ESTIMATING THE PREVALENCE OF CREAKY VOICE: A FUNDAMENTAL FREQUENCY-BASED APPROACH

Katherine Dallaston and Gerard Docherty

Griffith University
katherine.dallaston@griffithuni.edu.au, gerry.docherty@griffith.edu.au

ABSTRACT

Anecdotal claims of increasing prevalence of creaky voice in varieties of English, particularly among younger female speakers, have piqued the interest of sociophonetic researchers, speech pathologists, and public commentators alike. However, studies quantifying creaky voice prevalence are few in number and modest in scale, possibly because manual annotation of creaky voice – the method most often used for its detection – is time-intensive. Since low F0 characterizes most manifestations of creaky voice, it is conceivable that it can be detected, with a high degree of approximation, using an automated F0-based method. This paper describes such an approach, drawing on previous work by Dorreen [7], and explores its application and validity across male and female speakers of Australian English and across speaking tasks. Our findings suggest that our approach is an effective means of estimating creaky voice prevalence, with potential for generating new insights in an area where a reliable evidence base is much-needed.

Keywords: phonation, fundamental frequency, creaky voice, prevalence

1. INTRODUCTION

Creaky voice (‘vocal fry’, ‘glottal fry’, or simply ‘creak’) is a phonation type that auditorily manifests as a low-pitched and impressionistically ‘rough’-sounding voice quality [12]. In public and academic discourse, it is common to encounter anecdotal claims that in some varieties of English the prevalence of creaky voice has recently increased, particularly among younger female speakers, and particularly in the United States [14, 17, 23]. However, quantitative studies on the prevalence of creaky voice in English are scarce and tend to sample small numbers of speakers and only short stretches of speech [5]. Thus, presently, the prevalence of creaky voice in spoken English is a much-discussed but under-investigated topic, and commentary on the phenomenon is commonly rooted in anecdotal rather than empirical evidence.

The benefits of complementing anecdotal observations of creaky voice prevalence patterns with a reliable evidence base are many, and cross-disciplinary. For example, knowledge of creaky voice prevalence patterns across speakers and speaking contexts may provide sociophonetic researchers with new insights into its social and communicative functions [18]. As another example, for speech pathology researchers, measurement of creaky voice prevalence would provide an opportunity to evaluate the often-stated hypothesis that ‘overuse’ of creaky voice presents a risk to vocal health [1, 9, 2].

Implicit to the quantification of creaky voice prevalence is the assumption that it is appropriate to categorically classify phonation as either creaky voice or not creaky voice. This kind of categorical delineation of phonation types is grounded in long-established phonetic theory [13] and is widely-practised in speech science [8], but we do acknowledge that not all research on creaky voice takes this approach. For example, some studies have investigated speakers’ ‘creakiness’ using continuous acoustic measures [e.g. 20] or qualitative perceptual scales [e.g. 19]. Our interpretation of these studies is that their approaches are no less valid than ours, but do not measure prevalence as we intend it here – i.e. the percentage of phonation realised as creaky voice as opposed to some other phonation type.

When the methodological realities of creaky voice prevalence research are considered, the limitations of previous research are not altogether surprising. Creaky voice is typically defined according to auditory criteria (as we have done in this introduction), and so detection of creaky voice is usually achieved through human auditory perception and annotation [5]. This makes analysis of long speech recordings across a large sample of speakers time- and labour- intensive. It also makes duplication of findings and cross-study comparisons difficult due to issues of intra- and inter-rater reliability. Though some automated methods of detecting creaky voice have been proposed [10, 11], none are yet well-enough established that they are used routinely in quantitative creaky voice prevalence research [5].

Since most manifestations of creaky voice are characterised by low fundamental frequency (F0) [12], it is conceivable that phonation can be classified as being either creaky voice or not creaky voice, with a high degree of approximation, using an automated F0-based method.

There are two important considerations underpinning the success of such an F0-based approach. The first is sufficiently accurate detection of F0. This has previously been difficult to achieve because many widely-used pitch trackers such as Praat [3] do not reliably track F0 during intervals of very low frequency (i.e. those characteristic of creaky voice)[7]. REAPER [21] is a relatively new pitch tracker that has been shown to detect glottal closure instants (GCIs) – from which F0 can be calculated – with a high degree of reliability, even during intervals of creaky voice [7, 15]. Once GCI time points are known, the duration of each glottal cycle can be calculated, and then, each cycle's F0. The added usefulness of REAPER's GCI analysis is that it can be used to quantify the total phonation duration (calculated as the summed duration of all glottal cycles), which is needed to calculate what *percentage* of phonation realised as creaky voice, or in other words, to calculate creak voice prevalence.

The second condition upon which an F0-based method relies is the selection of an appropriate F0 value to delineate low F0 glottal cycles – i.e. those which are likely to be creaky voice – from those with higher F0. However, selecting an appropriate threshold is not simple as reports of creaky voice F0 ranges vary across studies [2] and may differ across speakers depending on speaker sex [16]. If we were to semi-arbitrarily select an F0 value to delineate creaky from non-creaky glottal cycles and apply that threshold uniformly across different speakers, we risk the detection of creaky voice being more accurate for some speakers than others. A *speaker-specific* F0 threshold therefore seems most appropriate. Recent work by Dorreen [7] has shown potential for a speaker's F0 *antimode* to be used as a threshold for delineating creaky voice from modal (or non-creaky) phonation. As Dorreen [7] explains, when speakers produce both modal and creaky phonation, F0 distributions are typically bimodal (one peak in the modal F0 range, and one in the creak F0 range). The antimode is the F0 value that occurs with the lowest frequency between these two modes.

In this paper, drawing on previous work by Dorreen [7], we describe an F0-based method for detecting occurrences of creaky voice to estimate its prevalence, and investigate the effectiveness of this across male and female speakers of Australian English and across speaking tasks.

2. METHODS

2.1. Speakers and speech material

Audio recordings were obtained from the AusTalk corpus [4], an Australian corpus of high-quality (44.1 kHz sample rate, 16-bit resolution) recordings made between 2011 and 2016 of approximately 900 speakers from across the country, all of whom received the entirety of their primary and secondary schooling in Australia (and thus deemed to be speakers of Australian English). We sampled speakers using the following criteria: aged 18-50 years old, living in the Perth area (i.e. recorded at AusTalk's Perth site), who reported no speech or hearing problems and no voice-impacting health problems. Recordings of two speech tasks were obtained for analysis: (1) *Read Story*, in which the participants read aloud an Australianised version of 'Arthur the Rat'; and (2) *Retold Story*, in which participants were asked to retell that same story in their own words. After excluding speakers who did not have intact recordings of both speech tasks (n=16), the final sample consisted of 42 speakers; 28 females (aged 19–47, M=28), and 14 males (aged 19–45, M=25). From this point onwards we refer to speakers as their AusTalk participant codes.

2.2. Analysis

2.2.1. Preparing the audio files for analysis

A total of 84 .wav files (42 speakers, 2 tasks) were manually edited to remove sections incongruent with the task description, e.g. dialogic speech before, after, or during the task, speech from the data collector, overlapping speech, laughter, background noise, and long portions of silence. After editing, the *Read Story* recordings were on average 203 seconds in duration, ranging 159–267 seconds, and *Retold Story* recordings were on average 47 seconds in duration, ranging 16–159 seconds. Each .wav file was divided into 10-second segments, with the final segment being 10 seconds plus the remainder.

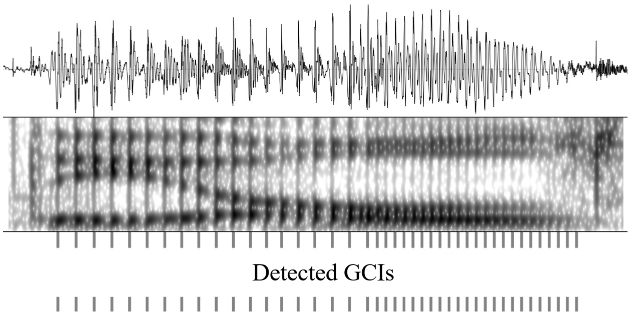
2.2.2. Detecting GCIs to calculate F0

The segmented .wav files were then processed through REAPER to obtain the GCI analyses, which are contained in REAPER's 'pitch mark' output. We used REAPER's default settings, but reduced the 'minimum F0 to look for' from 40 to 20 Hz because our priority was accurate detection of low F0 [15]. Once retrieved, the GCI analyses of segmented .wav files were merged, resulting in 84 pitch mark files corresponding to 84 .wav files. To catch any procedural errors at this stage, GCI time points were written to Praat textgrids for visual inspection alongside their respective .wav files (Fig. 1). Then, for each glottal cycle we calculated its duration (the inverse of the time between one GCI and the next), and F0 value (its duration / 1).

2.2.3. Calculating the F0 Antimode

We analysed the 84 F0 distributions using the R package 'modes' [6] to locate, in this order, the following: (1) the mode likely to be the *modal*

Figure 1: Speaker *3_926* saying “the old”, showing REAPER’s detection of Glottal Closure Instants (GCIs) of low and high F0 glottal cycles.



phonation mode (the most frequent F0 value or global maximum); (2) the mode likely to be the *creaky phonation mode* (the tallest local maximum with an F0 value below the global maximum), and; (3) the *antimode* (the least frequent F0 value, or smallest local minimum, between the two modes). In some cases, this three-step procedure returned an antimode that we deemed to be a ‘false’ antimode due to the creaky phonation mode being detected as a local maximum on the left side of the modal phonation distribution curve. After some experimentation, we decided to include in the automated process a condition that if the antimode selected had a density of >0.005 (i.e. was not a ‘convincingly low valley’), the creaky phonation mode was recalculated as the *next* tallest local maximum with an F0 value below the global maximum, and this was repeated until an antimode with sufficiently low-density was found. All antimodes were inspected visually, as in Fig. 2, to ensure this automated process had identified plausible antimodes for speaker and task.

2.2.4. Evaluating the effectiveness

To measure the effectiveness of the method, a textgrid file was created for each .wav file and intervals of glottal cycles with F0 values below the antimode were annotated as *+Creak*; all other sections were annotated as *–Creak*. Then, one of the authors listened to a random sample of *+Creak* and *–Creak*, accruing to a total of 15% of all *+Creak* and *–Creak* predictions (84.41 and 675.8 seconds, respectively), and noted how much of each interval was correctly predicted. Accuracy was judged according to the criterion that creaky voice is ‘a rough quality with the additional sensation of repeating impulses’ [10], and decision making was augmented by checking the spectrogram for widely- and/or irregularly spaced vertical striations. We made no conceptual distinction between suprasegmental creaky voice and phonemic glottalisation (e.g. of voiceless plosives or vowel onsets/hiatuses); any speech in which creaky-sounding phonation was perceivable was classified auditorily as creaky voice.

3. RESULTS

Antimodes were detected in 80 of the 84 speech recordings. For two speakers (*2_330* and *3_1212*), an antimode was detected in only one task, and for one (*4_767*), in neither task. Due to limited space, here we report results relating only to the 39 speakers for whom an antimode was detected in both tasks.

3.1. Antimodes across speakers and tasks

Antimodes values were considerably variable across speakers (Fig. 2), particularly between males and females (Fig. 3), yet relatively stable within-speakers. The difference between speakers’ antimodes across tasks ranged 1.38–13.41 Hz for males (M=8.53, SD=3.78), and 2.16–28.17 Hz for females (M=10.28, SD=6.65). The amount of phonation that speakers produced in their ‘intra-antimode window’ – the F0 range between their two antimodes – was small, ranging 0.5–2.23% of total phonation in each task for males (M=0.83, SD=0.63), and 0–5.3% for females (M=0.64, SD=0.94).

Figure 2: F0 distributions, antimodes (•), and estimated prevalence of creaky voice (%< •) for 39 speakers across two tasks.

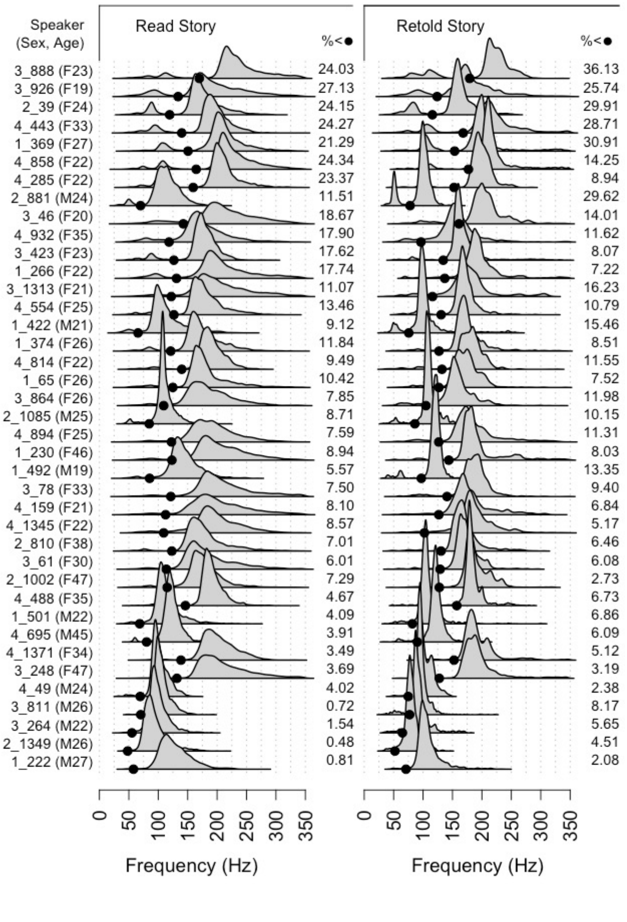
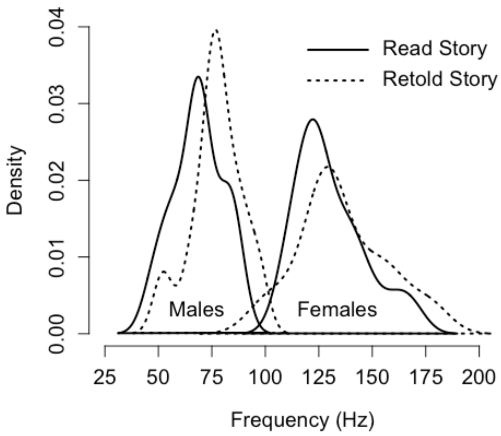


Figure 3: Distribution of antimode values for 39 speakers, grouped by speaker sex and task.



3.2. Accuracy of the prediction

Auditory analysis of randomly sampled +*Creak* (intervals of glottal cycles with F0s below the speaker's antimode) found 81.38% of +*Creak* was accurately predicted (81.22% accurate for males, as a group, and 81.43% accurate for females). When sections of .wav files were incorrectly predicted to be +*Creak*, these were largely in the context of glottalized realisation of plosives or glottal stops, in which one or two low F0 glottal cycles occurred but were so momentary that no auditory impression of creaky voice was perceivable. Additionally, some sections of the .wav file were incorrectly +*Creak* due to inaccurate GCI detection during, for e.g., air puffs into the microphone.

Auditory analysis of randomly sampled –*Creak* (sections of the .wav files *not* comprised of glottal cycles with F0s below the speaker's antimode) found 97.72% of –*Creak* was accurately predicted (96.89% accurate for males, as a group, and 98.27% accurate for females). Correctly predicted intervals of –*Creak* consisted of phonation but of another type (modal phonation, typically), voiceless consonants, or silence between words. When sections of the .wav file were predicted to be –*Creak* but were in fact creaky voice, according to auditory analysis, this tended to be a result of REAPER's occasional failure to detect GCIs of very low F0 and very low intensity. Additionally, some intervals of creaky voice were not detected because the interval was characterised by *irregular* rather than low F0.

4. DISCUSSION

Overall, this automated F0-based approach appears to be effective at achieving *coarse-grained* estimates of prevalence across and within speakers. Our auditory analysis indicated that this method divided audio files into sections likely and not likely to be creaky voice with a high level of accuracy. That there was similar accuracy across male and female speakers, despite the antimodes of female speakers being considerably higher than male speakers, supports the use of a speaker's F0 antimode as a speaker-specific criteria for the automated detection of creaky voice. We see this method having many potential applications, including in the selection of stimuli for experimental creaky voice perception studies. Because it is automated, it is efficient enough to be used with large numbers of speakers and long stretches of speech, potentially generating new insights in an area where a reliable evidence base is much-needed.

Our finding that individual speakers' antimodes were similar but not identical across tasks, despite tasks being different in style and duration, echoes Dorreen's [7] previous finding that bilingual speakers' F0 antimodes are similar but not identical across languages. When antimodes were not identical across tasks, the glottal cycles with F0 values between the speaker's two antimodes were classified as +*Creak* in one speaking task, but not in the other. However, the percentage of phonation produced in speakers' 'intra-antimode window' was low for all speakers. This suggests that this method may be used to estimate within-speaker variability in creaky voice prevalence, even when there are cross-task differences in antimode values. Small variability in antimodes may indicate that, for some speakers, there is a *range* of F0 values between their modal and creaky phonation distributions in which they tend not to (and possibly cannot) produce phonation.

Importantly, the automated classification of +*Creak* and –*Creak* did not always agree with our auditory analysis. This may be in part because not *all* manifestations of creaky-sounding phonation are characterised by low F0 [12]; a future refinement could be to incorporate a measure of F0 irregularity. Additionally, the method's effectiveness could be further assessed, such as by rendering the human rater blind to the automated prediction during auditory analysis, and/or by examining the variability of its accuracy across individual speakers.

5. REFERENCES

[1] Behrman, A., Akhund, A. 2017. The effect of loud voice and clear speech on the use of vocal fry in women. *Folia Phoniatr. Logop.* 68, 159–166.

[2] Blomgren, M., Chen, Y., Ng, M. L., Gilbert, H. R. 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *J. Acoust. Soc. Am.* 103(5 Pt 1), 2649–2658.

[3] Boersma, P., Weenink, D. 2018. *Praat: doing phonetics by computer* [Computer program]

[4] Burnham, D., Estival, D., Fazio, S., Viethen, J. Cox, J., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow. M., Lewis, T., Butcher, A., Hajek, J. 2011. Building an audio-visual corpus of Australian English: large corpus

collection with an economical portable and replicable Black Box. *Proc. Interspeech* Florence, Italy, 841–844.

[5] Dallaston, K. In Preparation. The prevalence of creaky voice in varieties of spoken English: A systematic review.

[6] Deevi, S. and 4D Strategies. 2016. *modes: Find the Modes and Assess the Modality of Complex and Mixture Distributions, Especially with Big Datasets.* R package version 0.7.0. http://CRAN.R-project.org/package=modes

[7] Dorreen, K. 2017. *Fundamental frequency distributions of bilingual speakers in forensic speaker comparison.* Master's thesis, The University of Canterbury, Christchurch, New Zealand.

[8] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29(4), 383–406.

[9] Gottliebson, R. O., Lee, L., Weinrich, B., Sanders, J. 2007. Voice problems of future speech-language pathologists. *Journal of Voice* 21(6), 699–704.

[10] Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N. 2008. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 47–56.

[11] Kane, J., Drugman, T., Gobl, C. 2013. Improved automatic detection of creak. *Computer Speech & Language* 27(4), 1028–1047.

[12] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th ICPhS* Glasgow, Scotland.

[13] Laver, J. 1980. *The phonetic description of voice quality*. New York: Cambridge University Press.

[14] Lawson, R. 2016. A different drum: Social media and the communication of sociolinguistic research. In: Lawson, R., Sayers, D. (eds), *Sociolinguistic research: application and impact.* Routledge.

[15] Liberman, M. 2015. *REAPER*. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=17590

[16] Melvin, S. 2015. *Gender variation in creaky voice and fundamental frequency.* Honor's thesis, The Ohio State University, Ohio, USA.

[17] Mendoza-Denton, N. 2011. The semiotic hitchhiker's guide to creaky voice: circulation and gendered hardcore in a chicana/o gang persona. *Journal of Linguistic Anthropology* 21(2), 261–280.

[18] Podesva, R. J., Callier, P. 2015. Voice quality and identity. *Annual Review of Applied Linguistics* 35, 173–194.

[19] Stuart-Smith, J. 1999. Glasgow: Accent and voice quality. In: Foulkes, P., Docherty G. J. (eds), *Urban voices.* London: Arnold, 203–222.

[20] Szakay, A. 2012. Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. *Journal of Sociolinguistics* 16(3), 382–397.

[21] Talkin, D. 2015. *REAPER: Robust Epoch And Pitch EstimatoR*. Retrieved from https://github.com/google/REAPER

[22] Wolk, L., Abdelli-Beruh, N. B., Slavin, D. 2012. Habitual use of vocal fry in young adult female speakers. *Journal of Voice* 26(3), e111–e116.

[23] Yuasa, I. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech: A Quarterly of Linguistic Usage,* 85(3), 315–337.