

# Prediction of Creaky Speech by Recurrent Neural Networks Using Psychoacoustic Roughness

Julián Villegas , Senior Member, IEEE, Konstantin Markov , Member, IEEE, Jeremy Perkins , Member, IEEE, and Seunghun J. Lee 

**Abstract**—The use of a psychoacoustic roughness model as a predictor of creaky voice is reported. We found that the roughness temporal profile of vocalic segments can predict the presence of creakiness in speech. Using a simple bi-directional Recurrent Neural Network (rnn), we were able to predict the presence of creakiness in vocalic segments from only roughness traces with an accuracy similar to that obtained with rnns trained on at least 12-dimensional input data (including amplitude difference between the first two harmonics, residual peak prominence, etc.). Training rnns with the combination of roughness and multidimensional input data improved the performance of the predictor, but not significantly. Likewise, augmenting the dataset by time derivatives of the input features did not improve the predictor’s performance. The proposed roughness-based predictor eases interpretation and comparison of creakiness among corpora and suggests that roughness prediction models could be successfully used for classification of creaky intervals in speech.

**Index Terms**—Creakiness, Psychoacoustic Roughness, Recurrent Neural Networks, Phonation, Tone Classification.

## I. INTRODUCTION

CREAKY voice, vocal or glottal fry, creak, creakiness, laryngealization, pulse register phonation, and other terms are used in different disciplines to describe a kind of phonation often characterized by irregular pulses of the glottis at low frequencies ( $\leq 70$  Hz) [1], [2].

Creakiness has been observed in patients with pathologies such as spasmodic dysphonia [3] and Parkinson’s disease [4]. It also plays different roles in several languages: It is used in Jalapa Mazatec as a means of main contrast between tones [5]. In other languages such as Mandarin, it has a supportive or secondary role accompanying acoustic features such as fundamental frequency ( $F_0$ ) [6], and it is used in Finnish for marking turn transitions [7].

Manuscript received May 1, 2019; revised August 22, 2019 and October 14, 2019; accepted October 16, 2019. Date of publication October 24, 2019; date of current version April 8, 2020. This work was supported in part by the JSPS Kakenhi Grant 15K16745 and in part by the Strategic Japanese-Swiss Science and Technology Programme of JSPS and SNSF. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Tan Lee. (Corresponding author: Julian Villegas.)

J. Villegas is with the Computer Arts Laboratory, University of Aizu, Aizu-Wakamatsu 965-8580, Japan (e-mail: julian@u-aizu.ac.jp).

K. Markov is with the Human Interface Laboratory, University of Aizu, Aizu-Wakamatsu 965-8580, Japan (e-mail: markov@u-aizu.ac.jp).

J. Perkins is with the Centre for Language Research, University of Aizu, Aizu-Wakamatsu 965-8580, Japan (e-mail: jperkins@u-aizu.ac.jp).

S. J. Lee is with the International Christian University, Tokyo 181-8585, Japan, and also with the University of Venda, Thohoyandou 0950, South Africa (e-mail: seunghun@icu.ac.jp).

Digital Object Identifier 10.1109/JSTSP.2019.2949422

Moreover, in other languages creakiness has been associated with marking parenthetical information [8], conforming with a specific demographic group’s way of speaking [9], etc. The focus of this research is primarily on the phonemic role of creakiness, i.e., on how it is used for distinguishing words.

From a physiological perspective, there seems to be a variety of ways to produce this phonation, and some authors have suggested that rather than a single phenomenon, creakiness is a set of differently produced phenomena [10]. Speakers who routinely use creaky speech may differ in the way they produce it, but ultimately, they are capable of successfully modifying their production in such a way that their interlocutors are able to distinguish creaky from non-creaky speech. Blomgren *et al.* [11] reported that listeners were capable of classifying modal and creaky (fry) utterances with accuracy  $\geq 95.5\%$  (1100 responses for each kind of phonation). Additional evidence supporting the idea that listeners can distinguish between modal and creaky phonation regardless of how it is produced is provided by Gerratt and Kreiman [12]. Thus, perceptual attributes of speech, as opposed to unprocessed acoustic attributes or physiological correlates, could be good predictors of creakiness.

Among perceptual attributes of sound in general, psychoacoustic roughness seems to be related to the perception of creakiness. In the literature, both terms have been used to describe the other. For example, Titze posed that “creaky voice seems to be perceived as some combination of low pitch and roughness” [13]. Conversely, when describing the relationship between roughness and the number of audible beats, Helmholtz [14, p. 171] mentioned that “slow beats give a coarse kind of roughness which can be described as rattling or jarring.” In alternative translations of the German word ‘knarrend,’ the term ‘creaking’ has been used instead of ‘jarring’ [15]. Additionally, De Bruijn and Whiteside found that roughness and creakiness are strongly correlated when used for voice quality evaluation. This correlation was observed in ratings of language therapists and seems unaffected by differences in their years of experience [16].

Despite these commonalities, note that psychoacoustic roughness is a prothetic sensation (i.e., a percept on a continuum) [17], whereas creakiness is a metathetic sensation (i.e., a categorical percept) [12], so the two perceptual attributes are different, and in this study the former was used to predict the latter.

Also note that in the framework of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) [18], the term ‘roughness’ was used as one of six voice quality features, and it was defined as “perceived irregularity in the voicing source.”

In this article, however, we use ‘roughness’ to refer to the psychoacoustic feature exclusively.

Recently, roughness has been linked to the perception of dysphonic voice [19]. The authors of that study found that psychoacoustic roughness models used to predict roughness of sinusoids, narrowband noise, etc., could also model reported roughness of dysphonic voice. In this study we investigated the possible role of psychoacoustic roughness in the perception of creakiness. To the extent of the authors’ knowledge, this is the first study relating these phenomena.

Concretely, the purpose of this research is two-fold: 1) to investigate the feasibility of psychoacoustic roughness models to predict creaky intervals in speech, and 2) to compare roughness-based predictions with those made by state-of-the-art predictors. An automatic predictor of creakiness based on psychoacoustic roughness could provide a better understanding of the use of phonation as a means of contrast in some languages since it focuses on the reception/feedback end of the speech chain [20], rather than on the production side. Furthermore, it could also help to improve current detection methods of non-modal phonation, for the same reasons.

The rest of the article is organized as follows: Section II discusses the current understanding of creaky phonation and psychoacoustic roughness, as well as some of their prediction models. Section III describes the corpus used in our research and discusses the feasibility and accuracy of roughness as a creaky predictor. In Section IV, the results of our experiments are discussed along with future work. Section V concludes the manuscript with a summary of our main findings.

## II. BACKGROUND

### A. Creakiness

Creaky phonation has been considered in some accounts as a point along a continuum between a fully open glottis (breathiness/voicelessness) and a fully closed glottis (glottal stop/closure) [21]–[24]. In this school of thought, such phonation is produced when the vocal folds have weak longitudinal tension while greatly adducted (i.e., arytenoid cartilages pressed together), thereby reducing the overall glottal opening and contributing to a slow and sometimes irregular vibration of the vocal folds [25]. Compared to modal voice, this glottal constriction is also reflected in a lower rate of airflow and Open Quotient (OQ)—the ratio of the open phase to a complete cycle of vocal fold vibration [26].

In a different view, creakiness has been regarded not as a region on a continuum between a fully open and closed glottis, but as the combined effect of different glottal mechanisms to control airflow through the throat [27]. These mechanisms are selectively used by speakers to modulate their phonation; for example, creakiness is presumably produced by vocal fold adduction and abduction, and upward and forward sphincteric compression of the arytenoids. By additionally engaging ventricular incursion, harsh speech may be produced.

Different articulation settings translate into different acoustic features. E.g., for low OQ (when vocal folds remain closed longer than in modal phonation), the amplitude of  $F0$  relative to the next

harmonic decreases. As a consequence, various metrics have been used to derive the presence of creakiness from acoustic signals including spectral tilt,  $F0$ ,  $F0$  jitter, and Harmonic-to-Noise Ratio (HNR).

Among spectral tilt measurements, the aforementioned amplitude difference between  $F0$  (the first harmonic) and the second harmonic is known as  $H1-H2$ , and its formant-corrected version as  $H1^*-H2^*$ . These differences have commonly been used to predict creakiness, and the latter has seemed to yield better correlations with observed creakiness across different language corpora [10], [28]–[30]. Other differences such as  $H2-H4$ ,  $H1-A1$ , ( $A1$  being the tallest harmonic amplitude within the first formant), or peaks in higher formants ( $A_n$ ) have been explored, but they did not outperform  $H1^*-H2^*$  in general, even after compensating for the influence of formant energies [31].

Computing a linear regression between quefrency and cepstral magnitude has revealed larger deviations (prominences) from this line for periodic signals compared to those observed on aperiodic signals [32]. Low Cepstral Peak Prominence (CPP) compared to that of modal phonation has also been associated with the presence of creakiness, arguably because of irregularities in the vibration of the vocal folds [33]. These irregularities have also manifested in high  $F0$  jitter, low HNR, and high SHR—Sub-harmonic to Harmonic Ratio (the ratio between the magnitude of harmonics below the fundamental frequency and those above it) [34].

Different articulation settings are the basis for sub-classifications of creakiness. Depending upon the predominant articulation within a group of people, the performance of acoustic features as predictors of creaky segments varies. This variation in performance has hindered study and comparison of creakiness in speech since researchers have often chosen different acoustic features to discuss their findings.

Recently, tools merging different acoustic features and heuristics have been developed for automatic prediction of creakiness. These tools are, in general, more robust than predictions made by individual acoustic features, and some of them are discussed in following.

### B. Creakiness Prediction Models

High speed video [35], electroglottographic recordings [36], and other methods have been proposed to detect creakiness, but audio recordings are perhaps the most popular source for creakiness detection. A brief selection of the methods to estimate creakiness from audio recordings is presented here.

1) *Vishnubhotla’s Method*: Vishnubhotla and Espy-Wilson [37] proposed an extension to an aperiodicity, periodicity, and pitch detector. This extension was capable of detecting creakiness in running speech with no prior information about vocalic segments in the speech signal. To achieve that, the signal is split into frequency channels via a filter-bank, and periodic structures are sought in each channel. With this information, vocalic segments are determined, and creakiness within a vocalic segment is detected through a characterization of each frame based on a number of features. The authors claimed an 87%

TABLE I  
BASIC FEATURES USED IN COVAREP TO PREDICT CREAKINESS

Name	Description
H2H1	$H2-H1$ (dB)
res_p	Residual peak prominence (samples)
ZCR	Zero-crossing rate (samples)
IFP	Intra-frame periodicity contour (samples)
IPS	Inter-pulse similarity contour (samples)
PwP.fall	Power Peak fall
PwP.rise	Power Peak rise
F0	Fundamental frequency from the SHR algorithm (Hz)
F0mean	Mean Fundamental frequency from SHR algorithm (Hz)
enerN	Normalized energy contour (dB)
pow_std	Standard deviation of power contour frames (samples)
creakF0	$F0$ as output by the $H2-H1$ algorithm (Hz)

correct recognition rate of creakiness on continuous running speech.

2) *Ishi's Method*: Creakiness introduces long fundamental periods, longer than the window-size commonly used in  $F0$  analyses, yielding them ineffective. Ishi *et al.* [38] tackled this problem by performing a pulse-synchronized analysis of the signal. Additionally, they proposed to filter the signal between 0.1–1.5 kHz, select possible creaky frames considering local power peaks, and determine frame creakiness based on intra-frame periodicity and inter-pulse similarity. They reported a 74% correct detection of creakiness with this method.

3) *Kane's Method*: More recently, an automatic detection of creakiness based on the presence of secondary excitation peaks in the residual signal of a linear prediction filter and residual peak prominence was proposed by Kane *et al.* [39]. In this method, both features are used as input of a decision tree classifier. They also used a routine to improve creakiness detection by excluding unvoiced and non-speech intervals, and by considering the duration of the creaky segments, etc. The authors reported that their method achieved significant improvements on the classification accuracy of creaky segments in clean recordings (measured as F-score) relative to that obtained with Ishi's derived methods.

4) *Covarep Classifier*: In addition to the acoustic features discussed by Kane *et al.* [39], the same authors acknowledged in [40] the importance of speech features such as  $H1-H2$  in the characterization of creaky voice. They included these features along with those proposed by Ishi *et al.* [38] in an automatic detector of creaky voice. An implementation of this method is included in Covarep (v. 1.4.2) [41], a Matlab [42] library comprising several routines for speech analysis.

In Covarep's implementation, an Artificial Neural Network (ANN) is used for the creakiness decision. This ANN is fed with 36 features: 12 basic features summarized in Table I, along with their first- and second-order time derivatives.  $F0$  and  $F0$  mean are computed with a method based on the Summation of Residual Harmonics (SRH).

5) *Mori's Method*: Mori *et al.* [43] analyzed spontaneous Japanese speech. Vowels were marked as 'creaky,' 'breathy,' or 'modal' by two experts. They considered 15 basic features (intensity,  $F0$ ,  $F1$ ,  $F2$ ,  $F3$ ,  $F0$  jitter, HNR,  $H1-H2$ ,  $H1-A1$ , among others). Additionally, a 364-feature vector was extracted with openSMILE [44]. These features were fed to two machine

learning algorithms: Random Forest—RF and Support Vector Machine—SVM. Using the mix of basic and extra features, they obtained an Area Under the Receiver Operating Characteristic Curve AUC = .903 for breathy voice, and an AUC = .872 for creaky voice with the RF algorithm. Although the AUC was high, the F-score was 0.62 for breathy and 0.59 for creaky voice, indicating that the RF performance was not as good as that achieved by experts.

### C. Interim Discussion

From the surveyed literature, it seems that the Covarep classifier is the most popular (see for example, [45]–[47]). This classifier outputs creakiness probability and binary decision per frame, every 10 ms. However, there is no consensus on how to determine the creakiness of a segment based upon the creakiness of its constituent frames. First of all, the threshold used in Covarep on the binary decision of a frame's creakiness seems to be corpus-dependent [48]. Hence, gauging the creakiness of a segment becomes problematic. Kuang [49] used the mean of frame binary decisions to determine a segment's creakiness, while other alternatives exist, such as computing the log of creaky probability for each segment and computing their mean, etc. [50].

Furthermore, many languages are known to have creaky phonation focused at specific times in vocalic segments. For example in Burmese, it has been reported that creakiness usually appears only in the second half of a creaky tone [68]. Thus, means taken from the entire vocalic segment may not be appropriate in these cases.

### D. Psychoacoustic Roughness

Roughness is a psychoacoustic attribute of a sound (not only speech) comparable to pitch, loudness, sharpness, etc. In the same way that any sound has some degree of loudness or pitch, it also has some degree of roughness. Roughness produces continuous and quantitative changes associated with rapid amplitude modulations (between 15–300 Hz). Perceived roughness reaches a maximum for modulation frequencies around 70 Hz [51].

One asper (the unit of roughness) is defined as the roughness elicited by a 100% Amplitude-Modulated (AM) 1 kHz sinusoid at a modulation frequency of 70 Hz, presented at 60 dB (SPL) [52]. By manipulating the modulation index of this AM sinusoid, the absolute threshold of roughness perception was found to be 0.07 aspers, and its just noticeable difference  $\Delta R/R = 17\%$ . I.e., in order to perceive a change of roughness  $\Delta R$ , it must be at least 17% of its current value  $R$  [51].

Apart from the modulation index and modulation frequency in AM sinusoids, roughness is influenced by the Sound Pressure Level (SPL), frequency deviation (in frequency modulated sinusoids), etc. [52]. Roughness is not exclusive of periodic amplitude modulations. Random modulations (like those found in narrowband noise) yield high values of roughness as well.

Important contributions to the understanding of roughness came originally from Helmholtz [14], who observed that musical consonance could be explained in terms of the roughness produced by the interactions between frequency components of

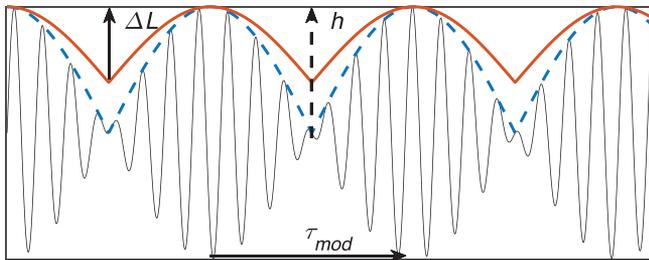


Fig. 1. Roughness dependencies: discrepancy between the temporal masking depth  $\Delta L$  and the crest-trough difference  $h$  appears for rapid modulation frequencies. This difference varies with modulation period  $\tau_{mod}$ , the inverse of the modulation frequency  $f_{mod}$ .

simultaneous complex waves. Later, Terhardt [51] found that relative amplitude fluctuation almost exclusively explained roughness of amplitude- and frequency-modulated sounds, while Plomp and Levelt [53] linked the maximum roughness elicitation to a separation of  $\sim 25\%$  of a critical bandwidth (in terms of critical bands as reported by Zwicker *et al.* [54]). More recently, Pressnitzer and McAdams [55] found that phase and temporal asymmetries of a sound wave also contribute to the perception of roughness. Roughness seems to play a major role in sensory pleasantness [56], sound quality [57], musical dissonance [58], [59], psychoacoustic annoyance [52], and speech intelligibility [60].

### E. Roughness Prediction Models

Roughness prediction models can be categorized according to two views: spectral approaches, which are based exclusively on the spectral segregation made by the basilar membrane [61], and temporal approaches, which also take into account temporal aspects of the signal such as the phase-lock of nerve cells to the period of the stimuli [62], [63]. Among the former group, some of the most important models correspond to those of Helmholtz [14], Plomp and Levelt [53], Sethares [64], and Vassilakis [65]. In domains where the beating of harmonics is supposed to be the most important source of roughness (e.g., musical dissonance), spectral approaches have been widely adopted. Temporal approaches have more commonly been used in other applications such as measurement or prediction of auditory annoyance, etc.

Creaky speech features temporal envelope modulations of around 20–70 Hz. Hence, in this study we opted for using a temporal model to predict roughness. In this kind of model, the roughness  $R$  is considered to be dependent chiefly on frequency and temporal resolution of the hearing system [52]. This is approximated as

$$R \sim f_{mod} \Delta L \quad (1)$$

where  $f_{mod}$  is the frequency of amplitude modulation and  $\Delta L$  is the temporal masking depth, as illustrated in Fig. 1.  $\Delta L$  accounts for the fact that rapid changes in the amplitude envelope of a signal are not accurately perceived, i.e.,  $\Delta L < h$ , the true crest-trough difference of the modulation.

More precise prediction models of roughness compute the temporal masking depth through different auditory channels

$$R = 0.3 \frac{f_{mod}}{\text{kHz}} \int_0^{24 \text{ Bark}} \frac{\Delta L_E(z) dz}{\text{dB/Bark}}, \quad (2)$$

and different models vary upon the computation of  $\Delta L_E$ , the temporal masking depth of a given auditory channel.

Von Aures [66] estimated global roughness  $R$  by computing specific roughness in 24 disjoint auditory channels corresponding to the bands of the Bark scale, as shown in Eq. (2). He used cross-correlation between adjacent bands to diminish the effect of random-like noise on the reported values of roughness. Daniel and Weber [15] optimized von Aures' model mainly by increasing (and overlapping) the number of auditory channels.

### F. Roughness Prediction Model Used in This Study

In the current study, we measured objective roughness using a Matlab implementation of Daniel and Weber's model [67]. This implementation was found to closely match empirical results reported by von Aures [66] on the roughness of AM sinusoids at different modulation frequencies and different frequency bands (Pearson's product-moment correlation  $r = .971$ ,  $p < .001$ ).

For the computation of roughness temporal profiles, we used frames that were 50 ms long, Blackman-windowed, and 80% overlapped (i.e., a 10 ms hop between adjacent analysis frames). These frames were divided into 47 auditory channels (one bark width with a half-a-bark overlap between channels), covering the audible spectrum from 20 Hz to 15.5 kHz. Specific roughness for each channel was computed and total roughness of each frame was obtained as a weighted sum of each channel's specific roughness.

## III. METHODS

In this section, prediction of creakiness by roughness is detailed. First, we describe the nature of the Burmese language corpus used here. Next, we compare roughness, spectral tilt, and Covarep prediction temporal profiles, to illustrate the feasibility of roughness as a predictor. Finally, we present a series of experiments with Recurrent artificial Neural Networks (RNNs).

### A. Materials

1) *Burmese*: Burmese has four contrastive tones that are associated with vowels: Creaky, checked, high, and low. Every vowel has one of these four tones with the exception of vowels in minor syllables, not discussed here. Tone is contrastive, meaning that it is important in distinguishing words from one another. Table II illustrates this with four words that differ only in their tone.

Tone in languages is commonly associated with the contrastive use of different  $F_0$  contours; however, in Burmese, not only  $F_0$ , but also duration, intensity, and phonation are involved in the four-way tone contrast, as shown in Table II.

Creaky tone involves a short-duration vowel with a high falling  $F_0$  contour characterized by gradually increasing creakiness, especially in the latter part of the vowel. Checked tone

TABLE II  
EXAMPLES OF BURMESE WORDS VARYING THEIR MEANING WITH TONE

IPA	Meaning	Tone	$F_0$	Duration	Phonation	Intensity
[kuʔ]	royal order	checked	high, sharp fall	very short	late creakiness	very high
[kû]	to give medical treatment	creaky	high, sharp fall	short	late creakiness	high
[kú]	to cross over	high	mid rising or falling	long	modal or breathy	medium
[kù]	to help	low	low level	long	modal	low

differs from creaky tone in that it has an even shorter duration and a more abrupt final laryngeal constriction (a glottal stop coda), usually accompanied with some creak. In fact, checked tones are often viewed as a distinct syllable type, rather than a tone due to this final glottal stop. While checked tones are all spoken with glottal stop codas, the Burmese writing system encodes for obstruent codas with different places of articulation. Historically, these codas could vary in place of articulation, but this distinction has reportedly been neutralized in the modern spoken language [68].

High tone is described as having a long duration and sometimes as breathy. Low tone is described as having a similarly long duration but with modal phonation. While low tone has a low  $F_0$  throughout its duration, high tone has a relatively higher peak  $F_0$  with a contour that varies depending on its context; in citation form, it has a falling contour. High tone has a more moderate  $F_0$  relative to creaky and checked tones, i.e., it is closer to a mid tone and the use of ‘high’ is to contrast it with the low tone. Regarding intensity, checked tone has the highest intensity, followed in order by creaky, high, and low tone.

Previous acoustic studies of phonation in Burmese tones have met with varying levels of success. Spectral tilt (measured as  $H1^* - H2^*$  or similar spectral differences) showed a poor ability to predict creakiness [69]. Gruber noted that this result may have partly been due to the fact that creakiness is located towards the latter part of the vowel in Burmese, and may not be detectable at the midpoint, where these previous studies attempted to measure it [68]. He also found that measuring spectral tilt closer to the endpoint of the vowel yields a higher correlation with creaky tone status.

2) *Corpus*: The corpus used in our experiments comprised single words uttered by twelve native speakers (six of each gender) of mostly Yangon Burmese. One of the speakers (BRM510) was from Magway, Myanmar. Gruber [68] found that the phonation contrast in Burmese is neutralized in carrier sentences and is only seen in citation form (i.e., words read in isolation). We elicited the same word list in citation and carrier sentences, verifying Gruber’s finding that creakiness is only produced in citation form. Thus, we focused only on the citation form for this study.

All recordings were made in a quiet room where speakers wore a head-mounted unidirectional microphone (Shure WD30)

connected to a solid-state recorder (Marantz PMD661 MKII) which stored the audio at a sampling rate of 44.1 kHz.

78 monosyllabic Burmese words repeated five times were produced in isolation by each speaker, resulting in a corpus of 4,679 tokens (one token was discarded). The word list was nearly balanced across the four language tones: 18 words had either creaky, low, or high tone, and 24 words had a checked tone. Examples of these tones uttered by the same female speaker are presented in Fig. 2; additional multimedia examples can be found at <http://onkyo.u-aizu.ac.jp/software/creakbyr>.

In addition, the word list was balanced for coda type (no coda vs. nasal coda) and vowel quality ([i], [u], and [a]). We varied the orthographic obstruent coda, just in case this did have an effect on the spoken forms. Onset consonants were mostly alveolar (62 words) with some velars (15 words) and a single word with a palatal onset. Most words had obstruent onsets (59 words), 19 words had sonorant onsets (eleven with liquid [l] and eight with nasal onsets). Bilabial onsets were not used.

The vocalic segments of the audio files (one per token) were manually labeled in Praat [70]. Onset of regular glottal pulses was used to mark the beginning of each vocalic segment. The offset of the vocalic segments was placed at the end of the final glottal pulse. Finally, all boundaries were moved to the nearest zero-crossing.

3) *Validation*: To assess whether the pronunciation of our speakers was similar to that described by the dictionary, we randomly selected a number of utterances per tone and speaker for manual verification. This review was independently performed by three of the authors, and we considered the spectrogram, waveform, and audio of each utterance to determine whether it was creaky or not. In total, the judges reviewed 492 utterances (about 11% of the corpus). The sample size for each combination of speaker and tone was ten, except in the case of checked tone (the most numerous among the four tones) where it was eleven. With this sample size, the margin of error was  $m = \pm 3 [t(9) = 2.262]$  with a confidence interval  $CI = 95\%$ , a finite population correction of 0.889, and assuming a standard deviation  $SD = 0.05$ .

Fleiss’ Kappa index  $\kappa$  was computed to assess the inter-rater agreement with the library `irr` [71] in R [72]. It was found that there was very good agreement between the three judges,  $\kappa = .94 [z = 36.1, p < .001]$ . Fig. 3 presents the results of the manual verification: Percentages were computed from the ratio of the number of utterances rated as creaky by the judges and the total number of ratings per speaker and tone. The best agreements between dictionary entries and expert judgements were obtained for non-creaky tones (i.e., high and low), with a discrepancy of 2.3%, while for creaky tones (i.e., checked and creaky), these discrepancies amounted to 8.3%.

Discrepancies between dictionary entries and expert judgements were more abundant for speakers BRM509 and BRM510. Besides their creakiness opinions, judges were also asked to comment on each utterance. From these comments, it seems that speaker BRM509 was some times exaggerating her pronunciation to the point where it was difficult to determine which tone was being used. On the other hand, BRM510 did not always use creakiness as expected: For checked and creaky

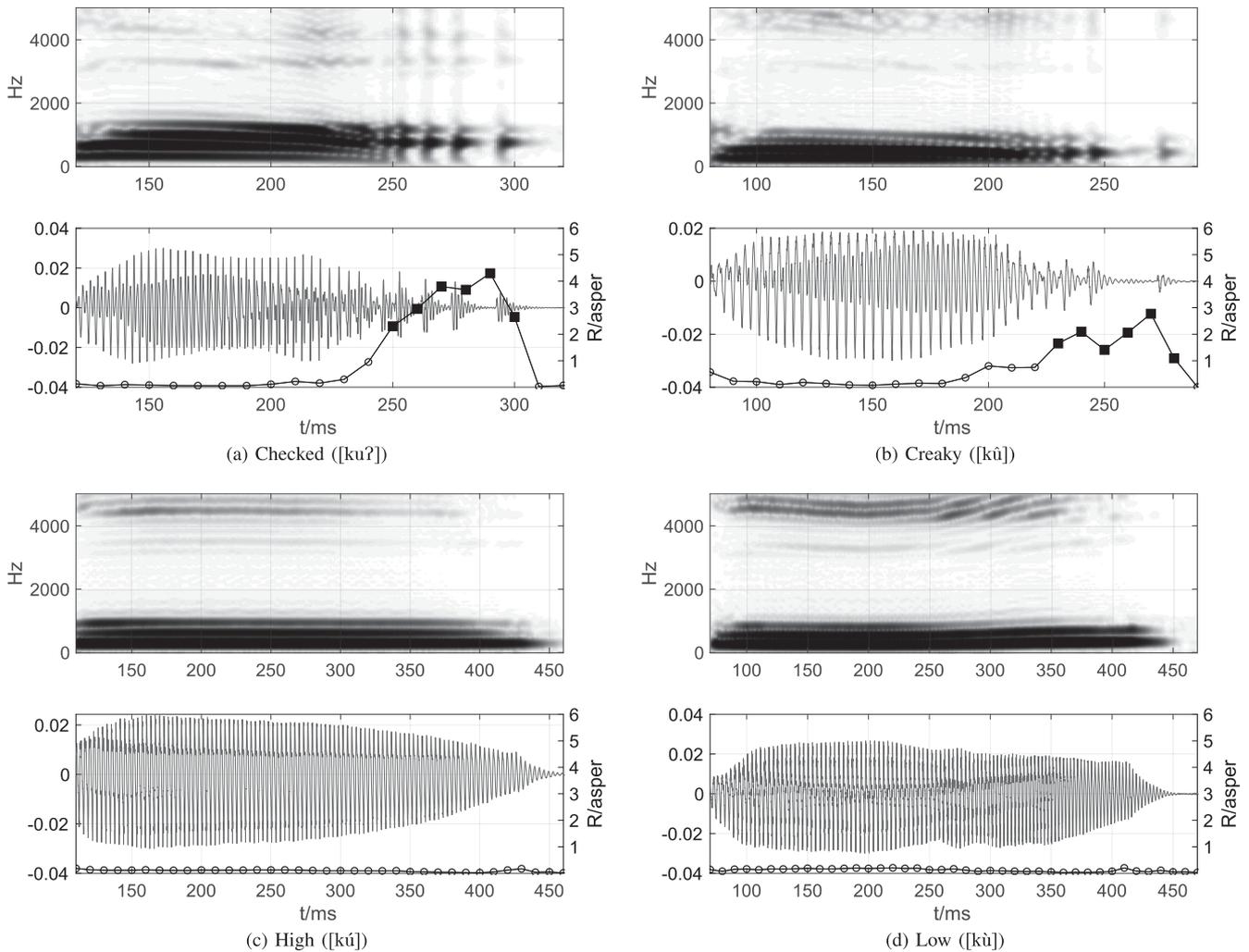


Fig. 2. Examples of the vocalic segments for the four Burmese words presented in Table II produced by a female speaker (BRM506). Spectrograms on top, waveforms and roughness temporal profiles on bottom. Creaky frames, as determined by the roughness-based prediction, are indicated by filled squares.

tones, he used short vocalic segments with a sharp fall of  $F_0$ , but without creakiness in many cases. This speaker came from a different region than the rest of the speakers, so it is possible that dialectal variation could explain his pronunciation. In all cases, the discrepancies found between dictionary entries and expert judgements could be considered normal in the context of inter-speaker or dialectal variation. We decided to preserve all utterances, which makes our corpus consequently noisy.

### B. Comparison of the Temporal Profiles of Several Predictors

We compared two common alternatives to estimate creakiness with a roughness-based prediction on the vocalic segments of the Burmese corpus as a way to assess the feasibility of the latter. These alternatives were spectral tilt and the method implemented in Covarep (creakiness probability and binary decision).

1) *Alternative Methods:* Spectral tilt was measured as  $H1^* - H2^*$ . For each utterance in our corpus, spectral tilt was measured every 10 ms with Voicesauce (v. 1.36) [73]. Covarep

measurements were computed at the same rate with no modifications to the algorithm. We registered the probability of creakiness and binary decisions for creakiness as output by the program.

2) *Creakiness Prediction Based on Objective Roughness:* For the roughness-based prediction of creakiness, we used the roughness implementation described above in Section II-F. Additionally, the following adjustments were made: Monophonic audio recordings were resampled at 16 kHz to minimize the effect of high frequency roughness. Note that similar resampling is performed in Covarep and other methods, arguably to reduce computational load. Recordings were also DC-filtered, and amplitude peak-normalized to be 0 dB (re. Full Scale—FS). Since the actual pressure level at which each utterance was produced is unknown, it was assumed that each audio frame was produced at a sound pressure level of 80 dB, effectively eliminating SPL differences between frames.

With these settings, preliminary visual inspection of roughness temporal profiles suggested that the vocalic segment of creaky tokens displayed frames with either large values of

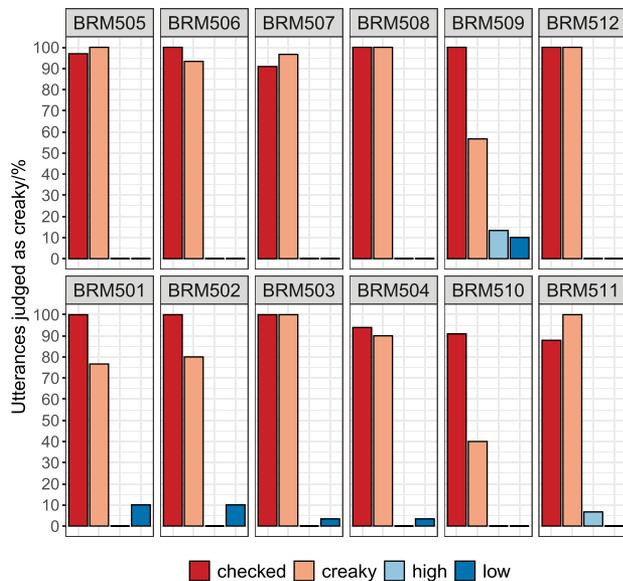


Fig. 3. Percentage of utterances rated ‘creaky’ by three experts for each tone in a sample of words randomly selected from our corpus. Top and bottom rows correspond to female and male speakers, respectively.

roughness or extreme roughness changes from frame to frame. In light of this, we arbitrarily set an absolute roughness threshold of 4.0 aspers, above which a given frame was considered creaky. In similar fashion, a frame was considered creaky if it was 1 asper higher than any of those in a 5-frame vicinity (frames were 50 ms long, overlapped 80%). This was implemented via a Hankel-like matrix with five columns and as many rows as needed depending on the length of the vocalic segment. Roughness traces with estimated creaky frames in the vocalic segments of the words presented in Table II, as uttered by a female speaker, are shown in Fig. 2.

3) *Results: Vocalic segment traces* corresponding to the same tone, speaker, and feature (Covarep’s probability and binary decision; spectral tilt; and roughness binary decision) were first time-normalized (i.e., setting the time to zero at the beginning of each vocalic segment, and dividing each frame time within it by the length of the vocalic segment) and then used to compute smooth conditional means. The smoothing was done with a generalized additive model using cubic splines and a span factor of 0.1% via the function ‘geom\_smooth’ [74] in R. 95% confidence intervals were also computed and are shown in gray around the smooth lines in Fig. 4.

As shown in Fig. 4(a) and 4(b), predictions made with Covarep were not always accurate, regardless of the output used (probability or binary). For our corpus, the two Covarep outputs produced very similar results: High and low tones yielded greater creaky frame probabilities (or number of creaky frames for the binary decision) than checked and creaky tones for some speakers (e.g., BRM505, BRM508). In other instances, differences between the traces were minimal (e.g., for speakers BRM503, BRM506, and BRM509). In fewer instances, Covarep results were in agreement with the tones (e.g., for speakers BRM501, BRM512).

The results obtained with spectral tilt measurements were somewhat better: Creaky tones had a lower spectral tilt than high and low tones. These differences were not always clear throughout the time course of each vocalic segment, making creakiness judgement time-dependent, as shown in Fig. 4(c) (see speakers BRM504, BRM507, and BRM510, for example).

Except in the case of speaker BRM510, roughness-based classifications consistently yielded a number of creaky frames for checked and creaky tones that was higher than that of high and low tones. This was especially clear in the second half of the vocalic segment. As mentioned before, a manual revision of BRM510’s production revealed that this speaker did not always produce creaky and checked tones with creaky phonation.

In agreement with [68], the roughness-based detector shows that Burmese creaky and checked tones have late creakiness in words in isolation. Finally, this experiment confirmed the feasibility of a roughness-based predictor of creakiness, details of which are provided in the following section.

### C. Classification Based on Recurrent Neural Networks

As discussed in Section III-B, there was clear evidence that roughness contours could be good predictors of creakiness. In that section, the criteria for determining the binary creakiness of a frame was based on visual inspection of roughness contours. It is very likely that we missed some patterns in the profiles that could improve the creakiness prediction made with psychoacoustic roughness. For that reason, we decided to experiment with Recurrent Neural Networks (RNNs) as a binary creaky segment classifier. The task given to the classifier was simply to decide whether a given vocalic segment is creaky or not. As previously mentioned, the length of the vocalic segment is in general different for each utterance. The roughness contour is a sequence of values (in aspers), one for each speech frame, and RNNs are especially well suited for modeling such kinds of temporal data series.

The dataset for this experiment consisted of voiced segments from all 12 speakers’ utterances. In order to get speaker independent results, we used the leave-one-speaker-out strategy, or 12-fold cross validation where the test data for each fold came from different speakers. The data from the remaining 11 speakers were randomly split into training and validation sets with a 10:1 ratio. Validation sets were used to tune some of the RNN hyper-parameters, such as batch size, optimizer, learning rate, etc.

1) *RNN Trained Exclusively on Roughness:* We experimented with various RNN structures and hyper-parameter combinations, but a simple bi-directional RNN with a few layers and several dozens of Gated Recurrent Units (GRUs) turned out to be the most suitable. The RNN input is a one-dimensional roughness contour and the output is also a one-dimensional sigmoid node for binary prediction. The results given in all of the following tables are obtained with batch size  $bs = 10$ , Adam optimizer with learning rate  $lr = 0.001$ , Binary Cross-Entropy (BCE) loss function, and a maximum of 50 training epochs. The validation data loss was monitored during training and the model of the epoch with the smallest loss was saved for evaluation with the test data.

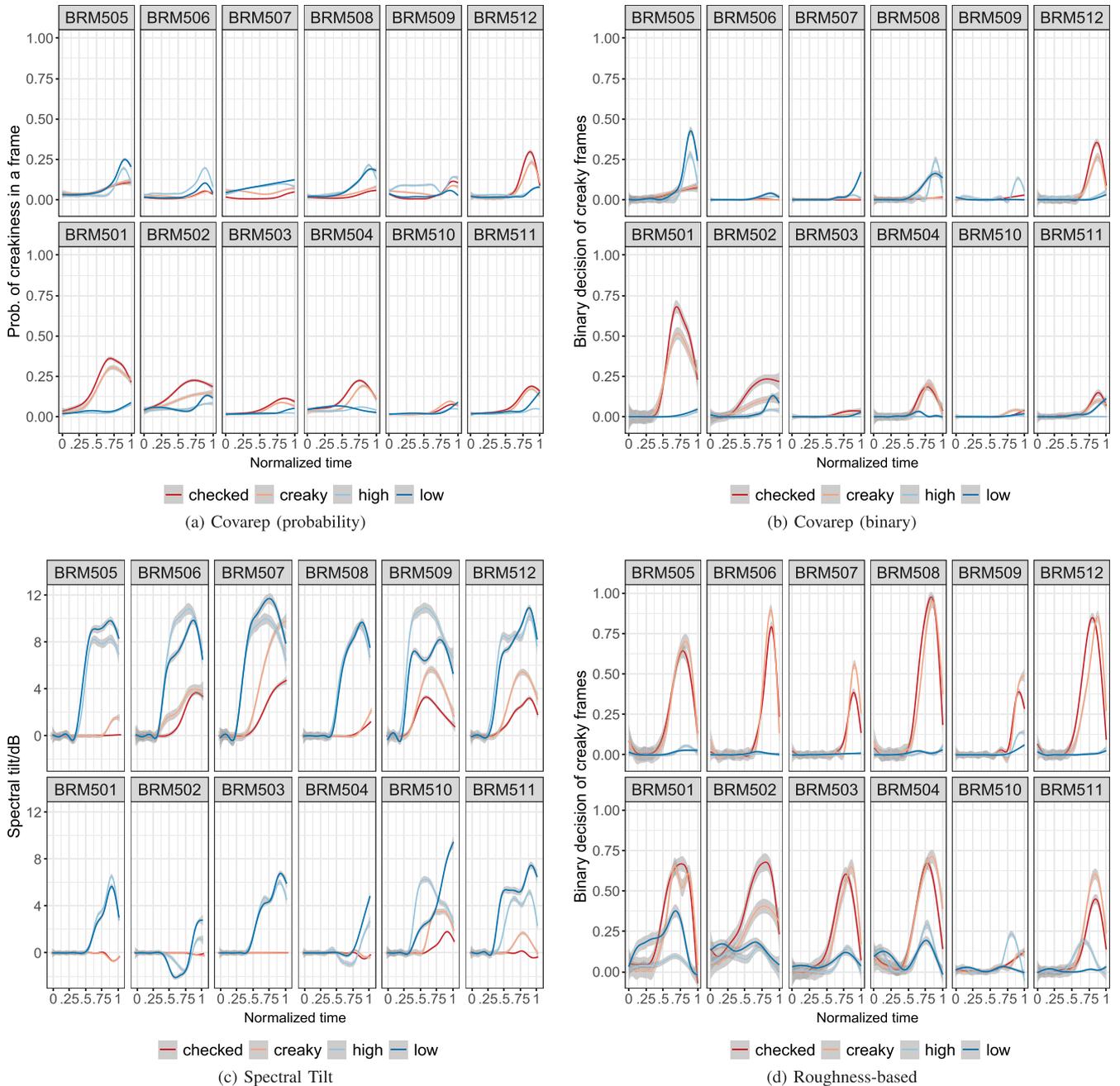


Fig. 4. Smooth conditional means and their corresponding 95% confidence intervals (in gray), computed for Covarep outputs, spectral tilt (i.e.,  $H1^* - H2^*$ ), and roughness-based prediction. Normalized time is used to account for the length differences between utterances. In each panel, top and bottom rows respectively correspond to female and male speakers.

In Table III, we summarize classification accuracy, precision, and recall results in terms of mean and standard deviation ( $SD$ ) of the 12 folds. The best accuracy result of  $94.5\% \pm 3.2\%$  was achieved with a 2-layer RNN and 32 GRU nodes per layer.

2) *RNN Trained on Covarep Features*: As mentioned earlier, the Covarep toolkit can predict the probability of creakiness of a single frame from the 12 features shown in Table I and their first- and second-order time derivatives. To obtain segment-level decisions using Covarep predictions, we used majority rule voting or accumulated log probability score. Unfortunately, in both cases the classification results were close to random,

i.e. about 50%. Apparently, the reason is that the Multi-Layer Perceptron (MLP) used in the toolkit had been trained on quite different data. To perform a fair comparison between Covarep and our approach, we trained a similar RNN using the same 12 features calculated from our data for the same voiced segments. The results we obtained are shown in Table IV. As can be seen, training an RNN with Covarep features yielded slightly better results than those obtained with only roughness contours. However, there are two main differences to take into account: i) roughness contours are unidimensional features (computed from several channels), while Covarep uses 12 disparate features, i.e.,

TABLE III

RNN PERFORMANCE IN TERMS OF ACCURACY, PRECISION, AND RECALL (MEAN AND SD) USING 1-DIMENSIONAL ROUGHNESS DATA

# of nodes	# of RNN layers		
	1	2	3
	Accuracy (%)		
8	93.4 ± 4.7	93.5 ± 4.6	92.4 ± 4.6
16	92.8 ± 5.9	93.0 ± 5.6	92.7 ± 5.7
32	93.3 ± 5.0	<b>94.5 ± 3.2</b>	93.3 ± 4.1
	Precision		
8	0.944 ± 0.040	0.942 ± 0.039	0.933 ± 0.036
16	0.934 ± 0.041	0.939 ± 0.041	0.938 ± 0.036
32	0.942 ± 0.044	<b>0.949 ± 0.032</b>	0.942 ± 0.040
	Recall		
8	0.938 ± 0.046	0.934 ± 0.048	0.923 ± 0.047
16	0.920 ± 0.062	0.928 ± 0.061	0.928 ± 0.052
32	0.933 ± 0.060	<b>0.940 ± 0.039</b>	0.931 ± 0.060

TABLE IV

RNN PERFORMANCE IN TERMS OF ACCURACY, PRECISION, AND RECALL (MEAN AND SD) USING 12-DIMENSIONAL COVAREP DATA

# of nodes	# of RNN layers		
	1	2	3
	Accuracy (%)		
32	93.4 ± 7.1	92.1 ± 8.3	94.4 ± 5.5
64	93.9 ± 6.7	94.5 ± 5.1	<b>94.9 ± 4.8</b>
128	94.1 ± 5.1	93.8 ± 5.2	94.2 ± 5.6
	Precision		
32	0.946 ± 0.049	0.941 ± 0.052	0.956 ± 0.035
64	0.952 ± 0.050	0.953 ± 0.039	<b>0.959 ± 0.046</b>
128	0.950 ± 0.041	0.945 ± 0.046	0.946 ± 0.042
	Recall		
32	0.934 ± 0.074	0.927 ± 0.077	0.948 ± 0.047
64	0.938 ± 0.076	0.947 ± 0.046	<b>0.950 ± 0.063</b>
128	0.941 ± 0.053	0.935 ± 0.060	0.939 ± 0.056

a 12-dimensional feature vector; ii) the standard deviation of the accuracy, precision, and recall are noticeably lower in the case of the RNN trained with roughness contours. Since the standard deviation is sensitive to outliers, skewness, etc. [75], this finding suggests that the RNN trained with roughness contours is more robust against speaker variability.

3) *RNN Trained on Roughness and Covarep Features*: Since the roughness contour and the Covarep features differ in their meaning and extraction methods, there was a possibility that they could convey somewhat different information. If this assumption were true, then when used together, they could improve classification accuracy. To confirm this hypothesis, we trained an RNN with concatenated roughness and Covarep features resulting in a 13-dimensional network input. All the other experimental parameters were the same as in the previous experiments. The results are presented in Table V. Indeed, the classification accuracy of the RNN trained with the combined data improved to 95.6%, but the standard deviation deteriorated with respect to the previous experiments ( $SD = 5.3\%$  cf. 3.2% and 4.8% for the roughness- and Covarep-trained RNNs, respectively).

Additionally, we trained RNNs, where inputs were augmented by the first- and second-order time derivatives of the corresponding features, i.e. roughness contour, 12 Covarep features,

TABLE V

RNN PERFORMANCE IN TERMS OF ACCURACY, PRECISION, AND RECALL (MEAN AND SD) USING 13-DIMENSIONAL DATA (CONCATENATING ROUGHNESS AND COVAREP DATA)

# of nodes	# of RNN layers		
	1	2	3
	Accuracy (%)		
32	94.2 ± 6.1	<b>95.6 ± 5.3</b>	93.1 ± 5.1
64	95.4 ± 5.1	93.2 ± 7.4	93.9 ± 6.8
128	94.8 ± 5.9	94.1 ± 6.1	94.7 ± 5.6
	Precision		
32	0.950 ± 0.040	<b>0.963 ± 0.041</b>	0.941 ± 0.041
64	0.961 ± 0.032	0.948 ± 0.052	0.950 ± 0.048
128	0.957 ± 0.061	0.951 ± 0.046	0.956 ± 0.042
	Recall		
32	0.941 ± 0.064	<b>0.956 ± 0.055</b>	0.931 ± 0.054
64	0.955 ± 0.042	0.932 ± 0.078	0.938 ± 0.069
128	0.949 ± 0.061	0.941 ± 0.064	0.949 ± 0.059

and their combination, resulting in 3-, 36-, and 39-dimensional vectors. Since RNNs are very good at modeling temporal dependencies, as we expected, there was no noticeable change in the performance with respect to the cases without derivatives.

#### IV. DISCUSSION

The combined results of the experiments discussed in Sections III-B and III-C suggest that models used to predict psychoacoustic roughness could also be used as predictors of creaky episodes in speech. Contrasting with the results of a roughness-based classifier, the creakiness detection routine implemented in Covarep failed in several cases to distinguish between creaky and non-creaky tones in our corpus, as shown in Fig. 4. These results persisted regardless of the Covarep's output used.

When using RNNs, the performance of the roughness-based classifier was very similar to that achieved by using the same input as the Covarep predictor. However, we believe that our approach has several advantages with respect to other methods: 1) it uses 1-dimensional data for the prediction, as opposed to multi-dimensional data; 2) having a unique well-defined unit (i.e., asper) eases the comparison of creakiness among different studies, corpora, voices, etc.; perhaps more importantly, 3) psychoacoustic roughness is a perceptual feature, and we argue that it is more related to the phonemic classification made by listeners than other acoustic features: There seems to be several ways in which speakers can produce creaky voice. These articulation variations correlate differently with different acoustic features, such as  $H1-H2$ , CPP, etc. Regardless of how it is produced, creakiness is perceived, nonetheless, under a seemingly single category [12]. Therefore, focusing on later stages in the speech chain (i.e., in the auditory process) may be appropriate to describe and study phonation and phonemic contrast.

We hypothesize that the psychoacoustic roughness prediction of creakiness works best when creakiness in speech is manifested as amplitude modulation (damped pulses, etc.) in vocalic segments, since that is what the psychoacoustic roughness model uses for its predictions. These modulations need not be periodic,

so a roughness-based predictor should perform similarly for both creaky and rough speech, the latter as understood in the context of the CAPE-V framework—that is, “perceived irregularity in the voicing source.”

Having a single dimension, a creakiness predictor based on psychoacoustic roughness, needs no ablation studies since all the output variance can be attributed to its single feature. For multi-dimensional predictors, on the other hand, ablation studies become indispensable to gain insights on the impact or relative weight that a single feature or set of features may have on the output.

Psychoacoustic roughness increases with pressure level, so that “for an increase in sound pressure level by 40 dB roughness increases by a factor of about 3” [52, p. 260]. In our case, we assumed that each frame in the analysis was produced at 80 dB (SPL). Setting the frame SPL to a different value (or making it relative to the maximum amplitude of the signal) would produce changes in the roughness traces as well. For roughness-based classifications not using RNNs, it would be necessary to adjust the thresholds to determine the creakiness of a frame; for RNN implementations, retraining the network would also be necessary. Note that perturbations on phonation such as  $F_0$ -jitter, amplitude shimmer, etc., tend to increase when speech is produced at lower sound pressure levels [76]. Equalizing the frame intensities increases the measured roughness at quiet frames, such as those found towards the end of the vocalic part of our corpus (see Figs. 2(a) and 2(b)).

The roughness model used in this research could predict with fair accuracy the presence of creakiness in speech. This model however, has been amended since originally proposed to account for the size of the auditory filters, i.e., replacing the Bark scale with narrower Equivalent Rectangular Bands—ERBs [77]; the effect of phase (such as those observed by the elicited roughness of a reverse saw-tooth signal) [55]; etc. We would expect that more sophisticated roughness prediction models yield more accurate classifications of creaky voice.

Throughout our research, we relied upon classifications of creaky tokens from dictionary entries of the Burmese language. We manually inspected a sample of words and found that their actual production was, in general, the intended one. Performing an exhaustive review of all tokens and including other corpora may help to improve the accuracy achieved by the proposed classifier. Confirming the robustness of our roughness-based predictor is an ongoing project, including releasing freely available routines for creakiness prediction based on psychoacoustic roughness.

## V. CONCLUSION

It was possible to adequately predict creaky episodes in vocalic segments of speech using a psychoacoustic roughness model. A predictive model based on a Recurrent Neural Network using only psychoacoustic roughness yielded results comparable to those obtained with RNNs trained with higher dimensional data. However, the roughness-based approach yielded lower standard deviation, suggesting robustness to speaker variation.

Additionally, no apparent advantage of augmenting the training data by time derivative features was found, and likewise, including roughness along with the acoustic features used in the multidimensional RNN did not improve prediction performance significantly.

## ACKNOWLEDGMENT

The authors would like to thank K. Otsuka for helping with construction of the word list for recording, and M. Abdar, and K. Naya for helping with annotations. The authors also want to thank M. Cohen, and the anonymous reviewers for offering invaluable comments and suggestions to improve previous versions of this manuscript.

## REFERENCES

- [1] I. R. Titze, *Principles of Voice Production*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.
- [2] D. G. Childers and C. Lee, “Vocal quality factors: Analysis, synthesis, and perception,” *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.
- [3] R. Reiter, T. K. Hoffmann, A. Pickhard, and S. Brosch, “Hoarseness—causes and treatments,” *Deutsches Ärzteblatt Int.*, vol. 112, no. 19, pp. 329–337, 2015.
- [4] S. Sapir, L. O. Ramig, and C. M. Fox, “Intensive voice treatment in Parkinson’s disease: Lee Silverman voice treatment,” *Expert Rev. Neurotherapeutics*, vol. 11, no. 6, pp. 815–830, 2011.
- [5] M. Gordon and P. Ladefoged, “Phonation types: A cross-linguistic overview,” *J. Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.
- [6] S. R. Moisiuk, H. Lin, and J. H. Esling, “A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS),” *J. Int. Phonetic Assoc.*, vol. 44, no. 1, pp. 21–58, 2014.
- [7] R. Ogden, “Non-modal voice quality and turn-taking in Finnish,” in *Sound Patterns in Interaction: Cross-Linguistic Studies From Conversation*. Amsterdam, The Netherlands: John Benjamins, 2004, pp. 29–62.
- [8] S. Lee, “Creaky voice as a phonational device marking parenthetical segments in talk,” *J. Sociolinguistics*, vol. 19, no. 3, pp. 275–302, 2015.
- [9] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, “Habitual use of vocal fry in young adult female speakers,” *J. Voice*, vol. 26, no. 3, pp. e111–e116, 2012.
- [10] P. Keating, M. Garellek, and J. Kreiman, “Acoustic properties of different kinds of creaky voice,” in *Proc. 18 Int. Congr. Phonetic Sci.*, 2015, pp. 0821.1-1–0821.1-5.
- [11] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, “Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers,” *J. Acoust. Soc. Amer.*, vol. 103, no. 5, pp. 2649–2658, 1998.
- [12] B. R. Gerratt and J. Kreiman, “Toward a taxonomy of nonmodal phonation,” *J. Phonetics*, vol. 29, no. 4, pp. 365–381, 2001.
- [13] I. R. Titze, “Acoustics of creaky voice,” *J. Acoust. Soc. Amer.*, vol. 86, no. S1, pp. S26–S26, 1989.
- [14] H. von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York, NY, USA: Dover, 1954.
- [15] P. Daniel and R. Weber, “Psychoacoustical roughness: Implementation of an optimized model,” *Acta Acustica United Acustica*, vol. 83, pp. 113–123, 1997.
- [16] C. de Bruijn and S. Whiteside, “Effect of experience levels on voice quality ratings,” in *Proc. Phonetics Teaching Learn. Conf.*, 2007, pp. 17e.1-1–17e.1-3.
- [17] S. S. Stevens and E. H. Galanter, “Ratio scales and category scales for a dozen perceptual continua,” *J. Experimental Psychol.*, vol. 54, no. 6, pp. 377–411, 1957.
- [18] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, “Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,” *J. Speech Hearing Res.*, vol. 36, pp. 21–40, 1993.
- [19] D. A. Eddins, L. M. Kopf, and R. Shrivastav, “The psychophysics of roughness applied to dysphonic voice,” *J. Acoust. Soc. Amer.*, vol. 138, no. 6, pp. 3820–3825, 2015.
- [20] P. B. Denes and E. N. Pinson, *The Speech Chain*, 2nd ed. New York, NY, USA: WH Freeman, 1993.

- [21] J. C. Catford, *Phonation Types: the Classification of Some Laryngeal Components of Speech Production*. Harlow, U.K.: Longmans, 1964.
- [22] P. Ladefoged, *Preliminaries to Linguistic Phonetics*. Chicago, IL, USA: University of Chicago Press, 1971.
- [23] J. H. Esling, J. G. Harris, and J. Romero, "An expanded taxonomy of states of the glottis," in *Proc. 15th Int. Congr. Phonetic Sci.*, 2003, vol. 1, pp. 1049–1052.
- [24] J. Laver, *The Phonetic Description of Voice Quality* (Cambridge Studies in Linguistics). Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [25] H. Eckert and J. Laver, *Menschen und ihre Stimmen (People and Their Voices)*. (in German). Weinheim, Germany: Beltz/Psychologie Verlags Union, 1994.
- [26] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, 1990.
- [27] J. A. Edmondson and J. H. Esling, "The valves of the throat and their functioning in tone, vocal register and stress: Laryngoscopic case studies," *Phonology*, vol. 23, no. 2, pp. 157–191, 2006.
- [28] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *J. Acoust. Soc. Amer.*, vol. 101, no. 1, pp. 466–481, 1997.
- [29] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 1064–1077, 1999.
- [30] H. M. Hanson, K. N. Stevens, H.-K. J. Kuo, M. Y. Chen, and J. Slifka, "Towards models of phonation," *J. Phonetics*, vol. 29, no. 4, pp. 451–480, 2001.
- [31] P. Keating, C. Esposito, M. Garellek, S. Khan, and J. Kuang, "Phonation contrasts across languages," in *UCLA Work. Papers Phonetics*, no. 108, 2010, pp. 188–202.
- [32] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, 2014.
- [33] B. Blankenship, *The Time Course of Breathiness and Laryngealization in Vowels*. Ph.D. thesis, Dept. Linguistics, University of California, Los Angeles, Los Angeles, CA, USA, 1997.
- [34] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 1, pp. I–333.
- [35] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert, W. Bigenzahn, and B. Schneider-Stickler, "A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, 2016, pp. 767–770.
- [36] H. Avelino, "Acoustic and electroglottographic analyses of nonpathological, nonmodal phonation," *J. Voice*, vol. 24, no. 3, pp. 270–280, 2010.
- [37] S. Vishnubhotla and C. Y. Espy-Wilson, "Automatic detection of irregular phonation in continuous speech," in *Proc. Interspeech*, 2006, pp. 949–952.
- [38] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 47–56, Jan. 2008.
- [39] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Comput. Speech Lang.*, vol. 27, no. 4, pp. 1028–1047, 2013.
- [40] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1233–1253, 2014.
- [41] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
- [42] Mathworks, "Matlab." Software, 2019. [Online]. Available: Available from www.mathworks.com (Oct. 14, 2019).
- [43] H. Mori, M. Fujimoto, T. Asai, and K. Maekawa, "Automatic classification of phonation type in the corpus of spontaneous Japanese," (in Japanese), in *Proc. Lang. Resources Workshop*, 2017, pp. 347–354.
- [44] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
- [45] D.-K. Mac, T.-L. Nguyen, A. Michaud, and D.-D. Tran, "Influences of speaker attitudes on glottalized tones: A study of two Vietnamese sentence-final particles," in *Proc. 18th Int. Congr. Phonetic Sci.*, Glasgow, U.K., Aug. 2015, pp. 0650.1-1–0650.1-5.
- [46] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and non-modal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 12, pp. 2281–2291, Dec. 2017.
- [47] E. Weed, R. Fusaroli, J. Mayo, and I.-M. Eigsti, "Creaky voice in adolescents with autism spectrum disorder: An acoustic, quantitative analysis," in *Proc. INSAR 2019*, 2019.
- [48] O. Murton, S. Shattuck-Hufnagel, J.-Y. Choi, and D. D. Mehta, "Identifying a creak probability threshold for an irregular pitch period detection algorithm," *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. EL379–EL385, 2019.
- [49] J. Kuang, "The influence of tonal categories and prosodic boundaries on the creakiness in mandarin," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL509–EL515, 2018.
- [50] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*. Shelter Island, NY, USA: Manning Publications, 2012.
- [51] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acustica*, vol. 30, pp. 201–213, 1974.
- [52] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin: Springer, 3rd ed., 2006.
- [53] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *J. Acoust. Soc. Amer.*, vol. 38, no. 4, pp. 548–560, 1965.
- [54] E. Zwicker, G. Flottorp, and S. Stevens, "Critical bandwidth in loudness summation," *J. Acoust. Soc. Amer.*, vol. 29, no. 5, pp. 548–557, 1957.
- [55] D. Pressnitzer and S. McAdams, "Two phase effects in roughness perception," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2773–2782, 1999.
- [56] W. von Aures, "Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgrößen (Sensory pleasantness as a function of psychoacoustic factors)," (in German), *Acustica*, vol. 58, pp. 268–281, 1985.
- [57] Y. Wang, G. Shen, H. Guo, X. Tang, and T. Hamade, "Roughness modelling based on human auditory perception for sound quality evaluation of vehicle interior noise," *J. Sound Vibration*, vol. 332, no. 16, pp. 3893–3904, 2013.
- [58] C. Stumpf, *Konsonanz und Dissonanz, Beiträge zur Akustik und Musikwissenschaft (Consonance and dissonance)*, (in German), vol. 1, 1898, pp. 1–108.
- [59] J. Villegas and M. Cohen, "Roughness minimization through automatic intonation adjustments," *J. New Music Res.*, vol. 39, no. 1, pp. 75–92, 2010.
- [60] J. Villegas and M. Cooke, "Maximising objective speech intelligibility by local f0 modulation," in *Proc. Interspeech*, Sep. 2012, pp. 1704–1707.
- [61] G. von Békésy, "Zur Theorie des Hörens. Die Schwingungsform der Basilarmembrane (The theory of hearing: oscillation of the basilar membrane)," (in German), *Physik. Zeits.*, vol. 29, pp. 793–810, 1928.
- [62] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, no. 4, pp. 128–34, 1951.
- [63] Y. I. Fishman *et al.*, "Consonance and dissonance of musical chords: Neural correlates in auditory cortex of monkeys and humans," *J. Neurophysiol.*, vol. 86, pp. 2761–2788, 2001.
- [64] W. Sethares, *Tuning, Timbre, Spectrum, Scale*, 2nd ed. Berlin, Germany: Springer, 2005.
- [65] P. N. Vassilakis, *Perceptual and Physical Properties of Amplitude Fluctuation and Their Musical Significance*. Ph.D. thesis, Dept. Ethnomusicology, University of California, Los Angeles, Los Angeles, CA, USA, 2001.
- [66] W. von Aures, "Model for calculating sensory euphony of various sounds," (in German), *Acustica*, vol. 59, pp. 130–141, 1985.
- [67] J. E. Schrader, "A MATLAB implementation of a model of auditory roughness," Master's thesis, Dept. Technologie Manage., Eindhoven University of Technology, Eindhoven, The Netherlands, 2002.
- [68] J. F. Gruber, *An Articulatory, Acoustic, and Auditory Study of Burmese Tone*. Ph.D. thesis, Dept. Linguistics, Georgetown University, Washington, DC, USA, 2011.
- [69] J. Watkins, "Can phonation types be reliably measured from sound spectra? some data from Wa and Burmese," *SOAS Working Papers Linguistics*, vol. 7, pp. 321–339.
- [70] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2019. Accessed: Oct. 14, 2019. [Online]. Available: www.praat.org.
- [71] M. Gamer, J. Lemon, and I. F. P. Singh, *Irr: Various Coefficients of Interrater Reliability and Agreement*, R package version 0.84.1, 2010. [Online]. Available: https://CRAN.R-project.org/package=irr. Accessed: Oct. 31, 2019.
- [72] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. Accessed: Oct. 14, 2019. [Online]. Available: Available from www.R-project.org.
- [73] Y. Shue, *The Voice Source in Speech Production: Data, Analysis and Models*. Ph.D. thesis, Dept. Elect. Eng., University of California Los Angeles, Los Angeles, CA, USA, 2010.
- [74] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York, NY, USA: Springer, 2009. [Online]. Available: http://had.co.nz/ggplot2/book.

- [75] J. Högel, W. Schmid, and W. Gaus, "Robustness of the standard deviation and other measures of dispersion," *Biometrical J.*, vol. 36, no. 4, pp. 411–427, 1994.
- [76] M. Brockmann-Bauser, J. Bohlender, and D. Mehta, "Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders," *J. Voice*, vol. 32, no. 2, pp. 162–168, 2018.
- [77] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, 1983.



**Jeremy Perkins** received the Ph.D. degree in phonology from Rutgers University, New Brunswick, NJ, USA, in 2013. Since then, he has been a Faculty member with the Center for Language Research, University of Aizu, Aizuwakamatsu, Japan, where he is currently a Senior Associate Professor. His research interests include the phonetics and phonology of tone in East Asian languages including Thai, Zhuang and Burmese. He is particularly interested in the role creaky phonation plays in tone languages and in the interactions between tone and consonants.



**Julián Villegas** received the Sc.B. degree in electronic engineering from the University of Valle-Cali, Colombia, and the M.Sc. and Ph.D. degrees in computer sciences from the University of Aizu, Aizuwakamatsu, Japan. He is currently an Associate Professor with the Computer Arts Laboratory, University of Aizu. His research interests include speech intelligibility, music and sound, psychoacoustics, and spatial sound.



**Seunghun J. Lee** received the Ph.D. degree from Rutgers University, New Brunswick, NJ, USA, in 2008. After teaching at Central Connecticut State University, he is currently Associate Professor of Linguistics with the International Christian University, Mitaka, Japan. His research interests include phonetics and phonology of less researched languages, including but not limited to Xitsonga (South Africa), Drenjongke (India), Burmese (Myanmar), as well as dialects in Korean and Japan. He regularly presents at international conferences such as ICPHS, ICCP,

SICSS. He is also a Reviewer for theoretical as well as descriptive linguistic journals.



**Konstantin Markov** (M'04) received the Ph.D. degree from the Toyohashi University of Technology, Toyohashi, Japan, in 1999. After that, he was a Research Scientist with the Advanced Telecommunications Research Institute, Kyoto, Japan, till 2009. Since then, he has been a Faculty of the University of Aizu, Aizuwakamatsu, Japan, where he is currently a Professor in Information Sciences. His research interests include machine learning, deep learning, and signal processing with applications to speech, music, and natural language processing. He has been a Program

Member of various international conferences such as Interspeech, ICASSP, EUSIPCO, and SpeCom as well as Reviewer for several IEEE and Elsevier scientific journals.