Speech Communication 147 (2023) 74–81

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Psychoacoustic features explain creakiness classifications made by naive and non-naive listeners☆

Julián Villegas [a,*], Seunghun J. Lee [b], Jeremy Perkins [c], Konstantin Markov [a]

[a] Department of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Fukushima, 965-8580, Japan
[b] Department of Psychology and Linguistics, International Christian University, Tokyo 181-8585, Japan
[c] Centre for Language Research, University of Aizu, Aizu-Wakamatsu, Fukushima, 965-8580, Japan

ARTICLE INFO

Keywords:
Creakiness
Psychoacoustic features
Subjective classification

ABSTRACT

We compared the classifications of words between creaky and non-creaky made by two groups of listeners whose native languages differ on whether creakiness is used for phonemic contrast. Japanese (naive group) and Vietnamese (non-naive group) listeners classified words produced by a single speaker of Du'an Zhuang, an unintelligible language for both groups that shares some tonal similarities with Hanoi Vietnamese. We found little differences between the accuracy achieved by the two groups, and that the non-naive group was more likely to rate words as creaky relative to the naive group. In addition, there seems to be no benefit of background linguistic knowledge on the accuracy with which the non-naive group classified words. The sensitivity to creakiness observed in classifications made by experts (inspecting the waveforms and spectrograms in addition to listening to the words) and made by a machine learning algorithm based on these kind of classifications was unmatched by both cohorts. This result challenges the perceptual validity of such refined classifications commonly used in phonation studies. In addition, we found a positive association between psychoacoustic roughness and the probability of a word to be judged as creaky. We also found a positive association of loudness with creaky judgments, whereas pitch was negatively associated. We found no evidence of sharpness and creaky association. Finally, the accuracy to predict subjective creakiness via a recurrent neural network classifier was best when the traces of all the considered psychoacoustic features were included as predictors.

1. Introduction

Researchers interested in phonation and its contrastive role in different languages often draw upon predicting models relating phonation with acoustic features like spectral tilt (the slope of speech's power spectral density) (Keating et al., 2010) or articulatory features (e.g., open quotient: the fraction of a glottal cycle during which the glottis is open) (Klatt and Klatt, 1990). It has been noted that there seem to be many ways in which speakers produce a given phonation: in the case of creaky voice, several subcategories can be identified according to fundamental frequency, spectral tilt, Subharmonic-to-Harmonic Ratio (SHR), etc. (Keating et al., 2015).

Differences in articulation are reflected in acoustic features, hence the performance of models predicting phonation is speaker-dependent and their accuracy varies depending on the predominant articulation within a group of speakers. This performance variation hinders comparison across studies since researchers may choose models that best suit their data. As an example, Kuang and Keating (2012) reported that Kong (2001) found that spectral tilt measured as the magnitude difference between the first and second harmonic (H_1–H_2) can be successfully used to distinguish between tense and lax phonation in several Tibeto-Burman languages such as Northern Yi, Zaiwa, and Jingpo. They also noted that measurements of spectral tilt as the difference between the first harmonic and the harmonic closest to the first or second formant (H_1–A_1 or H_1–A_2) are better predictors in Northern Yi. One can imagine situations where one spectral tilt (e.g., H_1–H_2) is used to study one language whereas a different tilt measurement (e.g., H_1–A_1) is used for a different language. While both studies may report their findings in terms of spectral tilt using the same units (dB), it would be difficult to directly compare their results as the spectral tilt was measured differently in each study.

To ameliorate this problem, models merging different acoustic features and heuristics have been proposed in what is known as automatic

☆ This research was supported by the JSPS Kakenhi Grant No. 20K11956.
* Corresponding author.
E-mail addresses: julian@u-aizu.ac.jp (J. Villegas), seunghun@icu.ac.jp (S.J. Lee), jperkins@u-aizu.ac.jp (J. Perkins), markov@u-aizu.ac.jp (K. Markov).

https://doi.org/10.1016/j.specom.2023.01.006
Received 27 May 2022; Received in revised form 16 January 2023; Accepted 20 January 2023
Available online 25 January 2023

prediction of phonation. For example, an Artificial Neural Network (ANN) is used for creakiness prediction in Covarep (Degottex et al., 2014), a Matlab (Mathworks, 2022) library comprising several routines for speech analysis. This ANN is based on a model proposed by Drugman et al. (2014) which, in addition to spectral tilt measurements (H_1–H_2), uses eleven additional features and their first- and second-time derivatives as inputs for the prediction. Some of the features included in their model are intra-frame periodicity and inter-pulse similarity (Ishi et al., 2008); the presence of secondary excitation peaks in the residual signal of a linear prediction filter; and the residual peak prominence (Kane et al., 2013). Although these automatic classifiers achieve great accuracy, they do not offer insights into the underlying features driving their classification.

An additional concern is raised from the fact that using acoustic or articulatory features to study phonation informs very little on the perceptual validity that level differences within a given feature may have. I.e., differences on the production side of the speech chain (Denes and Pinson, 1993) may not be auditorily relevant.

Speakers using phonemic contrast may differ physiologically in the way they produce such a contrast, but ultimately, their interlocutors are able to distinguish different phonations when necessary. Moreover, whereas different types of phonation are used for contrast in several languages, there is so far no evidence of languages using sub-types of the same phonation contrastively. Thus, auditory features of speech at the listener's end of the speech chain, as opposed to unprocessed acoustic, articulatory, or physiological features at the speaker's end, could be more suitable for predicting phonation.

The idea of using perceptual attributes for phonation classification resonates with that proposed by Kreiman et al. (2014). They propose a model linking production and perception of speech comprised of four components to describe different voice qualities. These components are associated with one or more voice synthesis parameters, all of which are derived from audio recordings directly (i.e., the acoustic signal) and not from actual auditory features which is the subject of our study.

Standardized psychoacoustic models map the non-linear relationships between physical quantities (frequency, pressure level, etc.) and their auditory counterparts (pitch, loudness, etc.) (Fastl and Zwicker, 2006). In this study we focus on loudness ("that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud" (American National Standards Institute, 2013)), pitch ("that attribute of auditory sensation by which sounds are ordered on the scale used for melody in music" (American National Standards Institute, 2013)), roughness ("subjective response to the perception of rapid—15–300 Hz—amplitude modulation of a sound" (American National Standards Institute, 2013)), and sharpness (subjective response to the spectral centroid of a sound (Fastl and Zwicker, 2006)).

Importantly, psychoacoustic models take into account the effect of interactions between acoustic features on auditory ones. E.g., the same sound pressure level yields different loudness values depending on the frequency content of a stimulus. Psychoacoustic models predict the most probable outcome reported by a subject for a given set of acoustic features. They also provide units (mel, sone, etc.) and scales that allow the direct comparison of disparate phenomena consistently. In addition, these models define absolute thresholds and Just-Noticeable Differences (JNDs) that are useful for assessing the perceptual relevance of a change in a physical quantity.

We have successfully demonstrated that accurate prediction of creakiness can be achieved using a single psychoacoustic feature, namely, roughness (Villegas et al., 2020). We found that a classification method that used only the roughness contours in vocalic parts of speech, yielded similar accuracy to that obtained with higher dimensional methods which also included time derivatives.

Despite the relative success of using roughness for creakiness prediction, it is uncertain whether this psychoacoustic feature is actively used by listeners to identify creakiness in speech. Moreover, psychoacoustic features are universal but the use of phonation for distinguishing between words is not. It is not clear whether listeners routinely exposed to creakiness in their own language for distinguishing between words are more accurate than listeners of languages that do not use it to detect creaky words in an unintelligible language. If no accuracy difference between the two groups is observed, could psychoacoustic features explain their classification? We hypothesize that the universality of psychoacoustic features prevails. I.e., regardless of native language, listeners assess creakiness in a similar fashion. Therefore, their classifications should be similar and should resemble those made by a psychoacoustic-based predictor.

To test these hypotheses, we conducted a series of experiments in which we asked Japanese and Vietnamese assessors (naive and non-naive groups, respectively) to distinguish between creaky and non-creaky words produced in an unintelligible language for the two groups.

This research is important because it uses well-known standardized psychoacoustic models for the prediction of phonation, therefore focusing on the listener end of the speech chain in a perception-based classification. The results presented here could be used for improving the automatic detection and classification of creaky speech, and it could be extended to other kinds of phonation.

2. Method

2.1. Apparatus

The experiments were conducted online using an open-source survey tool (LimeSurvey Project Team / C. Schmitz, 2016). Each cohort was presented with a version of the same survey in its respective language. I.e., all the instructions and feedback were either in Japanese or Vietnamese, depending on the group.

The Japanese cohort was tested first. This group was attending a class so data was collected in parallel (each participant in a different workstation) through the online survey under the supervision of one of the authors. The stimuli were reproduced using circumaural headphones (Sennheiser PC 131) attached to iMac computers in a quiet and otherwise regular classroom. The Vietnamese cohort was tested using the same online survey and recruited via social media, acquaintances, etc. because of the unavailability of assessors in our surroundings. Data from this group were mostly recorded after the beginning of the COVID-19 pandemic. We did not gather the participants at the same venue or supervise them. We asked participants to perform the task from a computer, in a quiet place, using any headphones available to them. It is possible that some Vietnamese assessors did not follow these instructions faithfully.

2.2. Materials

We used 129 words produced by a single male native speaker of Du'an Zhuang, a dialect of Zhuang, which is a Tai-Kadai language spoken in Guangxi, an autonomous region in South Central China. These words were selected from a larger corpus described by Perkins et al. (2016). Most words were recorded as WAV files at 48 kHz/16-bit; 35 of them were inadvertently recorded at 44.1 kHz/16-bit. Vocalic parts were manually annotated in Praat (Boersma and Weenink, 2022). Creakiness contours were estimated using the aforementioned ANN method available in Covarep. The creakiness contours of the words used in our experiments are presented in Fig. 1. Although it is uncertain whether creakiness is phonemically contrastive in Du'an Zhuang, creakiness was systematically used by this consultant. For example, while tone 3 and tone 6 were produced with similar falling pitch contours, tone 6 was notably creakier.

Words used in the experiments were independently inspected by the first three authors. The number of words per tone and the number of words deemed as creaky in each tone are summarized in Table 1. We listened (an unlimited number of times) and visually inspected

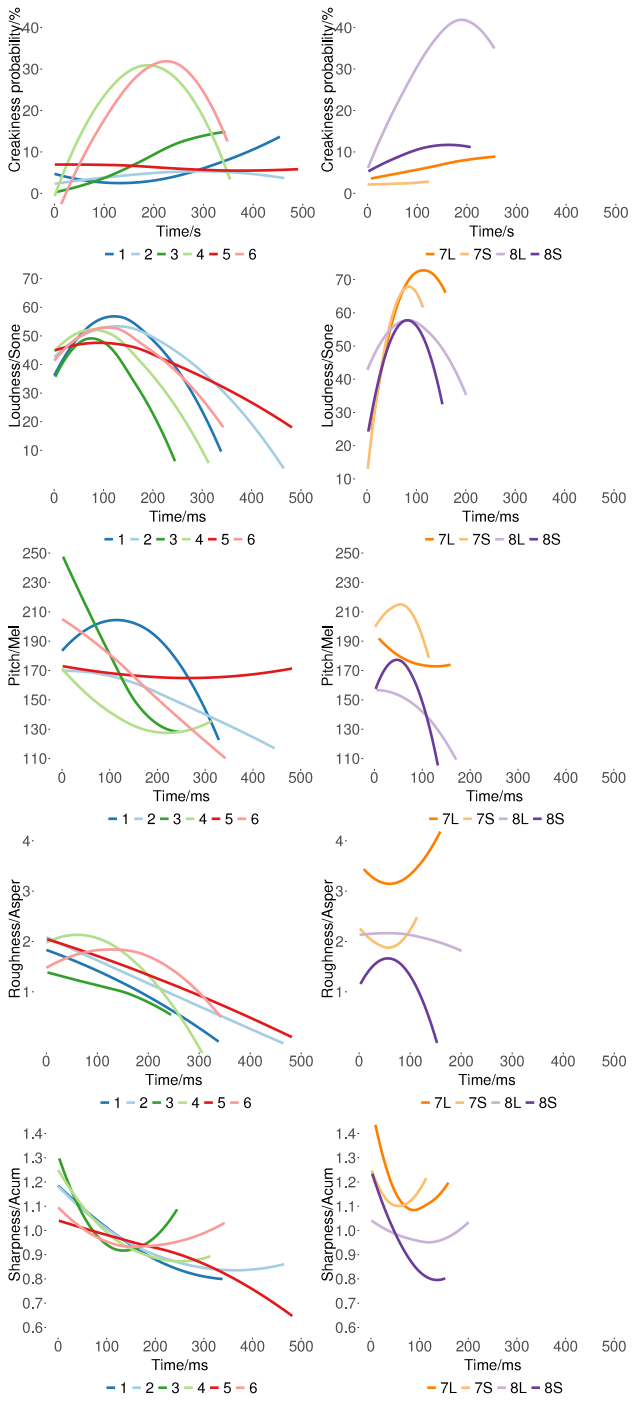

Fig. 1. Locally estimated scatterplot smoothings for creakiness, loudness, pitch, roughness, and sharpness produced by a Du'an Zhuang speaker. Unchecked and checked tones are in the left panels and right panels, respectively. Contours were computed only on the words used in the experiments.

waveforms and spectrograms of the vocalic part of each word to determine whether it was creaky or not. In 107 words (83%) we coincided unanimously in our judgment, in other cases, the final classification of a word was obtained by consensus. 62 words (48.06%) were classified as creaky and there was good agreement among the judges, as indicated by a Light's Kappa analysis ($\kappa = .774, z = 6.88, p < .001$). The materials used in our experiments are available upon request.

Table 1
Total number of words and number of creaky words per Du'an Zhuang tone in our sample.

Tone	1	2	3	4	5	6	7L	7S	8L	8S
Total	27	25	6	31	9	15	1	7	4	4
Creaky	8	4	5	30	0	12	0	0	2	1

2.3. Assessors

We recruited two groups sampled from populations whose language does or does not use creakiness as a phonemic contrast, namely Vietnamese and Japanese. Permission for performing these experiments was obtained following the University of Aizu ethics guidelines.

2.3.1. Naive group

The naive cohort was comprised of 41 volunteer Japanese students, recruited from a single academic course, with no apparent hearing issues (self-reported). They were on average 21 years of age (SD = 1.61), and they were mostly males (only one female). Additional data collected from two participants were excluded from further analysis; one of them reported having a hearing impairment, and the second did not finish the survey. Participants were compensated with credits for the course in which they were enrolled. Japanese does not use creakiness for phonemic contrast.

2.3.2. Non-naive group

The non-naive group comprised 40 Vietnamese assessors (19 of them were female) with no apparent hearing issues (also self-reported). Data from 48 additional participants was disregarded as they abandoned the survey without completing it. The remaining group was on average 29 years of age (SD = 9.56), 31 assessors (78%) lived (or had lived before moving abroad) in Hanoi, or cities with the same dialect. Three assessors (7.5%) lived in provinces speaking South Vietnamese, while three others lived in provinces speaking Mid-central Vietnamese. One assessor did not report this information. All assessors reported having no apparent accent, as perceived by their peers. The same claim was made regarding their parents. No compensation was given for their participation as a way to ensure that only genuinely interested assessors would take part in the experiment. This was crucial given the lack of control in the experimental setup for this cohort.

It has been reported that Vietnamese uses tones and phonation to distinguish between words (Loi and Edmondson, 1998; Pham, 2003; Kirby, 2011). The frequency of occurrence of creaky tones in this language (nặng and ngã) is about 27% according to Vo (Vo, 1997). Critically for this study, some tones in Hanoi Vietnamese are comparable to those found in the 129 words produced by the Du'an Zhuang consultant, as summarized in Table 2. The final column of this table shows the Du'an Zhuang tone number for this speaker that most closely resembled each Vietnamese tone. It should be noted that the pitch contours found in our limited corpus differ from those reported in previous research on Du'an Zhuang (Castro and Hansen, 2010; Li, 2011).

Hanoi Vietnamese is considered the *de facto* standard dialect of Vietnamese. National television is broadcast in this dialect, so even if assessors came from a different linguistic background, they are considered perceptually non-naive in these experiments. That is, regardless of their production differences (speakers of Southern Vietnamese use glottalization minimally) Vietnamese speakers are perceptually capable of distinguishing creaky from non-creaky tones (Brunelle, 2009).

2.4. Procedure

The main task for the assessors was to categorize words as creaky or non-creaky by listening to each word an unlimited number of times. Initially, assessors were presented with instructions about the experiment and answered a questionnaire about their linguistic background,

Table 2
Approximate correspondence between Hanoi Vietnamese tones and those produced by the Du'an Zhuang consultant.

Hanoi	Description	Chao Contour	Du'an
Ngang	Level	˧ (33)	5
Huyền	Mid falling	˨˧ (32)	2
Nặng	Low glottalized	˨ (22)	6
Nặng	Low checked	˨ (21)	8L
Hỏi	Low falling	˧˩˨ (312)	4

hearing impairment, age, biological sex, etc. Continuing with the experiment, two examples of creaky and non-creaky words were presented to the assessors without further definition nor additional information about creakiness. We included a practice session with 20 tokens (half of which were creaky) where feedback on their answers was provided. The practice words were from several languages where creaky voice is used as a contrastive feature (Kambaata, Zapotec, etc.) extracted mainly from the UCLA Phonetics Lab Data (Ladefoged, 2016). After the practice block, the Du'an Zhuang words were presented. Assessors progressed in the task at their own accord and they all finished in less than 30 minutes. Words within blocks were randomly permuted per participant.

2.5. Psychoacoustic features

We computed the temporal trends of psychoacoustic features within the vocalic parts of the words every 10 ms. Prior to this analysis, recordings were 0 dB (Full-scale—FS) normalized, removing any DC offset. When needed, recordings were resampled at 48 kHz. Since the actual Sound Pressure Level (SPL) on the recordings was unknown, we assumed that the recording level was such that a 1 kHz tone at 0 dB (FS) was produced by a 100 dB (SPL) tone. We computed traces of psychoacoustic loudness, sharpness, pitch, and roughness. For the former two features, the recordings were assumed to be free of reverberation (i.e., free-field recordings).

Loudness was computed according to the model proposed by Zwicker (International Organization for Standardization, 2017). Sharpness was computed according to the DIN 45692:2009 standard (German Institute for Standardization, 2009). To compute pitch, we first computed F_0 using the method proposed by Kawahara et al. (2016), as implemented in the TANDEM-Straight package. For this analysis, we assumed extreme values of 20 and 200 Hz. F_0 values were then transformed into pitch using the mel scale model (Stevens et al., 1937). Roughness was estimated using Daniel and Weber's model (Daniel and Weber, 1997), as implemented by Schrader (2002). This implementation was found to closely match empirical results reported by von Aures (1985) on the roughness of AM sinusoids at different modulation frequencies and different frequency bands.

3. Results

3.1. Subjective accuracy

Collected data were analyzed through a series of linear mixed models. Computation of these models was done using the glmer routine in the lme4 library (Bates et al., 2015) of R (R Core Team, 2022). The dependent variable was the accuracy with which an assessor classified a word (0: failure, 1: success) using as reference the classification made by the authors. Beginning with the dependent variable predicted only by assessor as a random factor, we increased the complexity of the predicting model by means of a stepwise forward selection procedure.

The final model included a random intercept per assessor and a random slope of group within word with correlated intercept (Barr et al., 2013). In addition, tone $(1, \ldots, 8S)$, group (naive vs. non-naive), and their interaction were included as fixed effects. The goodness of fit was confirmed with diagnostics available in the DHARMa library (Hartig, 2022).

Table 3
Contrasts for group accuracy per tone computed as the log-ratio between naive/non-naive probabilities. Degrees of freedom in these analysis were computed asymptotically.

Tone	Odds ratio	SE	z-ratio	p-value
8S	1.493	0.595	1.007	.314
8L	1.368	0.534	0.804	.421
7S	1.585	0.501	1.458	.145
7L	1.175	0.897	0.211	.833
6	0.824	0.178	−0.894	.371
5	1.897	0.526	2.308	**.021**
4	0.590	0.095	−3.280	**.001**
3	0.795	0.264	−0.691	.490
2	1.491	0.260	2.293	**.022**
1	1.365	0.240	1.772	.076

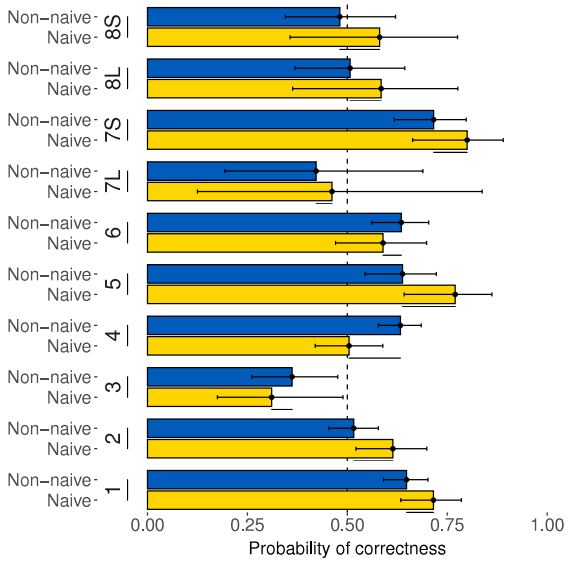

Fig. 2. Estimated marginal mean probability of correctness by tones and group. Here and in other figures, error bars correspond to the 95% CI.

Tone has a significant effect on accuracy $[\chi^2(9) = 45.5, p < .001]$, but Group does not $[\chi^2(1) = 0.4, p < .530]$ (in this study, significance level was set to $\alpha = .050$). Their interaction, however, has a significant effect too $[\chi^2(9) = 35.7, p < .001]$. Post hoc analysis based on Tukey's honest significant difference between estimated least-square means was computed with the library emmeans (Lenth, 2022). Because of the large number of observations, the degrees of freedom in these comparisons were computed asymptotically. We focused our post hoc analysis on the differences between groups in each tone, as summarized in Table 3 and Fig. 2.

Both groups performed similarly in the majority of tones. However, whereas the non-naive group was more accurate in judging tone 4, the naive cohort was more accurate in judging tones 2 and 5, as illustrated in Fig. 2. For tones 1, 5, and 7S both groups achieved an accuracy above 50%. In addition, the naive group also achieved accuracy higher than chance in judging tone 2, while the non-naive group did likewise for tones 4 and 6. For tone 3, on the other hand, both groups achieved an accuracy below 50%. Tone 3 is thus exceptional, as it was the only tone with accuracy below 50% in both groups.

3.2. Similarity between judgments

We were also interested in finding similarities between the classifications made by the two groups regardless of correctness. We built a

Table 4
Contrasts for group opinions per tone computed as the log-ratio between naive/non-naive probabilities. Degrees of freedom in these analysis were computed asymptotically.

Tone	Odds ratio	SE	z-ratio	p-value
8S	0.473	0.163	−2.173	**.030**
8L	0.559	0.188	−1.726	.084
7S	0.533	0.159	−2.111	**.035**
7L	0.843	0.494	−0.292	.770
6	0.648	0.154	−1.833	.067
5	0.452	0.123	−2.910	**.003**
4	0.566	0.120	−2.683	**.007**
3	0.423	0.131	−2.788	**.005**
2	0.414	0.090	−4.041	**<.001**
1	0.272	0.061	−5.848	**<.001**

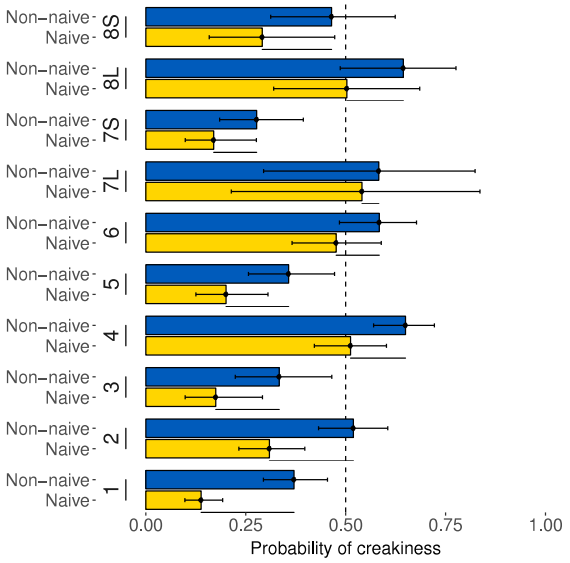

Fig. 3. Estimated marginal mean probability of subjective creakiness per group and tone. Non-naive group was in general more liberal than the naive group.

similar model as the one described in Section 3.1 using as dependent variable the binary classification made by the assessors and not its accuracy.

Group $[\chi^2(1) = 16.2, p < .001]$, Tone $[\chi^2(9) = 114.0, p < .001]$, and their interaction $[\chi^2(9) = 30.8, p < .001]$ have significant effects on the judgments. The post hoc analysis, summarized in Table 4 and Fig. 3, indicates that the non-naive group was more liberal than the naive group in their creakiness judgments except in tones 6, 7L, and 8L, where the opinions of both groups were similar.

3.3. Subjective creakiness and psychoacoustics

To weigh the total effect of different psychoacoustic features on the subjective classifications regardless of group, we computed the mean value of loudness, pitch, roughness, and sharpness for the vocalic portions of each word. With these values, we created a generalized linear mixed model as before with the outcome being the binary classification of the assessors. The effects of assessor and word were considered random, and those of the psychoacoustic features, fixed.

Psychoacoustic roughness $[\chi^2(1) = 18.306, p < .001]$, loudness $[\chi^2(1) = 20.058, p < .001]$, and pitch $[\chi^2(1) = 72.777, p < .001]$ have significant effects on the binary classifications made by the assessors; on the other hand, sharpness does not $[\chi^2(1) = 1.6525, p = .199]$. As shown in Fig. 4, higher mean values of loudness and roughness increased the probability of a word being rated as creaky. In contrast, higher mean pitch values increased the probability of a word being rated as non-creaky. This is not surprising since, as noted before, creaky voice is

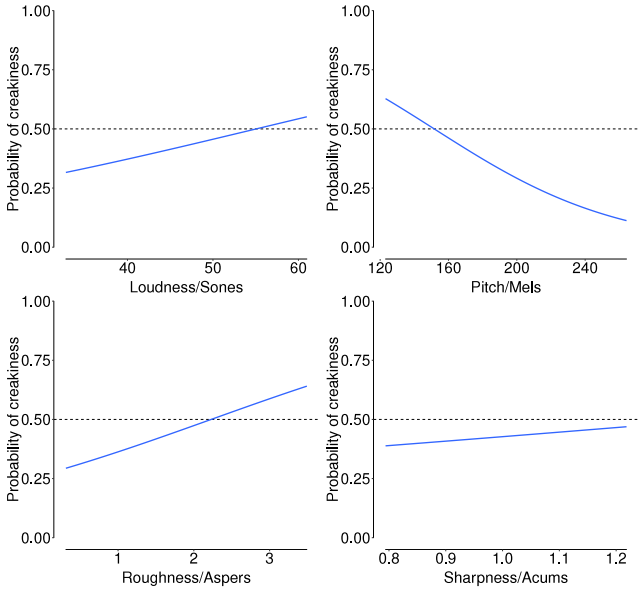

Fig. 4. Effect of the mean value of each psychoacoustic feature (computed across the vocalic portions of each word) on the probability of rating such a word as creaky. The effect of sharpness is not significant.

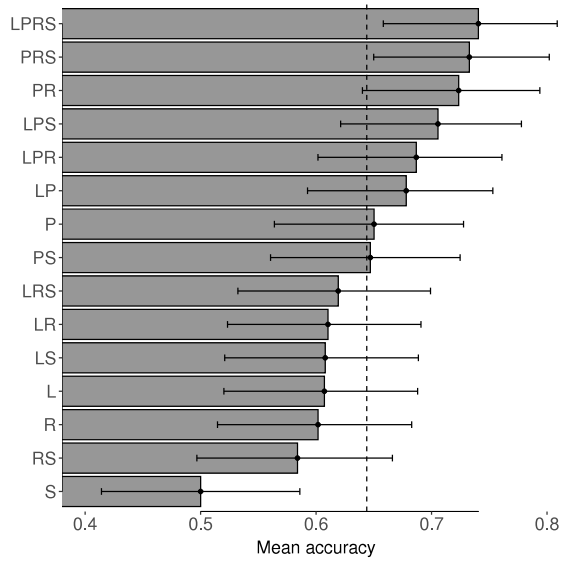

Fig. 5. Estimated marginal mean accuracy of a Recurrent Neural Network predictor of subjective creakiness trained with the computed time-contours of (L)oudness, (P)itch, (R)oughness, (S)harpness, and all possible combinations of these psychoacoustic features. A dashed line indicates the mean accuracy across all combinations of features.

generally accompanied by low F0 values, and F0 is the main acoustic correlate of pitch.

3.4. Automatic prediction of subjective creakiness

Psychoacoustic features could be used to predict the classifications made by the assessors in the same manner that they have been used to predict expert classifications in the past. We trained a Recurrent Neural Network (RNN) using all possible combinations of the four selected psychoacoustic features. This RNN is adequate to capture the time dependencies present in the computed contours which would be difficult to take into consideration otherwise. The scarcity of subjective data prevented us from conducting a fresh tuning of the hyper-parameters.

For this reason, we employed a model very similar to that reported by Villegas et al. (2020). We used a model whose input layer size depended on the number of features used (1 … 4), followed by two bidirectional long short-term memory (BiLSTM) layers with 32 nodes, a fully connected layer followed by a softmax layer, and a classification layer based on the cross-entropy loss. We used the same batch size (10), maximum training epochs (50), learning rate (0.001), and optimizer (Adam), as in (Villegas et al., 2020). In addition, we also evaluated the model using 64 nodes in both BiLSTM layers, as a way to reflect the increment of the number of input features with respect to the previous research.

We used 10-fold cross-validation. Each testing subset comprised a number of creaky words proportional to that of our corpus (48%). Creakiness of a word, in this case, corresponded to a binary classification based on the majority of opinions given by the assessors. For each validation set, we computed the balanced accuracy achieved by the model.

The best mean accuracy (64.39) across all combinations of features and evaluation sets was achieved with the 64-node RNN (cf. 62.28 for the 32-node RNN). A generalized linear model with the accuracy achieved by the RNN as outcome, validation set and combination as predictors indicated significant effects of validation set [$\chi^2(9) = 32.795, p < .001$] and combination [$\chi^2(14) = 34.118, p = .002$].

The top three combinations shown in Fig. 5, i.e., pitch and roughness [z-ratio = $1.741 p = .040$]; pitch, roughness, and sharpness [z-ratio = $1.958, p = .025$]; and the 4-way combination [z-ratio = $2.141, p = .016$]; yielded accuracies higher than the mean accuracy of all combinations (64.39), as indicated by a test performed in the logit scale of the estimated marginal means. A pairwise comparison (with Tukey adjustments) revealed no significant differences in accuracy among combinations except between the top three combinations of Fig. 5 and sharpness: pitch and roughness [z-ratio = $3.631, p = .023$]; pitch, roughness, and sharpness [z-ratio = $3.786, p = .012$]; and the 4-way combination [z-ratio = $3.918, p = .008$].

4. Discussion

Naive and non-naive groups achieved similar accuracy in judging the creakiness of Du'an Zhuang words, the non-naive group being significantly more liberal notwithstanding. The liberal tendency observed in the non-naive group is not justified by the frequency of occurrence of creaky tones in their own language (27% to 37%, if the Hỏi tone is also considered as glottalized). However, different studies (for example, Liu and Wang (2016) and Smith (2011)) have shown that Vietnamese respondents have higher acquiescence than Japanese ones, i.e., regardless of the contents of a question, they tend to select a positive response more frequently than not (Harzing, 2006), in alignment with our findings. This tendency could explain the differences between the accuracy achieved by the two groups judging tones 2, 4 and 5. When words were predominantly creaky, as in tone 4, the non-naive group achieved higher accuracy; on the other hand, when words were predominantly not creaky (tones 2 and 5), the naive group trumped the other. These results indicate that psychoacoustic features may not be the only factors driving the creakiness classifications, but that cognitive factors (culture, for example) also play an important role.

With exception of the classifications of tone 4 words, the non-naive group did not outperform the naive group in our experiments. According to these results, there seems to be little or no benefit of the linguistic background when judging the creakiness of unintelligible words, especially in consideration of the mentioned bias of the non-naive group. Instead, we found evidence in favor of psychoacoustic features better explaining the judgments of the two groups.

The accuracy results showed that both groups performed worst classifying tone 3. Only for this tone, classifications made by the two groups were significantly lower than chance. From the six words with tone 3 included in our survey, five of them were deemed as creaky (four of them unanimously classified). In contrast, the majority of assessors in both groups rated words with this tone as non-creaky, as shown in Fig. 3. The creakiness probability contours presented in Fig. 1 indicate that it increased towards the end of this tone, resulting in the third-largest peak among the unchecked tones. The high sensitivity exhibited by the expert and the ANN classifications is unmatched by the subjective classifications, questioning the perceptual relevance of such extreme sensitivities.

In contrast to the creakiness contours, the loudness and roughness contours (shown in Fig. 1) indicate that tone 3 has one of the lowest values across the vocalic portions of the words. Pitch contours for this tone started the highest among all tones. Recalling that loudness and roughness were found to be positively associated with creakiness judgments, whereas pitch was negatively associated, these findings suggest that listeners were using psychoacoustic features as a way of identifying creakiness in speech. In other words, our results indicate that building models to predict creakiness by means of auditory attributes could yield closer predictions to subjective judgements than when attributes used to describe the unprocessed acoustic signal (the speaker production) are used instead, as in the case of Covarep.

Whereas the roughness contours are drawn from a model that aims to predict the elicited roughness of a given acoustic stimulus (noise, harmonic tone, or whatever it may be), the creakiness contours are produced by a model that has been tailored to predict variations in phonation that characterize such kind of production. The former is an auditory model, the latter a production model. One of the central points of this article is that creakiness prediction models may be too focused on the acoustic signal, not on its representation inside the listening brain. In addition, manual classification of creakiness is aided by the inspection of spectrograms, waveforms, and careful listening of the stimuli an undetermined number of times. Regular listeners do not have these to their avail, and more often than not stimuli are presented in suboptimal conditions (e.g., in the presence of an energetic masker).

When the psychoacoustic contours were used to train an RNN, the accuracy achieved was far from perfect. This is a consequence of the scarcity of subjective judgments, and the use of a binary outcome instead of a continuous one (i.e., the proportion of creaky ratings). The use of a binary outcome was chosen to match the design of the experiments in (Villegas et al., 2020). That way, we could use the same hyper-parameters established in that research, as explained before. The analysis of the results obtained with the RNN model indicates that including pitch and roughness contours as input increases its accuracy. Note that including only these two features produced an accuracy greater than the average of all the combinations trained with the same model and that there was no significant difference in the accuracy achieved with these two predictors, and the accuracy achieved with more predictors.

Sharpness seems to have little or no effect on the subjective creakiness classification. This feature is related to the relative loudness of high frequency bands relative to low frequency ones. For this reason, it is possible that sharpness could be useful to distinguish breathy from normal or creaky phonation. In our corpus, breathy phonation was rare.

The Du'an Zhuang consultant recorded for this experiment produced creaky tones mainly with low and irregular pitch. His creaky phonation is ascribed to arguably the most common kind of creakiness described in the literature (Keating et al., 2015), and it was similar to that produced by Burmese consultants in a previous study (Villegas et al., 2020). Psychoacoustic features could be successfully used to distinguish creaky from non-creaky words, as in the previous study, or explaining subjective judgements, as in the present study. Other kinds of glottalization (fry, tense voice, etc.) could alter the relative weight that each psychoacoustic feature has on both the objective and the subjective predictions. Further studies including other kinds of glottalization could reveal their effectiveness as phonemic contrast and how different psychoacoustic features correlate with them.

5. Limitations

Differences in the apparatus could have influenced the results. Recall that for the Vietnamese cohort, there was no supervision, and we are not certain that all the assessors performed the task in a quiet venue in front of a computer, as requested. It is possible that in a less controlled environment, the Vietnamese cohort could have been distracted, or their classification capabilities hindered by using suboptimal headphones.

In addition, there are several implementations of psychoacoustic models. The loudness, sharpness, and pitch of models used in our analysis have been accepted and standardized, but they are not free from limitations and shortcomings. The roughness model that we used was found to be more accurate than that proposed by Fastl and Zwicker (2006) and International Organization for Standardization (2017), but it has limitations to predict the elicited roughness of signals with a similar magnitude spectrum but different phases (Pressnitzer and McAdams, 1999). It is possible that more sophisticated psychoacoustic models yield more accurate predictions, consequently improving the accuracy of phonation classifiers based on them.

This study shows that psychoacoustic features can predict subjective creakiness accurately, but it falls short in identifying the levels of each of these features, their intra-dynamic behavior, and the inter-feature relationships that are driving the subjective judgments. These important aspects require more data than we can currently afford and hence is deferred to further studies.

6. Conclusions

We found no evidence of a benefit related to the linguistic background of a listener when judging creakiness. The observed difference between naive and non-naive judgements could be explained by other factors such as culture. The auditory judgments made by the two groups are not always well predicted by expert judgments (usually based on visual and auditory inspections) or expert systems (e.g., machine learning algorithms) trained with the latter. The difference in predictions made by listeners (regardless of their linguistic background) and those made by experts highlights the importance of using perceptually valid methods to study phonetic phenomena. Psychoacoustic features, especially roughness and pitch, better predict auditory creakiness judgments and could be used to refine current methods for automatic prediction of phonation so that they become perceptually relevant.

CRediT authorship contribution statement

Julián Villegas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing, Funding acquisition. **Seunghun J. Lee:** Methodology, Investigation, Resources, Data curation, Funding acquisition. **Jeremy Perkins:** Investigation, Resources, Data curation, Review & editing. **Konstantin Markov:** Software, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

American National Standards Institute, 2013. Standard Acoustical & Bioacoustical Terminology Database. ANSI/ASA S1.1 & S3.20.

von Aures, W., 1985. Berechnungsverfahren für den sensorischen wohlklang beliebiger schallsignale (A Model for Calculating the Sensory Euphony of Various Sounds). Acustica 59, 130–141, In German.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Memory Lang. 68, 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Boersma, P., Weenink, D., 2022. Praat: Doing phonetics by computer. Version 6.2.12. Retrieved February 1, 2023. Available from www.praat.org.

Brunelle, M., 2009. Tone perception in Northern and Southern Vietnamese. J. Phonetics 37, 79–96. http://dx.doi.org/10.1016/j.wocn.2008.09.003.

Castro, A., Hansen, B., 2010. Hongshui He Zhuang dialect intelligibility survey. SIL Electronic Survey Reports 25, Translated by D. Huang.

Daniel, P., Weber, R., 1997. Psychoacoustical roughness: Implementation of an optimized model. Acta Acust. United Acust. 83, 113–123.

Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. Covarep - A collaborative voice analysis repository for speech technologies. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing. ICASSP, pp. 960–964. http://dx.doi.org/10.1109/ICASSP.2014.6853739.

Denes, P.B., Pinson, E.N., 1993. The speech chain: The physics and biology of spoken language. In: Science/Communication, second ed. New York, NY.

Drugman, T., Kane, J., Gobl, C., 2014. Data-driven detection and analysis of the patterns of creaky voice. Comput. Speech Lang. 28, 1233–1253. http://dx.doi.org/10.1016/j.csl.2014.03.002.

Fastl, H., Zwicker, E., 2006. Psychoacoustics: Facts and Models, third ed. Springer, Berlin, http://dx.doi.org/10.1007/978-3-540-68888-4.

German Institute for Standardization, 2009. Measurement technique for the simulation of the auditory sensation of sharpness. DIN 45692 (2009).

Hartig, F., 2022. DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. https://CRAN.R-project.org/package=DHARMa. R package version 0.4.5.

Harzing, A.W., 2006. Response styles in cross-national survey research: A 26-country study. Int. J. Cross Cultural Manag. 6, 243–266. http://dx.doi.org/10.1177/1470595806066332.

International Organization for Standardization, 2017. Acoustics–methods for calculating loudness–Part 1: Zwicker method. ISO 532-1:2017.

Ishi, C.T., Sakakibara, K.I., Ishiguro, H., Hagita, N., 2008. A method for automatic detection of vocal fry. IEEE Trans. Audio Speech Lang. Process. 16, 47–56. http://dx.doi.org/10.1109/TASL.2007.910791.

Kane, J., Drugman, T., Gobl, C., 2013. Improved automatic detection of creak. Comput. Speech Lang. 27, 1028–1047. http://dx.doi.org/10.1016/j.csl.2012.11.002.

Kawahara, H., Agiomyrgiannakis, Y., Zen, H., 2016. Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. In: 9th ISCA Wkshp. on Speech Synthesis, Sunnyvale, CA, USA. pp. 239–246. http://dx.doi.org/10.21437/SSW.2016-36.

Keating, P., Esposito, C., Garellek, M., Khan, S., Kuang, J., 2010. Phonation contrasts across languages. In: UCLA Working Papers in Phonetics. pp. 188–202.

Keating, P., Garellek, M., Kreiman, J., 2015. Acoustic properties of different kinds of creaky voice. In: Proc. of the 18 Int. Congress of Phonetic Sciences, 0821. pp. 1–5.

Kirby, J.P., 2011. Vietnamese (Hanoi Vietnamese). J. Int. Phonetic Assoc. 41, 381–392. http://dx.doi.org/10.1017/S0025100311000181.

Klatt, D.H., Klatt, L.C., 1990. Analysis synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87, 820–857. http://dx.doi.org/10.1121/1.398894.

Kong, J., 2001. On Language Phonation. Minzu University Press, (in Chinese).

Kreiman, J., Gerratt, B., Garellek, M., Samlan, R., Zhang, Z., 2014. Toward a unified theory of voice production and perception. Loquens 1, 1–19. http://dx.doi.org/10.3989/loquens.2014.009.

Kuang, J., Keating, P.A., 2012. Glottal articulations of phonation contrasts and their acoustic and perceptual consequences, Vol. 12. UCLA Working Papers in Phonetics, UCLA Phonetics Laboratory, pp. 3–161.

Ladefoged, P., 2016. UCLA phonetics lab data. [Software]. Retrieved February 1, 2023. Available from www.phonetics.ucla.edu.

Lenth, R.V., 2022. Emmeans: Estimated marginal means, aka least-squares means. R package version 7.2.

Li, X., 2011. [Du'an Zhuangyu Xingtai Bianhua Yanjiu](Du'an Zhuang Morphological Change Research). National Publishing House, Beijing, (in Chinese).

LimeSurvey Project Team / C. Schmitz, 2016. LimeSurvey: An open source survey tool. In: LimeSurvey Project. Hamburg, Germany, URL http://www.limesurvey.org.

Liu, M., Wang, Y., 2016. Interviewer gender effect on acquiescent response style in 11 Asian countries and societies. Field Methods 28, 327–344. http://dx.doi.org/10.1177/1525822X15623755.

Loi, N.V., Edmondson, J., 1998. Tones and voice quality in modern northern Vietnamese: Instrumental case studies. Mon-Khmer Stud. 28, 1–18.

Mathworks, 2022. Matlab. Software. Available from www.mathworks.com (February 1, 2023).

Perkins, J., Lee, S., Villegas, J., 2016. An interplay between F0 and phonation in Du'an Zhuang tone. In: Proc. 5 Int. Symp. on Tonal Aspects of Languages, Buffalo. pp. 56–59. http://dx.doi.org/10.21437/TAL.2016-12.

Pham, A.H., 2003. Vietnamese Tone: A New Analysis. Routledge, New York, London.

Pressnitzer, D., McAdams, S., 1999. Two phase effects in roughness perception. J. Acoust. Soc. Am. 105, 2773–2782. http://dx.doi.org/10.1121/1.426894.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Version 4.2.0. Retrieved on February 1, 2023. Available from www.R-project.org.

Schrader, J.E., 2002. A Matlab implementation of a model of auditory roughness (Master's thesis). Eindhoven University of Technology. Eindhoven. Department of technologie management.

Smith, P.B., 2011. Communication styles as dimensions of national culture. J. Cross-Cultural Psychol. 42, 216–233. http://dx.doi.org/10.1177/0022022110396866.

Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8, 185–190. http://dx.doi.org/10.1121/1.1915893.

Villegas, J., Markov, K., Perkins, J., Lee, S.J., 2020. Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness. IEEE J. Sel. Top. Signal Process. 14, 355–366. http://dx.doi.org/10.1109/JSTSP.2019.2949422.

Vo, X.H., 1997. Nghien Cuu Ve Chuc Nang Cua Thanh Dieu Tieng Viet Theo Phuong Phap Dinh Luong (a Statistical Study of the Function of Viemamese Tones) (Ph.D. thesis). University of Education, Hanoi, (in Vietnamese).