with which to build an accurate acoustic model. A mismatch between a word and its phone sequence is a major source of performance degradation. We constructed a pronunciation dictionary based on Unicode for Korean and used the dictionary to build acoustic models for each phone with HTK (Hidden Markov Toolkit) [7]. We adopted a knowledge-based approach to generate the dictionary; information about pronunciation variation was extracted from phonological rules of Korean [8]. The phonological rules include, among others, nasal assimilation, coda neutralization, and place assimilation. Acoustic models were trained for each phone, using speech samples produced by 40 speakers (20 male and 20 female in their 20's) of the Speech Corpus of Reading-Style Standard Korean. Figure 1 illustrates the output of the forced alignment system for an utterance in the corpus.

In order to examine whether the force-aligned segmental information is reliable enough for phonetic analyses, the force-aligned output of the first 8 out of 19 reading passages was manually corrected and the acoustic measurements based on the forced alignment were compared with those based on the manually corrected alignment. Manual correction was carried out by an undergraduate research assistant who took phonetics and phonology courses but did not know the purpose of the study.

2. Analysis

2.1. Data

Using the manually corrected and automatically generated phone-aligned data, all sequences of 'ta' were extracted. 412 occurrences of final 'ta' and 247 occurrences of medial 'ta' were found. Final 'ta' is always followed by a silent pause. As summarized in Table 1, 229 occurrences of medial 'ta' were found

[1]http://korean.go.kr

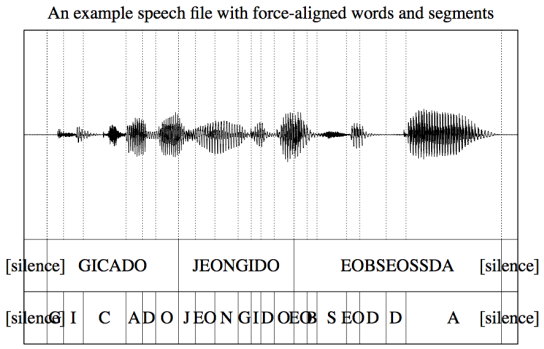


Figure 1: *Sample output of prototype Unicode- and HMM-based Alignment System for "A Speech Corpus of Reading-Style Standard Korean."*

before a consonant and 18 were found before a vowel.

Table 1: Number of tokens of sentence-final 'ta' and sentence-medial 'ta'.

Position	# of tokens	Next segments
final *ta*	412	Silent pause
medial *ta*	247	Consonant: 229 Vowel: 18

2.2. Acoustic features

For each token of the target vowel 'a', the acoustic measurements of duration, F0, F1, F2, H1-H2, H1-A1, H1-A2, and H1-A3 were extracted. All measurements except for vowel duration were taken at the center of the target vowel using a 30ms window. Duration and F0 are related to the prosodic characteristics of speech and based on previous studies, final 'ta' is expected to have longer duration and lower F0 than medial 'ta'. F1 and F2 are related to vowel quality, i.e., vowel height (F1) and backness (F2). The rest of the measurements are correlated with voice quality, with higher values signaling breathy voice and lower values signaling creaky voice. The Praat [9] script developed by Bert Remijsen [10] was used with some modification. It is known that the low harmonics (i.e., H1 and H2) are sometimes affected by the value of A1 of high vowels, which tend to have low F1 [12]. But, because the target vowel in our data is a low vowel, no formant-related correction to the harmonic-related measures was made. Figure 2 and Figure 3 illustrate the spectrogram and spectrum display of the vowel 'a' in medial 'ta' and final 'ta', respectively.

3. Results

3.1. Low vowel identity

As there is no previous study indicating that vowel quality is significantly changed in utterance-final position, we may expect the F1 and F2 values of the vowel in final 'ta' and medial 'ta' to be comparable. The average values of the first two formants (F1, F2) and their standard deviations (in parentheses) are presented in Table 2. For comparison, the average values of the two formants reported in [11] are provided. The provided val-

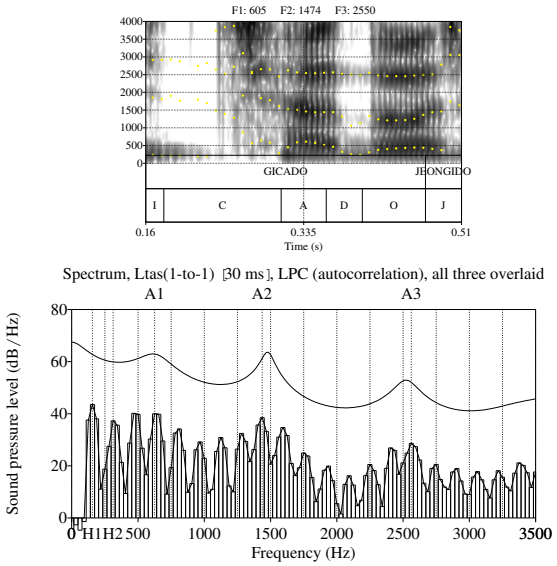


Figure 2: *Spectrogram displaying the target segment in sentence-medial position and the spectrum taken from the center of the target vowel 'a' using a 30ms window. The force-aligned segmental and word information is shown below the spectrogram. The spectrum indicates the locations of H1, H2, A1, A2, and A3, respectively. The values of the acoustic features are: duration (50ms); F0 (161Hz); F1 (605Hz); F2 (1474Hz); H1-H2 (6.2dB); H1-A1 (3.5dB); H1-A2 (5.0dB); H1-A3 (14.8dB).*

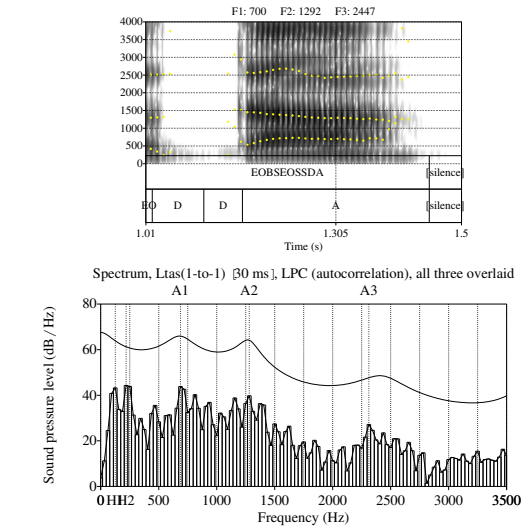
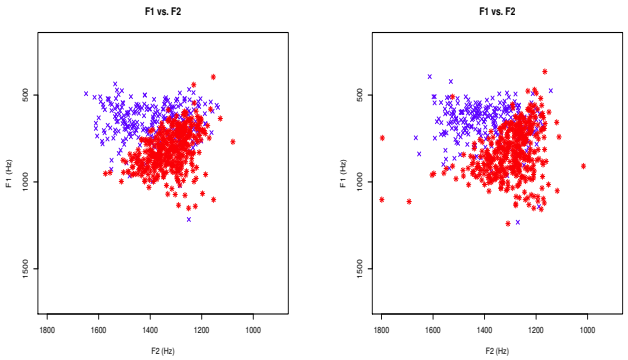


Figure 3: *Spectrogram displaying the target segment in sentence-final position and the spectrum taken from the center of the target vowel 'a' using a 30ms window. The values of the acoustic features are: duration (290ms); F0 (111Hz); F1 (700Hz); F2 (1292Hz); H1-H2 (-1.063dB); H1-A1 (-0.6dB); H1-A2 (3.4dB); H1-A3 (16.0dB).*



(a) manually-aligned data (b) force-aligned data

Figure 4: *Scatter plots of F1 and F2 frequencies for 'a' in medial (blue) and final 'ta' (red) in manually aligned data (a) and force-aligned data (b).*

ues are based on the vowel in a /h**a**da/ context produced by 10 male speakers (n=30).

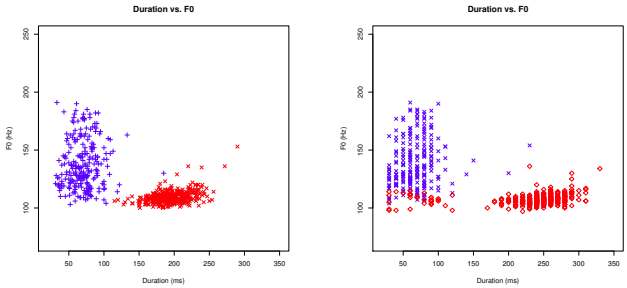
Table 2: Mean and standard deviation (in parentheses) of the first and second formants of the vowel 'a' from medial and final 'ta'. The data from [11] is provided for comparison.

	medial *ta*	final *ta*	Yang (1996)
F1	653 (119)	812 (113)	738 (87)
F2	1403 (141)	1323 (81)	1372 (124)

While there is a substantial overlap of formant values, Welch Two Sample t-tests for F1 and F2 in the force-aligned data indicate that the formant values differ significantly depending on whether the vowel occurs at the end of a sentence or in sentence-medial position (F1: t(463)=-12.9, p<.001, r=.51; F2: t(503)=7.18, p<.001, r=.30). The manually aligned data provides comparable results. Figure 4 provides scatter plots of F1 and F2 in the two contexts from the manually aligned data and the force-aligned data. While there is a greater overlap of the F2 values between the two positions, the F1 values of 'a' in sentence-final position is in general higher than the F1 values of 'a' in sentence-medial position. F1 is inversely related to vowel height or jaw height. That is, F1 tends to increase if the jaw lowers. Two possible accounts can be entertained: one is that final 'ta' is hyper-articulated (Eon-Suk Ko, personal communication), and the other is that it is affected by the following silent pause. For example, sentence-final 'a' occurs before a silent pause and the resting position of the tongue is lower than when the vocal tract is in speech mode. As a consequence, F1 of 'a' in sentence-final position may be expected to be higher than F1 of 'a' in sentence-medial position.

3.2. Boundary related cues

Utterance boundaries in declarative sentences in Korean are marked by falling or low pitch [2] and phrase-final lengthening in the vicinity of the intonational phrase boundary [1]. Our results reaffirm these previous findings. Figure 5 provides scatter plots of F0 and vowel duration measurements. The vowel 'a' in sentence-final position is longer in duration and lower in



(a) manually-aligned data (b) force-aligned data

Figure 5: *Scatter plots of duration and F0 for vowel 'a' in medial (blue) and final 'ta' in manually aligned data (a) and force-aligned data (b).*

F0 than 'a' in sentence-medial position. F0 of sentence-medial 'a' shows more variance than final 'a' and this is expected given the wide range of prosodic contexts medial 'a' may stem from. The measurements from the two contexts form non-overlapping clusters, especially in the manually corrected data, such that two clearly separate categories are visible for medial and final 'ta'. Taking together the findings of F1 and F2, this study demonstrates very clearly that the prosodic and segmental domains are independent of each other. That is, whereas the segmental cues of F1 and F2 taken from sentence-final and sentence-medial positions overlap, prosodic boundaries cues of F0 and duration form their own clusters depending on the position in utterance. In the force-aligned data, a substantial subset of the final 'ta' tokens (28 out of 412) show a duration less than 100ms and no comparable pattern was found in the manually corrected data. This discrepancy is a matter for further research.

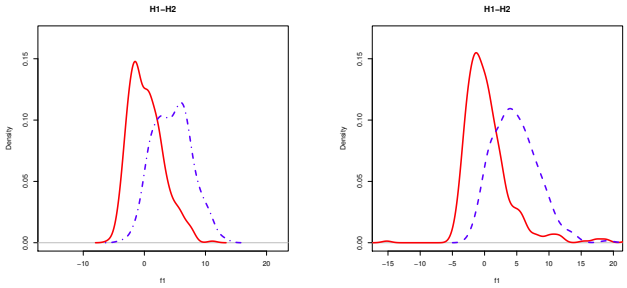
3.3. Features related to voice quality

One commonly used measure of voice quality is the amplitude difference in decibels between the first and second harmonics (=H1-H2, henceforth) [5, 12, 13]. The rule of thumb is that the value of H1-H2 is much larger in breathy voice (spread glottis) than that of creaky voice (constricted glottis). In English, non-modal phonation tends to be observed in utterance-final position [4, 5], even if it is not contrastively used. We may expect a similar pattern in Korean such that a non-modal phonation type is strongly correlated to the position in utterance in spite of the lack of its contrastive use in Korean.

Figure 6 provides the density plots of H1-H2 for medial and final 'ta' in manually corrected data and force-aligned data. Both plots indicate that creaky tokens are more likely to be observed in sentence-final position than in sentence-medial position, although modal or breathy voice was also commonly observed in sentence-final position for this speaker. Also, this voice quality difference is not an epiphenomenon of the F0 difference between medial and final position. It is also known that F0 and H1-H2 are not linearly related even though creaky voice tends to associate with low F0 [13]. This is shown in Figure 7, in which the scatter plot of H1-H2 and F0 in manually aligned and force-aligned data and the two measurements are only weakly correlated.

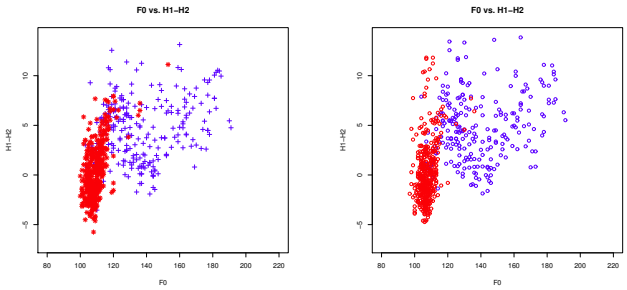
3.4. Correlation between manual and force aligned datasets

Table 3 shows the correlation of the acoustic measurements between the manually aligned data and the force-aligned data.



(a) manually-aligned data (b) force-aligned data

Figure 6: *Density plots of H1-H2 measurements for medial (blue) and final (red) 'ta' in manually aligned data (a) and force-aligned data (b).*



(a) manually-aligned data (b) force-aligned data

Figure 7: *Scatter plot of F0 and H1-H2 for medial (blue) and final (red) 'ta' in manually aligned data (a) and force-aligned data (b).*

Spearman's ρ coefficients were calculated for the acoustic features for vowels in medial and final 'ta'. Strong correlation was found for all acoustic measurements except for duration. Also, overall, the forced alignment system performed better with the sentence-medial tokens than sentence-final tokens.

Table 3: Correlation coefficients using Spearman's ρ for each acoustic feature between manually aligned dataset and force-aligned dataset at sentence-medial and sentence-final positions.

	Sentence-medial data	Sentence-final data
F0	0.9548786	0.7952297
H1-H2	0.896881	0.7900871
F1	0.8522826	0.728002
F2	0.9142722	0.7072893
Duration	0.5849893	0.4758266

4. Discussion and Conclusions

The paper examined the phonetic characteristics of 'ta' in sentence-final position in comparison with sentence-medial 'ta'. The phonetic studies revealed that the phonetic characteristic of 'ta' is affected by the prosodic structure. In spoken language, the segmental properties (as measured by F1 and F2) and prosodic boundaries cues (as measured by F0 and duration) show differences between sentence-final 'ta' and sentence-medial 'ta'. The role of voice quality is less clear. The findings observed from one male speaker may be an individual acoustic characteristic, which warrants detailed phonetic analyses for more speakers. The performance of a Unicode-based HMM forced alignment system was tested by comparing the acoustic measurements extracted from force-aligned data with those extracted from manually aligned data. Acoustic measurements that were taken at the center of the target vowel show strong correlation. This suggests that the automatic alignment method can be used for other acoustic studies that rely on these latter measurements. Nevertheless, duration showed poorer correlation and direct duration modeling for more accurate time-alignment of segments is necessary to improve the accuracy of duration and other measures.

5. Acknowledgements

We are grateful for Yesle Kim, who spent the summer manually correcting the force-aligned data,for Eon-Suk Ko for valuable discussion and for Hoi-Ching So for proofreading the paper. This work is supported by the Natural Science and Engineering Research Council (NSERC) of Canada. Statements in this paper reflect the opinions and conclusions of the authors and are not endorsed by the NSERC.

6. References

[1] Cho, H.-S. and Hirst, D. (2006). "The contribution of silent pauses to the perception of prosodic boundaries in Korean read speech." In Proceedings of Speech Prosody 2006.

[2] Jun, S.-A. (1993). The phonetics and phonology of Korean Prosody. PhD dissertation, Ohio State University.

[3] Jun, S.-A. (2000). K-ToBI Labelling conventions. [available at http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html]

[4] Redi, L. and Shattuck-Hufnagel, S. (2001). "Variation in the rate of glottalization in normal speakers." Journal of Phonetics 29: 407-427.

[5] Epstein, M. (2002). Voice Quality and Prosody in English. Ph.D. dissertation, UCLA.

[6] Kim, H. and Cole, J. (2005). Acoustic expansion of accented vowels in American English. Presented at the 79th Annual Meeting of Linguistic Society of America (LSA).

[7] Young, S, Evermann, G, Gales, M, Hain, T, Kershaw D, Liu, X, Moore, G, Odell, J Ollason, D, Povey, D, Valtchev V and Woodland P (2010) The HTK Book, Microsoft Corporation and Cambridge University Engineering Department.

[8] Ahn, S.-C. (1998). An Introduction to Korean Phonology, Seoul: Hansin Munhwasa.

[9] Boersma, P. and Weenink D. (2010) Praat: doing Phonetics by Computer (version 5.2), [downloadable from http://praat.org].

[10] Remijsen, B. (2004) script to measure & check formants and voice quality measures [praat script downloaded from http://www.lel.ed.ac.uk/ bert/msr&check_spectr_indiv_interv.psc]

[11] Yang, B.-G. (1996). "A comparative study of American English and Korean vowels produced by male and female speakers." Journal of Phonetics, 24: pp 245-261.

[12] Stevens, K. N. and Hanson, H.M. (1995). "Classification of glottal vibration from acoustic measurements." In O. Fujimura and M. Hirano (eds. by). Vocal Fold Physiology: Voice Quality Control. San Diego, Singular Publishing Group: 147-170.

[13] Yoon, T.-J., Zhuang, X., Cole, J., and Hasegawa-Johnson, M. (2009). "Voice Quality Dependent Speech Recognition." In S.-C. Tseng (ed.) Linguistic Patterns in Spontaneous Speech (Language and Linguistics Monograph Series), Academia Sinica: pp. 77-100.