INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Optimizing an automatic creaky voice detection method for Australian English-speaking females

Hannah White[1], Joshua Penney[1], Andy Gibson[1], Anita Szakay[1], Felicity Cox[1]

[1]Centre for Language Sciences, Department of Linguistics, Macquarie University, Sydney, Australia

hannah.white2@hdr.mq.edu.au, joshua.penney@mq.edu.au, andy.gibson@mq.edu.au, anita.szakay@mq.edu.au, felicity.cox@mq.edu.au

Abstract

Creaky voice is a nonmodal phonation type that has various linguistic and sociolinguistic functions. Manually annotating creaky voice for phonetic analysis is time-consuming and labor-intensive. In recent years, automatic tools for detecting creaky voice have been proposed, which present the possibility for easier, faster and more consistent creak identification. One of these proposed tools is a Creak Detector algorithm that uses an automatic neural network taking its input from several acoustic cues to identify creaky voice. Previous work has suggested that the creak probability threshold at which this tool determines an instance to be creaky may vary depending on the speaker population. The present study investigates the optimal creak detection threshold for female Australian English speakers.

Results show further support for the practice of first finding the optimal threshold when using the Creak Detection algorithm on new data sets. Additionally, results show that accuracy of creaky voice detection using the Creak Detection algorithm can be significantly improved by excluding non-sonorant data.

Index Terms: creaky voice, creak detection, automatic methods, Australian English

1. Introduction

Creaky voice is a nonmodal phonation type, which is generally perceived to have low pitch and a rough and/or pulse-like quality [1, 2, 3]. Broadly speaking creaky voice is often produced with greater glottal constriction compared to non-creaky voice (e.g., modal voice or other nonmodal phonation types such as breathy voice) [4, 5, 6, 7, 8]. Although this is not true for every instance [6, 9]. There are different realizations of creaky voice characterized by a variety of cues which can be captured by a number of acoustic measures [3, 9, 10]. These include low fundamental frequency (f0), low spectral tilt (indicating high levels of glottal constriction), high noise/aperiodicity in the signal measured by the Harmonic-to-Noise Ratios (HNR) (high HNR indicates irregular f0), or low HNR, indicating the damping of glottal pulses [9].

In some languages such as Jalapa Mazatec and Sgaw Karen [5], creaky phonation functions as a phonologically contrastive voice quality. In English, creaky voice is non-contrastive and has been associated with various prosodic and social functions. For example, creaky voice functions as a marker of phrase-finality [3] and has been suggested to convey authoritativeness, genuineness or, in some cases, boredom [11, 12]. Additionally, the prevalence of creaky voice has been found to vary according to speaker background. There appears to be a general consensus, particularly in social commentary (e.g., [13]), that creaky voice is more common in women's speech than men's speech. Some studies of American English have provided support for this claim [12, 14], although others have found no gender differences [15]. Dallaston and Docherty [16] highlight how production studies have disproportionately focused on women's speech, making it difficult to empirically confirm these claims.

Researchers use a range of approaches in studying creaky voice. Typically, creaky voice is identified manually by human annotators using auditory or spectrographic auditory-visual cues [16]. However, manual annotation of creaky voice is a time-consuming and labor-intensive process that leaves opportunity for human bias and inconsistencies between and within annotators, particularly as we know little about variation in how humans perceive creaky voice [1] (though see [17-21]). As a result, and in response to the appeal of big data, there has been growing interest in automatic identification methods for detecting creaky voice [1, 22-27].

A metric commonly used to assess the performance of automatic creaky voice detection methods is the *F*1 score [22, 24] described in (1) below.

$$F1 = \frac{2 \times \text{True positives}}{2 \times \text{True positives} + \text{False positives} + \text{False negatives}} \quad (1)$$

The *F*1 score is a measure of a test's accuracy and is particularly useful for skewed datasets where the feature in question occurs sparsely. The *F*1 score returns a value between 0 and 1, with higher values indicating more accurate performance.

In [28], two automatic detection methods (AntiMode method [1] and the Creak Detection algorithm [22]) were assessed with respect to a human annotator using speech data from male and female Australian English (AusE) speakers. Each method was compared to the other and to a "union" method, which combined the output of the two methods. The results showed that combining the two methods resulted in significantly more accurate creaky voice detection than using either method alone. It was also found that limiting the analysis to vowel segments resulted in significantly better creaky voice identification. This was likely due to a number of false positives (or incorrect identification of creaky voice) by the automatic methods during voiceless segments and periods of silence.

The Creak Detection algorithm (CD) uses an artificial neural network implemented in Matlab [29] to identify creaky voice based on several acoustic cues [22]. For each 10ms frame of the input signal, CD assigns a probability of the speech in that interval being creaky. This value is converted into a binary creaky voice decision depending on the creak probability

http://dx.doi.org/10.21437/Interspeech.2021-711

threshold to which the algorithm is set (the threshold above which a frame is classified as creaky).

In [22] Drugman *et al.* conducted a threshold sweep to find the optimal threshold for eliciting the most accurate creaky voice detection using CD. The source data was both read and conversational speech from several different corpora. Male and female speakers of American English (m: 3; f: 2), Finnish (m: 1; f: 1), Swedish (m: 1; f: 1) and Japanese (f: 2) were included. The probability threshold was systematically varied between 0 and 1. Results showed that a threshold of 0.3 returned the highest CD performance. The authors also noted that there was little inter-database sensitivity with this threshold setting. They therefore set the default creak probability in CD at 0.3.

Murton *et al.* [24] conducted a similar threshold sweep with CD, but with a more homogeneous corpus, using spontaneous speech from eight American English-speaking females aged 18-22 years. Their results showed an optimal threshold of 0.02, much lower than the default of 0.3. The maximum *F*1 score returned at the 0.02 threshold in [24] was 0.549, while the default 0.3 threshold returned a score of 0.326. As a result, the authors suggest that CD performance might vary depending on the speaker group, a factor that should be considered when using this method on different data sets.

1.1. Aim of the present study

In light of the findings in [24], the aim of this study is to find the optimal creak probability threshold for a corpus of female AusE speakers. An additional aim is to examine whether restricting the analysis to only a subset of the data, e.g., vowels or sonorants rather than all available data, will have an impact on the optimal threshold and improve the outcome.

2. Methods

2.1. Data

The data for this study were extracted from the AusTalk [30] and AusVoices [31] corpora. Nine sentences common to both corpora were used in this analysis. The sentences were all declaratives ranging from 9 to 30 syllables.

The speech of 38 AusE-speaking females from Sydney aged between 18 and 34 at the time of recording were used in this study. Most of the sentences included were speakers' first attempts; however, when mistakes were made, second attempts were included if they were available. Although most speakers had a usable recording of each sentence, there were seven cases when sentences were excluded from analysis due to file corruption (n=4), laughter (n=1), missing words (n=1) or coughing (n=1). At least seven sentences were included for each speaker, except one speaker who had five.

All were processed by the MAUS automatic aligner [32] set to an AusE model, which returned Praat textgrids with marked phoneme boundaries. Phoneme boundaries were hand-corrected by the first author using cues in the waveform and spectrogram. Sentences were manually annotated for creaky voice in Praat [33] by the first author, using visual-auditory cues in the spectrogram such as low f0 and irregularly spaced striations [1, 12]. A subset of sentences (10%) was manually annotated by a second annotator and re-annotated by the first author. The Kohen's Kappa measure was used to calculate inter- and intra-annotator agreement using the irr [34] package in R [35]. The kappa value was 0.834 for inter-annotator agreement and 0.854 for intra-annotator reliability, indicating high levels of agreement for both comparisons. All sentences per speaker were concatenated into a single file before processing through CD (mean length: 50.2 seconds; standard deviation: 12.6 seconds).

2.2. Procedure and analysis

Each file was processed through CD 30 times, adjusting the probability threshold each time between 0.01 and 0.3 in increments of 0.01. The decision to stop the threshold sweep at 0.3 rather than progressing to 1 was made based on the findings in [24] showing that the optimal threshold was considerably smaller than 0.3 at 0.02 and that *F*1 scores continued to decrease past the 0.3 point.

A total of 1140 files (38 speakers x 30 thresholds) were processed. As mentioned, CD returns a binary creak decision for every 10ms frame. These decisions were compared to manual coding and each was assigned one of the following:

- a true positive (TP: manual coder and CD agree creak is present)
- a true negative (TN: manual coder and CD agree creak is not present)
- a false positive (FP: CD identifies creak when manual coder does not)
- a false negative (FN: manual coder identifies creak when CD does not)

Following [22] and [24] we used the *F*1 score to evaluate the performance of CD. In [28], it was shown that *F*1 scores can be significantly improved by restricting the analysis to vowel segments alone. The present study therefore included a comparison of *F*1 scores calculated over all available data versus sonorant-only segments versus vowel-only segments. These will henceforth be referred to as data subsets.

3. Results

*F*1 scores were calculated across all speakers at each threshold from 0.01 to 0.3. Figure 1 illustrates the *F*1 scores for each of the data subsets: all available data, sonorants-only and vowels-only. It shows that while the sonorant-only and vowel-only *F*1 scores pattern very similarly to each other across the creak probability thresholds, the *F*1 scores calculated using all available data are much lower and pattern differently. The optimal creak probability thresholds for each data subset and the *F*1 scores they returned are shown in Table 1. For comparison, the creak probability thresholds and *F*1 scores identified in Murton *et al.* [24] are also included in this table.

Table 1: *Peak F1 scores and optimal creak probability thresholds for each data subset and Murton et al.'s [24] results, which included all available data.*

Data subset	Optimal threshold	Peak *F*1
All	0.18	0.611
Sonorants	0.04	0.767
Vowels	0.03	0.783
Murton *et al.*	0.02	0.549

*F*1 scores were then calculated for each speaker using the output of CD with the optimal threshold settings shown in Table 1. The decision was made to restrict the comparison to the sonorants-only data and all available data for this analysis as, in Figure 1, the difference in *F*1 scores between sonorant-only and

vowel-only data appears to be small, and including non-vocalic sonorant segments increases the number of datapoints that the *F*1 score equation is calculated over for each speaker.

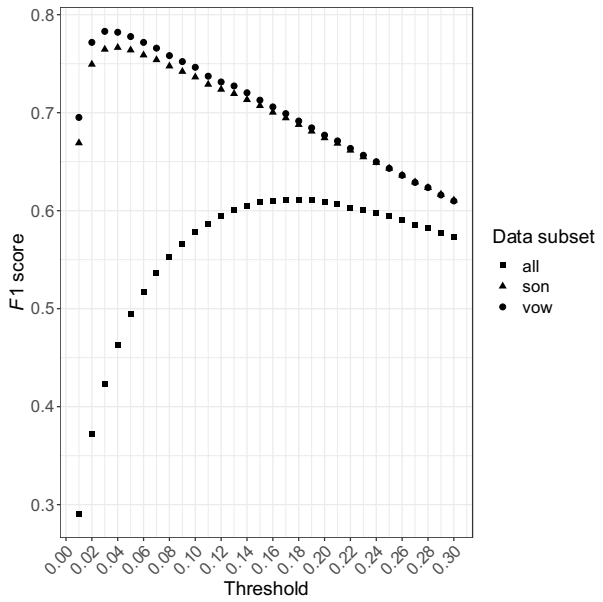

Figure 1: *F1 scores across each creak probability threshold for each data subset. Squares represent all available data, triangles represent sonorant-only subset and circles represent vowel-only subset.*

A comparison of the raw means and standard deviations of the *F*1 scores calculated for each speaker using all available data, sonorant-only data and the results of Murton *et al.* [24] are given in Table 2.

Table 2 shows that while the optimal creak probability thresholds of the present study's full data set and that of [24] (who did not subset their data), are different, they achieved very similar mean *F*1 scores. Mean *F*1 scores in the present analysis using sonorant-only data were substantially higher.

Table 2: *Mean and standard deviation of F1 scores calculated per speaker at the optimal creak probability threshold for all available data and sonorant-only segments from the present study's speech corpus, and Murton et al.'s [24] results, which included all available data.*

Data subset	Optimal threshold	Mean *F*1	SD
All	0.18	0.566	0.14
Sonorants	0.04	0.743	0.084
Murton *et al.*	0.02	0.539	0.084

A linear mixed effects model was run using the lme4 package [36] in R to determine whether speakers' *F*1 scores were significantly affected by data subset. The model included *F*1 score as the dependent variable and the two-level factor data subset (all available data or sonorant-only data) as the independent variable, with speaker as a random intercept. The main effect of data subset was significant ($p<0.0001$) and is illustrated by the bold black points in Figure 2.

Figure 2 shows that *F*1 scores calculated using sonorant-only segments are significantly higher than those calculated using all available data. The raw values in Figure 2, represented by the grey points, show that this was the case for every speaker, albeit to varying degrees.

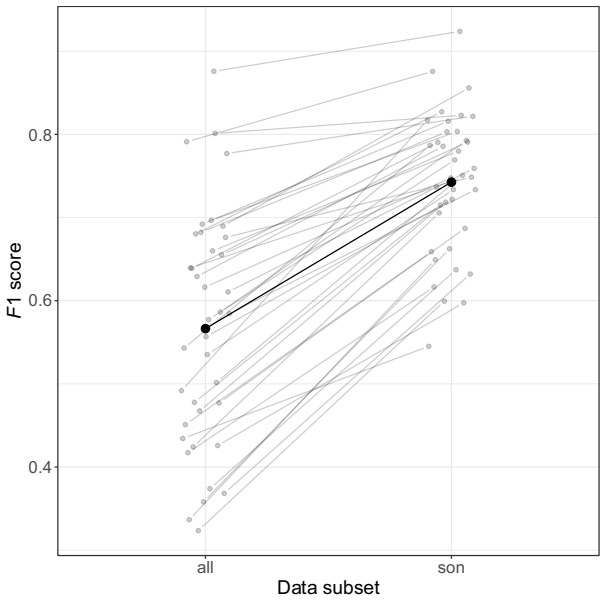

Figure 2: *Predicted F1 scores of the linear mixed effects model for all available data versus sonorant-only data, and raw F1 scores of each speaker. Model predictions are shown by bold black points, and raw values are grey points. Raw F1 scores for each data subset are connected by lines for each speaker.*

4. Discussion

The results of this study support the findings in [24] that the optimal creak probability threshold used in CD may vary depending on the speaker population on which it is being used. Here we have shown that for the female AusE-speakers in this study, when using all available data, the optimal creak probability threshold of 0.18 was between the default setting in the algorithm used in [22] (0.3) and that found in [24] (0.02) for American English-speaking females. The differences in optimal thresholds across the different speaker populations in these studies suggest that there is a need to manually annotate a subset of data and identify the optimal threshold for a particular population before applying CD (and perhaps other automatic detection methods) to different corpora. This is highlighted by the fact that speakers in both the present study and [24] were English-speaking females; however, even in these closely related varieties (AusE and American English), optimal thresholds were found to be different.

Results also show that by limiting the analysis to sonorant-only segments, the *F*1 score can be significantly improved. In this analysis, the optimal creak probability threshold was lower using the sonorant-only segments compared to all available data (0.04 versus 0.18, respectively). As the threshold is lowered, it increases the chance that the algorithm will identify a segment as creak, in turn increasing the risk of identifying false positives (i.e., returning creak results for non-creaky segments).

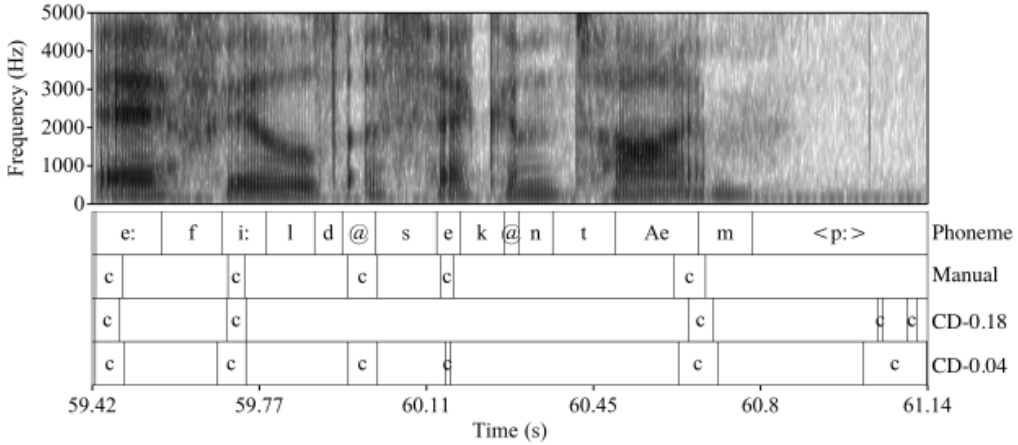

Figure 3 shows a wide-band spectrogram and an aligned phonemic transcription from part of a sentence taken from this study. This figure also shows three levels of creaky voice annotation, based on manual identification (Manual), and based on CD with a threshold level of 0.18 (CD-0.18) and 0.04 (CD-0.04). It is possible to see that the creaky voice segments identified by CD at the 0.04 threshold generally appear to align with the manual coding more accurately than those identified by CD at the 0.18 threshold. Both of these threshold settings identified creaky voice during periods of silence, as can be seen for example at the right edge of Figure 3, with the 0.04 threshold doing so to a greater extent. Using all available data at the 0.18 threshold allowed these false positives to contribute to the *F*1 scores for this data subset. In the sonorant-only analysis at the 0.04 threshold, the false positives shown in the fourth tier of Figure 3 would have been excluded from contributing to the *F*1 score equation. Additionally, the 0.04 threshold captured a slightly wider interval of creaky voice in /i:/ which extended back into the /f/. This false positive (within the fricative segment) would have also been excluded in the sonorant-only analysis. This speaker's *F*1 score increased from 0.492 using all available data at the 0.18 threshold to 0.817 using sonorant-only data at the 0.04 threshold.

In light of these findings, it appears researchers may be able to improve the performance of automatic creaky voice detection methods by first identifying and labelling the individual segments at the phoneme level, then excluding all non-sonorant segments from analysis. It is important that studies of creaky voice are able to accurately reflect the use of creaky voice in different populations. This involves aiming for the most accurate methods of detection possible. While the prospect of manually annotating creaky voice across large corpora may lead researchers to turn to automatic detection methods, it is important that these methods are optimized. This study has shown two ways in which this can be achieved: through performing a threshold sweep on a subset of data, and by limiting analysis to sonorant segments.

5. Conclusion

This study followed [24] in performing a sweep of creak probability thresholds in the CD algorithm [22] to find the threshold that resulted in the most accurate creaky voice detection for a group of AusE-speaking females. Results provided further evidence that CD performance may be maximized at different thresholds depending on the speaker population under examination. Based on the findings of [24] and the present study, using CD to accurately detect creaky voice may require a threshold sweep to be performed in the initial stages of the analysis, which in turn requires a subset of data to be manually annotated.

Results have also shown that limiting the analysis to a sonorant-only subset of data excluded many false positives from being included in the *F*1 score equation, resulting in higher accuracy.

In future work, we plan to add male speech from the same corpora to this analysis to investigate whether speaker sex also impacts the optimal threshold. We are also planning to undertake further automatic method comparisons including methods such as the measure of roughness introduced in [27].

6. Acknowledgements

This research was supported by a Macquarie University Research Excellence Scholarship to the first author and by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the fifth author. We thank Jidde Jacobi for his assistance with MATLAB scripting, Benjamin Purser for annotation support, and the MQ phonetics lab for helpful discussions around this study.

7. References

[1] K. Dallaston and G. Docherty. "Estimating the prevalence of creaky voice: A fundamental frequency-based approach," In Proc. 19th ICPhS, Melbourne, Australia, Aug. 2019, pp. 581.1-5.

[2] H. Hollien, P. Moore, R. W. Wendahl and J. F. Michel. "On the nature of vocal fry," J. Speech and Hearing Research, vol. 9, no. 2, 1966, pp. 245-247.\

[3] L. Redi and S. Shattuck-Hufnagel. “Variation in the realization of glottalization in normal speakers,” J. Phonetics, vol. 29, no. 4, pp. 407-429, Oct. 2001, doi: 10.1006/jpho.2001.0145.

[4] J. Esling, S. Moisik, A. Benner and L. Crevier-Buchman, *Voice quality: The laryngeal articulator model*. Cambridge, UK: Cambridge University Press, pp. 37-82, 2019.

[5] C. M. Esposito and S. Dowla Khan. “The cross-linguistic patterns of phonation types,” Lang. Linguist Compass, vol. 14, pp. e12392, Nov. 2020, doi: 10.1111/lnc3.12392.

[6] M. Garellek. “The phonetics of voice,” in The Routledge Handbook of Phonetics, W. Katz and P. Assmann, Eds., New York, NY, USA: Routledge, 2019.

[7] M. Gordon and P. Ladefoged. “Phonation types: A cross-linguistic overview,” J. Phonetics, vol. 29, no. 4, pp. 383-406, Oct. 2001, doi: 10.1006/jpho.2001.0147.

[8] J. Laver, *The phonetic description of voice quality.* Cambridge, UK: Cambridge University Press, 1980.

[9] P. Keating, M. Garellek and J. Kreiman. “Acoustic properties of different kinds of creaky voice,” In Proc. 18th ICPhS, Glasgow, UK, Aug. 2015, pp. 0821.1-5.

[10] M. Garellek and P. Keating. “Phrase-final creak: Articulation, acoustics, and distribution,” presented at 89th Ann. Meeting of Linguistic Soc. Amer., Portland, OR, USA, Jan. 8-11, 2015.

[11] C. Gobl and A. Ní Chasaide. “The role of voice quality in communicating emotion, mood and attitude,” Speech Communication, vol. 40, no. 1-2, pp. 189-212, Apr. 2003, doi: 10.1016/S0167-6393(02)00082-1.

[12] P. Yuasa. “Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?,” American Speech, vol. 83, no. 3, pp. 315-337, Aug. 2010, doi: 10.1215/00031283-2010-018.

[13] N. Wolf. “Young women, give up the vocal fry and reclaim your strong female voice.” The Guardian. https://www.theguardian.com/commentisfree/2015/jul/24/vocal-fry-strongfemale-voice (accessed Feb. 26, 2021).

[14] R. Podesva. “Gender and the social meaning of non-modal phonation types,” In Proc. 37th Annu. Meeting Berkeley Linguistics Soc., C. Cathcart, I. Chen, G. Finley, S. Kang, C. S. Sandy and E. Stickles, Eds., 2013, pp. 427-448.

[15] N. Abdelli-Beruh, T. Drugman and R. H. Red Owl. “Occurrence frequencies of acoustic patterns of vocal fry in American English speakers,” J. Voice, vol. 30, no. 6, pp. 759.e711-759.e720, Nov. 2016, doi: 10.1016/j.jvoice.2015.09.011.

[16] K. Dallaston and G. Docherty. “The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature,” PLoS ONE, vol. 15, no. 3, e0229960, Mar. 2020, doi: 10.1371/journal.pone.0229960.

[17] L. Davidson. "The effects of pitch, gender, and prosodic context on the identification of creaky voice," Phonetica, vol. 76, no. 4, pp. 235-262, July 2019, doi: 10.1159/000490948.

[18] L. Davidson, "Perceptual coherence of creaky voice qualities," in Proc. 19th ICPhS, Melbourne, Australia, Aug. 2019, pp. 147-151.

[19] M. Garellek and S. Seyfarth. “Acoustic differences between English /t/ glottalization and phrasal creak,” in Proc. INTERSPEECH 2016, San Francisco, CA, USA, Sept. 2016, pp. 1136-1140.

[20] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," Journal of Phonetics, vol. 29, no. 4, pp. 365-381, 2001, doi: 10.1006/jpho.2001.0149.

[21] J. Kreiman, Y. Lee, M. Garellek and R. Samlan. “Validating a psychoacoustic model of voice quality,” J. Acoust. Soc. Am., vol. 149, no. 1, pp. 457-465, Jan. 2021, doi: 10.1121/10.0003331.

[22] T. Drugman, J. Kane and C. Gobl. “Data-driven detection and analysis of the patterns of creaky voice,” Comput. Speech and Lang., vol. 28, no. 5, pp. 1233-1253, Sep. 2014, doi: 10.1016/j.csl.2014.03.002.

[23] P. Martin. "Automatic detection of voice creak," in Speech Prosody 2012, Shanghai, China, May 2012, pp. 43-46.

[24] O. Murton, S. Shattuck-Hufnagel, J. Choi and D. D. Mehta. “Identifying a creak probability threshold for an irregular pitch period detection algorithm,” J. Acoust. Soc. Am., vol. 145, no. 5, EL379, May 2019, doi: 10.1121/1.5100911.

[25] N. Narendra and K. S. Rao. "Automatic detection of creaky voice using epoch parameters," in Proc. INTERSPEECH 2015, Dresden, Germany, Sept. 2015, pp. 2347-2351.

[26] L. Tavi, T. Alumäe and S. Werner. "Recognition of Creaky Voice from Emergency Calls," in Proc. INTERSPEECH 2019, Graz, Austria, Sept. 2019, pp. 1990-1994.

[27] J. Villegas, K. Markov, J. Perkins and S. J. Lee. “Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness,” IEEE J. Sel. Topics in Signal Process., vol. 14, no. 2, pp. 355-366, Feb. 2020, doi: 10.1109/JSTSP.2019.2949422.

[28] H. White, J. Penney, A. Gibson, A. Szakay and F. Cox. “Assessment of automatic creaky voice detection tools for sociolinguistic applications,” presented at the Australian Linguistic Soc. Conf., online, Dec. 14-15, 2020.

[29] MATLAB. (R2020a), The MathWorks Inc. Available: https://au.mathworks.com/products/matlab.html/

[30] D. Burnham et al. “Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box,” In Proc. INTERSPEECH 2011, Florence, Italy, Aug. 2011, pp. 841-844.

[31] F. Cox and S. Palethorpe. “Timing differences in the VC rhyme of Standard Australian English and Lebanese Australian English,” In Proc. 17th ICPhS, Hong Kong, Aug. 2011, pp. 528-531.

[32] F. Schiel, C. Draxler and J. Harrington. “Phonemic segmentation and labelling using the MAUS technique,” Presented at the New tools and methods for very-large-scale phonetics research workshop, Philadelphia, PA, USA, Jan. 29-31, 2011.

[33] P. Boersma and D. Weenink. “Praat: doing phonetics by computer,” (Version 6.1.39, 2021). [Computer program]. Available: http://www.praat.org/

[34] M. Gamer, J. Lemon, I. Fellows and P. Singh. “irr: Various Coefficients of Interrater Reliability and Agreement,” (R package version 0.84.1). Available: https://CRAN.R-project.org/package=irr

[35] R: A Language and Environment for Statistical Computing. (2020), R Core Team, R Found. for Statistical Comput. Available: https://www.R-project.org/

[36] D. Bates, M. Maechler, B. Bolker and S. Walker. “Fitting Linear Mixed-Effects Models using lme4,” J. Statistical Software, vol. 67, no. 1, pp. 1-48, 2015, doi: 10.18637/jss.v067.i01.