USING MACHINE LEARNING TO MODEL THE THREE-WAY LARYNGEAL CONTRAST IN KOREAN

Jeremy Perkins[a], Yu Yan[b], Dahm Lee[c], Seunghun J. Lee[d]

University of Aizu[a], Ritsumeikan University[b], , Seoul National University[c], International Christian University[d], IIT Guwahati[d]

jperkins@u-aizu.ac.jp, yuyan@fc.ritsumei.ac.jp, dahm1021@snu.ac.kr, seunghun@icu.ac.jp

ABSTRACT

Machine learning via random forest was used to model learning of the laryngeal contrast of Korean obstruents. The model was trained on eight acoustic landmarks: f0, VOT, spectral tilt, psychoacoustic roughness and duration measures of closure and frication. The release duration (frication and aspiration) and aspiration duration of affricates and fricatives were also included. This method identifies which measures form necessary and sufficient conditions for successful machine learning, offering an additional way to identify potential contrastive cues using acoustic data. The results confirmed previous findings where the learning of the lenis-aspirated contrast depends on f0, and that of the tense stops depends on VOT. Release duration was the primary acoustic cue for learning tense affricates, and fricatives were learned using a combination of two measures: frication duration and either spectral tilt or release duration. Our results identify a minimal set of acoustic cues for learning the obstruent contrast in Korean.

Keywords: Korean, laryngeal contrast, machine learning, random forest

1. INTRODUCTION

The three-way laryngeal contrast among obstruents in Seoul Korean has received much past attention as it is a rare exception to the nearly universal two-way voicing contrast seen in most languages [1, 2]. In early research, these obstruents differed in VOT, with tense obstruents having short positive VOT, aspirated obstruents having long positive VOT and lenis obstruents having intermediate VOT [3]. However, among younger speakers currently, VOT differences between aspirated and lenis obstruents have decreased and are no longer clearly distinct. At the same time, f0 differences in the following vowel have supplanted VOT as the primary difference between aspirated and lenis obstruents [4, 5, 6, 7].

This change has provided linguists with an interesting test case then, since a cue shift in progress can be studied. Researchers have asked whether and to what extent f0 and VOT are used as cues in identifying each of the three laryngeal types. Studies that compared younger and older speakers found that younger speakers were more sensitive to f0 than VOT in the lenis-aspirated contrast [8]. Younger Seoul Korean listeners tend to use lowered f0, but not VOT, as a primary cue to recognize lenis obstruents, but tense and aspirated obstruents are recognized based on some combination of f0, VOT and spectral tilt [9, 10, 11].

Regarding the tense series in particular, Korean tense obstruents occur with reduced values of H1*−H2* and H1*−A1* at the onset of a following vowel [12]. However, other spectral tilt measures are also sometimes used (e.g. H1*−A2*, H1*−A3* and H2*−H4* [13, 14, 15]). In addition, recent work has shown that psychoacoustic roughness correlates with creaky voice in Burmese [16]. It's unclear which of these measures correlates most strongly with laryngeal constriction in a given language. Therefore, one of this paper's goals was to assess whether roughness can identify laryngeal constriction associated with the tense series, and also whether there were differences among roughness and the various spectral tilt measures in the relative importance of each measure. This result would be useful to inform future researchers on which of these various measures is most important in the production of laryngeal constriction of tense obstruents.

This paper's goal is to use supervised machine learning for new evidence on which parts of the acoustic signal may be used to learn the three-way laryngeal distinction among obstruents in Seoul Korean. The Random Forest (RF) for language modeling [17], which is a natural extension of the decision tree language models [18], is adopted. The main reason that the RF method was chosen is that RFs yield high accuracy compared with other

machine learning methods such as Support Vector Machine for speech recognition [19]. Although machine learning algorithms differ from human learners (in sensitivity and bias), they provide a way to quantify which and to what degree particular acoustic characteristics differ within a given sound contrast, allowing a glimpse of which features may play a contrastive role in linguistic contrasts.

2. METHOD

In this paper, acoustic data is collected, with a subset of that data used to train a network using the random forest technique. The trained network is then applied to another disjoint subset of the acoustic data to test whether it can accurately identify the three laryngeal obstruent types based on the acoustic data. This method identifies which parts of the acoustic signal are necessary and which are sufficient in learning each binary sound contrast; it can also compare candidate acoustic measures and assess which are more useful in learning a contrast.

2.1. Data collection

Twenty-four Seoul Korean participants (14 females), age 20 to 27 were recorded reading 66 words, all nested in a carrier sentence, 단어 X 는 무슨 뜻인가요?, meaning "What does the word X mean?". The words were bisyllabic CVCV words (most nonce) and the first consonant, which is the locus of the main research question, allowed all possible variations of laryngeal setting, manner and place among the Korean obstruents. The second consonant was one of the lenis or aspirated stops and only the vowel [a] was used. The 66 sentences were presented twice in random order via slides displayed on a screen. Participants wore a head mounted Shure WH-30 microphone connected to a Tascam MK-2 with an XLR cable. The audio files were recorded in mono at a sampling rate of 44.1 kHz. The files were segmented for consonant closure, frication and aspiration with all boundaries moved to the nearest zero-crossing using Praat [20].

F0, formant-normalized spectral tilt and segment duration were extracted from the vocalic portion following the first consonant using VoiceSauce at 10 ms intervals. Psychoacoustic roughness was also measured in the same way using a Matlab routine [16]. For f0, spectral tilt and roughness, only the initial measurement from each word was retained in the analysis in order to get a single measure, maximally adjacent to the consonant. F0 was normalized within each speaker relative to that speaker's median f0, and then converted to cents. Duration measures included closure duration for stops and affricates, frication and aspiration duration for affricates and fricatives, and VOT for stops. Release duration (aspiration plus frication) was also included for affricates and fricatives.

2.2. The Random Forest classification model for the Korean laryngeal contrast

The data was partitioned by manner and RFs were run separately for three subsets, each containing one of the three binary laryngeal contrasts (i.e. (1) aspirated & lenis, (2) aspirated & tense, (3) lenis & tense). Since there is a three-way laryngeal contrast among stops and affricates, but only a two-way contrast among fricatives, this resulted in seven separate models (three for affricates, three for stops, one for fricatives). The training data for each model involved randomly selecting one of the two repetitions from each item from each speaker. The other repetition for that item and speaker was assigned to the test data. As such, the testing and training data were equally balanced with respect to each other both by item and by speaker. All models were run with five layers and 200 nodes per layer.

The full set of acoustic measurements were f0, H1*−H2*, H1*−A1*, H1*−A2*, H1*−A3*, H2*−H4*, psychoacoustic roughness, VOT, closure duration, frication duration, aspiration duration, and release duration. RFs were run containing all possible permutations of these acoustic measurements. However, RFs that included more than one measure of spectral tilt were excluded since these various measures reflect the same physical characteristics (laryngeal constriction). A single spectral tilt measure is chosen from among the five spectral tilt measures to maximize model accuracy in each case. Roughness was only included in models reported in section 3.2.

The acoustic measures also differed for each model, since, for example, closure duration is defined only for affricates and stops, but not for fricatives. Also, since tense affricates and fricatives don't contain aspiration, the four models with tense affricates and fricatives included release duration instead of aspiration duration. RFs can learn the tense series easily based on the fact that frication duration is equal to release duration only for tense, but not for lenis and aspirated obstruents; however this reflects the actual situation in the language and so it was left as is.

A successful model is taken here to be one that achieves an overall accuracy of 95%. Relative importance values (ranging from 0 to 1) that quantify the contribution each measurement made

in each model are reported. A measure or group of measures was treated as sufficient to learn a given contrast if all models that contained those measures were successful. A measure or group of measures was treated as necessary to learn a given contrast if the only models that were successful contained those measures.

3. RESULTS

The RF model testing results based on seven acoustic measures are shown in Fig. 1. In the figure, the fully-loaded RFs with all relevant acoustic measures are arranged along the x-axis with accuracy plotted along the y-axis. Relative importance values are reflected for each acoustic measure via stacked bar plots. All fully-loaded RFs resulted in successful learning.

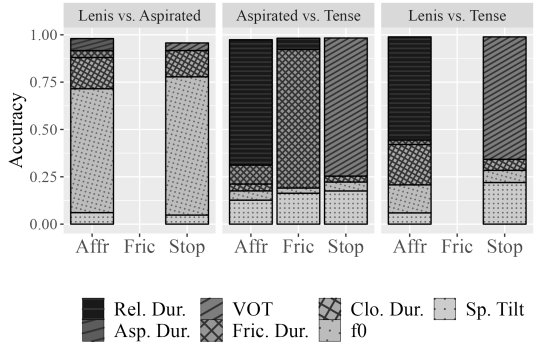

Figure 1: RF model accuracy and importance of each acoustic measure by binary laryngeal contrast and manner

Table 1 shows which acoustic measures were necessary and sufficient in each model. The results generally confirmed past findings [4, 5, 6, 7], while also offering some interesting insights which are discussed in the following subsections.

3.1. Acoustic measure importance

Among both stops and affricates, f0 was necessary for the lenis-aspirated contrast, as expected. However, among the stops, f0 wasn't sufficient on its own, confirming that other secondary cues are also relevant. The most likely secondary cue is closure duration, which had significantly higher importance than VOT and spectral tilt.

Regarding the tense series, previous research has suggested some combination of VOT, spectral tilt and f0 is likely involved for contrasts [9, 10, 11]. Our results showed that for the aspirated-tense contrast, release duration in affricates and VOT in stops were necessary for successful learning.

Table 1: Summary of necessary and sufficient acoustic measures for each contrast

Contrast	Necessary	Sufficient
Len-Asp Affr	f0	f0
Len-Asp Stops	f0	f0, clo. dur. f0, sp. tilt, VOT
Asp-Tns Affr	rel. dur.	rel. dur., any other
Asp-Tns Fric	fric. dur.	fric. dur., rel. dur. fric. dur., sp. tilt
Asp-Tns Stops	VOT	VOT
Len-Tns Affr	none	rel. dur. f0, sp. tilt, clo. dur., fric. dur.
Len-Tns Stops	none	VOT f0, sp. tilt, clo. dur.

Among fricatives, frication duration was necessary and was sufficient when paired with either release duration or spectral tilt. Thus, for affricates, release duration plays the role that VOT does for stops, and for fricatives, frication duration does the same.

In the lenis-tense contrast, no single measure was necessary, but release duration (for affricates) and VOT (for stops) were sufficient on their own. In addition, among the stops, the combination of f0, spectral tilt and closure duration achieved success; among the affricates, the same three measures plus frication duration sufficed. This result suggests primary roles for VOT and release duration, with f0, closure duration and frication duration playing secondary roles. Notably though, spectral tilt was not necessary for successful learning of the tense series. In fact, among the affricates, relative importance values indicated that closure duration plays a larger secondary role in the lenis-tense contrast, and that frication duration and spectral tilt played similarly important secondary roles in the aspirated-tense contrast. However, spectral tilt was by far the most important secondary measure for the tense fricative and stops.

3.2. Spectral tilt and psychoacoustic roughness

Another aim of this paper was to assess which of the various measures of spectral tilt and psychoacoustic roughness are optimal for learning the laryngeal contrast. To make this assessment, models containing the acoustic measures that were sufficient to reach 95% accuracy were selected, with spectral tilt or roughness necessarily included. This resulted in a set of models with minimal acoustic features, at least some of which were sufficient for successful learning, and that differed only on which spectral tilt or roughness measure was included. The resulting

model accuracies are compared in Table 2, and their relative importance values are compared in Table 3. Importance values are taken as more meaningful, since accuracy may be subject to a ceiling effect, and since importance values dispersed more widely.

Table 2: Model accuracy for spectral tilt and roughness measures. The measure with highest accuracy is bold.

Measure	Lenis-Asp		Asp-Tense			Lenis-Tense	
	Aff	Stop	Aff	Fric	Stop	Aff	Stop
Roughness	.967	**.960**	**.965**	.942	**.982**	.944	.907
H1*−H2*	.967	.959	.944	.942	.976	.946	.923
H2*−H4*	.967	.952	.958	.942	.964	.953	.915
H1*−A1*	.951	.952	.963	**.971**	.973	.946	**.958**
H1*−A2*	**.972**	.950	.963	.942	.970	.944	.925
H1*−A3*	.967	.949	.963	.947	.972	**.958**	.931

Table 3: Model importance of spectral tilt and roughness measures. The measure with highest importance is bold.

Measure	Lenis-Asp		Asp-Tense			Lenis-Tense	
	Aff	Stop	Aff	Fric	Stop	Aff	Stop
Roughness	**.135**	**.116**	.160	.178	.152	.060	.142
H1*−H2*	.135	.085	.141	.211	.159	.073	.182
H2*−H4*	.133	.103	.177	.187	.114	.117	.193
H1*−A1*	.101	.085	.218	**.257**	**.278**	**.184**	**.398**
H1*−A2*	.096	.068	.218	.202	.250	.131	.328
H1*−A3*	.096	.074	**.235**	.218	.189	.182	.277

In contrasts involving the tense series, where laryngealization was expected to play an important role, the spectral tilt measures had higher importance. However, in the aspirated-lenis contrasts, where spectral tilt is less important, roughness had higher importance. This result can be explained by the fact that roughness (but not spectral tilt) inversely correlates with f0 (in Thai [21]). Our results showed the same trend: Roughness was higher throughout low-tone vowels following lenis obstruents for most speakers, albeit with a local lowering effect at the consonant boundary.

This line of reasoning also explains why roughness fared extremely poorly in the lenis-tense distinction. Spectral tilt is lowered following tense obstruents due to laryngeal constriction and raised following lenis obstruents due to breathiness, causing diverging effects. However, roughness is raised following tense obstruents due to laryngeal constriction but is also raised due to decreased f0 following lenis obstruents, causing converging effects that make roughness a poor indicator of the contrast. This suggests roughness should only be used for identification of laryngeal constriction between two categories that do not differ in f0.

The aspirated-tense contrast provides a suitable test case to compare roughness and spectral tilt because f0 is the same. The importance values for roughness were relatively closer to spectral tilt, but still lower almost across the board. However, for stops and affricates, the accuracy for the models with roughness were higher than those with spectral tilt, suggesting that roughness may still be useful as a measure of laryngeal constriction. This was not the case for the fricatives however, where H1*−A1* was dominant for both accuracy and importance. Future research is needed to discern why fricatives differ in this way.

Finally, among the spectral tilt measures, H1*−A1* had higher importance values in all of the contrasts involving tense stops and fricatives, indicating that it is generally the best measure of laryngeal constriction among Korean tense obstruents. However, H1*−A3* may be a better measure for the tense affricates. The model with H1*−A3* had the highest importance for the aspirated-tense contrast in affricates. H1*−A3* also arguably outperformed H1*−A1* in distinguishing lenis and tense affricates by virtue of its only marginally lower importance and its higher accuracy. It is unclear why H1*−A3* would be better than H1*−A1* among affricates, but not also fricatives.

4. CONCLUSION

This study used the random forest machine learning method to learn the Korean three-way laryngeal contrast and found that f0 was most important in the lenis-aspirated distinction, and VOT was most important for the tense series, confirming previous findings. Among affricates and fricatives, release duration and frication duration were most important respectively. Spectral tilt was not necessary for learning any of the contrasts, but was found to be a significant secondary measure for most contrasts involving the tense series, as expected. Among the various spectral tilt measures, H1*−A1* was found to be relatively more important than others, although some evidence suggested that H1*−A3* may be optimal for tense affricates. Finally, psychoacoustic roughness was most important in the lenis-aspirated distinction, due to its correlation with f0 rather than laryngealization; as an indicator for laryngealization, comparable results to spectral tilt should only be expected among contrasts where f0 doesn't vary.

5. ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant-in-Aid for Early Career Scientists Number JP19K13162 and also by the ILCAA joint research project “Phonetic typology from cross-linguistic perspectives (PhonTyp)” (2021-2023).

6. REFERENCES

[1] M. R. Kim and S. Duanmu, “"tense" and "lax" stops in korean,” Journal of East Asian Linguistics, vol. 13, no. 1, pp. 59–104, 2004.

[2] C. W. Kim, “On the autonomy of the tensity feature in stop classification,” Word, vol. 21, pp. 339–359, 1965.

[3] L. Lisker and A. S. Abramson, “Cross-language study of voicing in initial stops: acoustical measurements,” Word, vol. 20, pp. 384–422, 1964.

[4] D. J. Silva, “Acoustic evidence for the emergence of tonal contrast in contemporary korean,” Phonology, vol. 23, pp. 287–308, 2006.

[5] J. Wright, “Laryngeal contrast in seoul korean,” Ph.D. dissertation, University of Pennsylvania, 2007.

[6] K.-H. Kang and S. G. Guion, “Clear speech production of korean stops: Changing phonetic targets and enhancement strategies,” Journal of the Acoustical Society of America, vol. 124, no. 6, pp. 3909–3917, December 2008.

[7] Y. Kang, “Voice onset time merger and development of tonal contrast in seoul korean stops: A corpus study,” Journal of Phonetics, vol. 45, pp. 76–90, 2014.

[8] K.-H. Kang, “Generational differences in the perception of korean stops,” Phonetics and Speech Sciences, vol. 2, no. 3, pp. 3–10, 2010.

[9] M. R. Kim, P. S. Beddor, and J. Horrocks, “The contribution of consonantal and vocalic information to the perception of korean initial stops,” Journal of Phonetics, vol. 30, pp. 77–100, 2002.

[10] M. Kim, “Correlation between vot and f0 in the perception of korean stops and affricates,” in Eighth International Conference on Spoken Language Processing, 2004.

[11] J. Schertz, T. Cho, A. Lotto, and N. Warner, “Individual differences in phonetic cue use in production and perception of a non-native sound contrast,” Journal of Phonetics, vol. 52, pp. 183–204, 2015.

[12] T. Cho, S.-A. Jun, and P. Ladefoged, “Acoustic and aerodynamic correlates of korean stops and fricatives,” Journal of Phonetics, vol. 30, pp. 193–228, 2002.

[13] M. Gordon and P. Ladefoged, “Phonation types: a cross-linguistic overview,” Journal of Phonetics, vol. 29, pp. 383–406, 2001.

[14] P. Keating, M. Garellek, and J. Kreiman, “Acoustic properties of different kinds of creaky voice,” in Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, 2015.

[15] M. Garellek, The phonetics of voice, the routledge handbook of phonetics ed. Routledge, 2019, pp. 75–106.

[16] J. Villegas, K. Markov, J. Perkins, and S. J. Lee, “Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 355–366, 2020.

[17] Y. Su, F. Jelinek, and S. Khudanpur, “Large-scale random forest language models for speech recognition,” in Eighth annual conference of the international speech communication association, 2007.

[18] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, “A tree-based statistical language model for natural language speech recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 7, pp. 1001–1008, 1989.

[19] K. H. Oo, “Comparing accuracy between svm, random forest, k-nn text classifier algorithms for detecting syntactic ambiguity in software requirements,” in International Conference on Information Systems and Intelligent Applications. Springer, 2023, pp. 43–58.

[20] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version xxx),” 2022.

[21] J. Perkins, “Acoustic measurement of laryngeal constriction in thai consonants,” in 35th General Meeting of the Phonetics Society of Japan, September 25-26 2021.