Psychoacoustic roughness as proxy of creakiness in White Hmong

Julián Villegas*, Jeremy Perkins*, Seunghun J. Lee**
*University of Aizu, **International Christian University & University of Venda
*{julian, jperkins}@u-aizu.ac.jp, **seunghun@icu.ac.jp

ABSTRACT

Creakiness of vocalic regions in White Hmong (a Hmong dialect with a three-way phonation contrast: modal, creaky and breathy tones) was measured with a state-of-the-art software predictor and with one based on an objective model of psychoacoustic roughness. Similar results for the two classifiers were found when comparing creaky vs. modal tones, but roughness classifier performance discriminating breathy and creaky tones, in comparison with the other classifier, was found to be subpar. These results suggest that roughness could be a good predictor of non-modal phonation, but further analysis and modifications are needed to improve roughness-based prediction of creakiness.

1. Introduction

Creakiness, a phonation here understood as a slow and sometimes irregular vibration of the vocal folds, is easily detectable in a spectrogram as series of visible sharp vertical stripes, often irregularly spaced. But, its detection via signal processing of the acoustic signal has proved difficult.

We hypothesize that perceptual attributes of speech (as opposed to unprocessed acoustic attributes or physiological correlates) could be good predictors of creakiness. Among the perceptual attributes of sound, psychoacoustic roughness seems to be related to the perception of creaky speech.

2. Psychoacoustic roughness

Roughness is a sensation elicited by changes in the temporal envelope of a sound with modulation frequencies ranging between 15–300 Hz, approximately; this sensation reaches a maximum when these amplitude modulations are at about 70 Hz. Its unit 'asper' has been defined so that the sensation elicited by a 100% amplitude modulated 1 kHz tone, at a modulation frequency of 70Hz, presented at 60 dB (SPL), equals 1 asper. Several models to quantify roughness have been proposed; in this research we use an implementation of Daniel and Weber's optimization of von Aures' model [1].

Vocalic regions were segmented in 50 ms long frames, 80% overlapped. A token was considered creaky if at least three frames were found creaky. The criteria to classify a frame as creaky included: frames with roughness higher than 4 aspers, and frames in which roughness increased more than 1 asper compared to those within five consecutive measurements.

3. White Hmong Corpus

White Hmong (a dialect of Hmong language spoken mainly in Laos) is characterized by the use of seven tones and three phonation types [3]. In this research, we used monosyllabic tokens from the corpus provided by Esposito and Yang, as found in the "Production and Perception of Linguistic Voice Quality" project repository of UCLA.[1] A total of 1,881 tokens (369 breathy, 472 creaky, and 1,040 modal), produced by 31 speakers (10 of which were female), were used in the analyses.

4. Experiment

To assess the performance of the roughness-based creakiness classifier, we compared its results with those found using Covarep [2], a state-of-the-art classifier based on an Artificial Neural Network (ANN), which uses several acoustic features (H1-H2, intra-frame periodicity, etc.) We hypothesize that the roughness-based classifier would perform as good as the ANN-based one, regardless of phonation.

4.2. Results

Contingency matrices (i.e., confusion matrices) of the classifications are presented in Figure 1. For this analysis, an equal number of tokens (corresponding to the maximum number of the least numerous phonation) was randomly selected from the corpus to avoid sample size bias in the analysis. The expert classification (in rows) was based on the information provided in the corpus.

When all phonation types were considered, the two classifiers have very similar performance, the roughness-based classification having a more conservative bias (a tendency to produce false negatives). The same small bias was observed in the classification of tokens, regardless of whether they were creaky or modal. But, when the subset of breathy and creaky tokens was considered, this bias was larger and the ANN-based classifier outperformed the roughness-based one.

[1] http://www.phonetics.ucla.edu/voiceproject

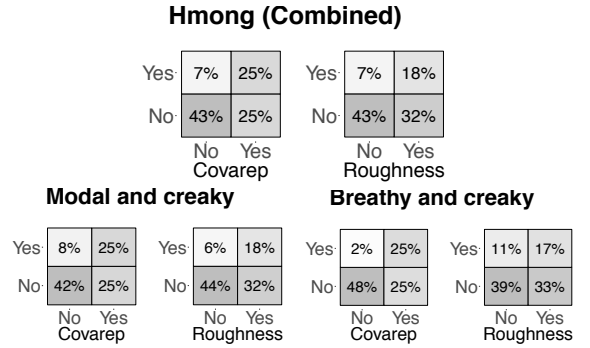

Figure 1. Is this token creaky? Comparisons of classifications made by experts (in rows) vs. automatic classifiers (in columns). An antidiagonal with 50% entries would denote a perfect match between expert and automatic classifications. All phonations considered (top panel), only modal and creaky (left) and breathy and creaky (right). An equal number of tokens was randomly selected to create these confusion matrices.

When areas under receiver operating characteristic curves were compared for each speaker (grouping modal and breathy tones as non-creaky and using as predictor the proportion of creaky frames within the vocalic segment), it was found that the roughness-based classifier performed significantly better ($p < .05$) for three speakers, and no significant differences in performance between the two classifiers were found for the additional 16 speakers.

5. Concluding remarks

The term creakiness has been used to describe a rather wide range of phonation, probably caused by different articulations. In this research, we used the phonemic classification found in the corpus, but it is not clear to the authors if this classification was based on expected tones from a dictionary (as opposed to manually verified), or if the creakiness produced by the speakers was similar. As illustrated in Figure 2, preliminary analysis suggests that when slow and irregular pulses of the glottis were present, the roughness-based classifier outperformed the ANN-based one, so a visual inspection of the analyzed tokens is being conducted to clarify the possible applications of roughness-based creakiness detection.

This revision would also help us to understand whether the rather poor performance of the roughness-based classifier distinguishing between breathy and creaky tokens could be partially caused by artifacts of the recordings (bursts of air could be picked up by the microphone increasing roughness in those regions) or a genuine characteristic of breathiness, in which case, it would be necessary to give different weights for roughness values depending on the frequency range, or to use additional correlates to discriminate between these two phonations.

Acknowledgments

This work was supported by JSPS Kakenhi Grant Numbers 15K16745, 16K02641, 16K00277 and by AKS Grant Number AKS-2016-LAB-2250004**.**

References

[1] P. Daniel and R. Weber, "Psychoacoustical Roughness: Implementation of an Optimized Model," Acta Acustica United with Acustica, vol. 83, pp. 113–123, 1997.

[2] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - A collaborative voice anal- ysis repository for speech technologies," in Ieee Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, pp. 960–964, 2014.

[3] Esposito, C. M. (2012). An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, *40*(3), 466-476.

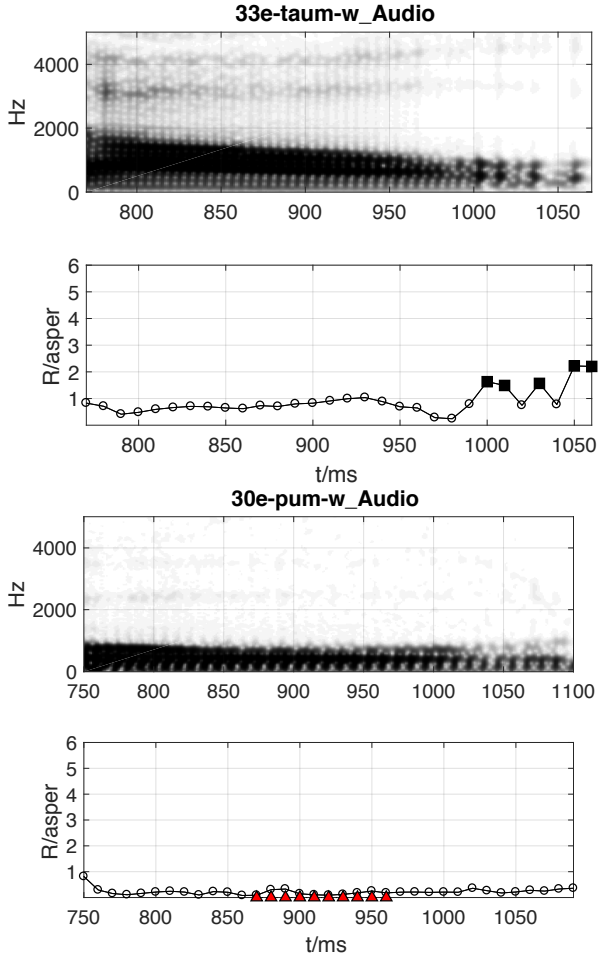

Figure 2. Spectrogram and time course of roughness for two creaky tokens. Above, vertical stripes in the spectrogram indicate irregular and slow pulses of the glottis; below, such stripes are not evident. Creakiness detected with the roughness-based method is denoted by squares; red triangles show creakiness found by the ANN-classifier.