Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

# Creak in the phonetic space of low tones in Beijing Mandarin, Cantonese, and White Hmong

*Seoyoung Kim*[1,*], *Claudia Matachana*[1,*], *Alex Nyman*[1,*], *Kristine M. Yu*[1]

[1]University of Massachusetts Amherst, USA

seoyoungkimk@umass.edu, cmatachanalo@umass.edu, anyman@umass.edu, krisyu@linguist.umass.edu

## Abstract

Low pitch, irregular pitch, and constricted voicing have been proposed as three independent perceptual properties of creaky voice quality, with corresponding acoustic correlates fundamental frequency, harmonics-to-noise ratio, and spectral tilt measure H1-H2. We examined how these three acoustic measures described the variability in a small corpus of multispeaker productions of low falling tones that are often creaky in Beijing Mandarin, Cantonese, and White Hmong. Using principal components analysis, we found that harmonics-to-noise ratios strongly dominated the first principal component (50-60% of the variance across languages), while fundamental frequency and H1-H2 were strongly correlated. Moreover, in all three languages, tokens identified as likely to be creaky by a neural network creak classifier (Drugman et al. 2014) clustered in the high noise region of the principal component space according to the first principal component. No systematic patterns of clustering with respect to fundamental frequency or spectral tilt were found. Principal component analysis on only tokens identified as having greater than a 50% likelihood of being creaky indicated a lack of statistical independence between the three acoustic measures across languages and no distinct clusters were found in the principal component space in any language.

**Index Terms**: voice quality, creak, tone, phonation

## 1. Introduction

Creaky voice quality is not a monolithic, unitary perceptual category; rather, there are distinct sub-types [1]. Most recently, [2, 3] proposed low pitch, irregular pitch, and constricted voicing to be three independent perceptual properties of creaky voice quality, with corresponding acoustic correlates of low fundamental frequency (f0), low harmonics-to-noise ratios (high noise), and low spectral tilt measure H1-H2 (difference in amplitude between first and second harmonic), and in addition, high subharmonics-to-harmonics ratio—a special kind of noise measure, an acoustic correlate of multiply pulsed creak. Here, we mean independent in the sense that having a single one of these properties alone has been proposed to be sufficient for inducing a creaky voice quality percept (sub-type). In this phonetic creak space, [2, 3] have also identified distinct kinds of creak, e.g., vocal fry (low pitch, constricted voicing, but not irregular pitch), unconstricted creaky voice (irregular pitch, low pitch, but not constricted). Given these identified sub-types, we might expect creaky exemplars to cluster in different regions over the proposed 3D acoustic space, rather than for exemplars to be uniformly distributed over the space. If the contrast in identified sub-types is linguistically important, e.g., [3] gives the example of one sub-type being associated with prosodic phrasing while another is associated with segmental identity, then sufficient acoustic separation between the two sub-types could be important for perceptual contrast. If one of the three perceptual properties is sufficient for a creaky percept, we might also expect the three corresponding acoustic dimensions to show statistical independence, e.g., naively, without consideration to correlations due to articulatory configurations, if low f0 is sufficient for a speaker to induce a creaky percept, then there is no need to also have high noise/irregularity.

In this paper, we examined how the acoustic measures underlying the proposed phonetic creak space (harmonics-to-noise ratios, H1-H2, and f0, as well as subharmonics-to-harmonics ratio) characterize the variability in a small corpus of multispeaker productions of low falling tones that are often creaky in Beijing Mandarin, Cantonese, and White Hmong. Table 1 shows the tonal inventories of the three languages, with tones ordered by level tones, rising tones, and then falling tones; low falling T21 (˩, boxed) occurs in all three inventories and has been shown to be often produced with creak in all three languages (Beijing Mandarin: [4], Cantonese: [5], White Hmong: [6], see also refs. within). In all three languages, some perceptual work has suggested that creak in these T21 low falling tones may be concomitant with the production of low f0 such that creak independent from low f0 does not necessarily bias listeners towards identifying T21 (Mandarin: [4, 7], White Hmong: [8], Cantonese: [9]); [4] also proposed that creaky voice quality could serve as a way to enhance low pitch cues.

Table 1: *Tonal inventories of languages*

| Language | Tonal inventory |
|---|---|
| Beijing Mandarin | ˥,˧,˨, ˩ |
| Cantonese | ˥,˦,˧,˨,˩, ˩ |
| White Hmong | ˥,˦,˧,˨,˩, ˩ |

Given these results, we might expect f0 and the other acoustic properties associated with creak to be strongly correlated in productions of T21 in these languages. For instance, we might expect a token with lower f0 to also have lower H1-H2, and higher noise. Moreover, we might expect there to be a gradient of increased creakiness in terms of these acoustic measures in the phonetic space over all T21 tokens—including ones that might be perceived/categorized as non-creaky. However, given the proposed perceptual independence of the three primary creak space dimensions, we might also expect, a priori, that they could be statistically independent, i.e., orthogonal to one another. We used principal components analysis (PCA) as an exploratory method to investigate the statistical indepen-

* These authors contributed equally to the work.



10.21437/SpeechProsody.2020-107

dence of and interactions between f0, H1-H2, and noise measures in characterizing the acoustics of T21 tokens in the three languages. PCA determines statistically independent dimensions that account for the most variability in the data over the input parameters. It also does not require any distributional assumptions; distributional assumptions of normality are needed only for inferential statistics with PCA. We also used the automatic creak detector of [10] as a starting point to observe how the likelihood of a token being creaky related to its position in the phonetic space. We performed PCA again on the subsets of T21 tokens detected to be more likely to be creaky than not.

# 2. Materials and Methods

### 2.1. Tonal production corpus

Data came from a subset of recordings described in [11, Ch. 1] that included Tone 21 as a sentence-medial target tone. Tone 21 was produced as either the first or second syllable of a bitone uttered as part of a sentential carrier phrase, where the bitone varied over all possible tone combinations in the language. In the recordings, the bitone was flanked by mid level tones in Cantonese and Hmong and by low and high tones in Mandarin (following [12]), and speakers produced five fluent repetitions of each tonal sequence combination. Segmental material for the target syllables was chosen to be sonorant with low vowel quality when possible to facilitate f0 and voice quality measure extraction (Mandarin: [ma], Cantonese: [lau], Hmong: [la]). For this paper, the data consisted of the following number of tokens: Beijing Mandarin (4 men, 4 women, 909 tokens), Cantonese (6 men, 6 women, 2095 tokens) and White Hmong (4 men, 2 women, 613 tokens). Vowels were manually segmented in Praat [13] in preparation for voice quality measurements.

### 2.2. Voice quality measurements

VoiceSauce [14] was used to automatically extract the acoustic measures proposed to be relevant for characterizing creak: fundamental frequency, noise measures, and spectral tilt measure H1-H2. In particular, we computed f0 using the STRAIGHT algorithm [15], formant frequencies and bandwidths using default settings for Praat, H1-H2, which was corrected for the effect of formants [16], the harmonics to noise ratios in bandwidths from 0-500 Hz (HNR05), 0-1500 Hz (HNR15), 0-2500 Hz (HNR25), and 0-3500 Hz (HNR35) [17], and the Subharmonic to Harmonic Ratio (SHR), that calculates the amplitude ratio between subharmonics and harmonics [18]. We configured the f0 limits specifically for each speaker, since with creaky voice irregularities in f0 are common. We set a floor f0 value of 40 Hz for all of the speakers since creaky regions can have low f0 values, while f0 ceiling values for each speaker were determined by manually estimating the highest f0 values uttered in high tones and adding an additional 15 Hz buffer.

### 2.3. Creak assessment

Following [19] and [20], we used the MATLAB creak detector [10] `detect_creaky_voice.m` available at the covarep Github repository to automatically estimate the likelihood that a token was creaky, with default settings. [19] found an 81.3% agreement rate between the detector and hand-classified sample of Mandarin. The algorithm uses acoustic features, including H1-H2, energy- and f0-based measures, as well as acoustic periodicity/irregularity features proposed in [21], in a neural network with a single hidden layer. It outputs a creak probability estimate with a 10ms frameshift. We estimated the overall likelihood of creak of a token as the mean creak probability estimated over all frames within the vowel. Up to this point we have not yet hand-validated the automated creak detection, but plan to in future work.

### 2.4. Statistical analysis

R [22] was used to conduct PCA with the mean extracted values over the vowel of the parameters chosen during VoiceSauce analysis. Each voice quality parameter mean was scaled within each speaker as z-scores to standardize the scales over all parameters for interpretation of PCA. We used `dplyr` [23] and `ggplot2` [24] packages for data extraction and plotting, and `factoextra` [25] for PCA visualization.

# 3. Results and Discussion

### 3.1. PCA space for all T21 tones

The first three PCs accounted for approximately 90% of the total variance across the T21 tokens in each language: for P1-P3, respectively: Cantonese: 58.7%, 18.2%, and 14.7%; Hmong: 55.6%, 19.5%, 13.7%, Mandarin: 61.1%, 18.0%, 11.4%. The variable correlation plots in Figures 1-3 show the correlation between the voice quality variables and PCs 1 and 2. The arrowhead points to the direction of positive correlation between a variable and a principal component axis; a small angle between two variable vectors means that they are strongly correlated, while variable vectors orthogonal to one another are uncorrelated, and a small angle between a variable vector and a PC axis means that that variable strongly influences that PC. The further away a variable vector is from the origin, i.e., the higher the squared cosine (cos2) value and redder the color of the vector, the more important it is for the first two components (see [26]). If the noise (HNRs), spectral tilt (H1-H2), and f0 dimensions each perfectly mapped onto separate PCs, we'd expect to see one of the corresponding vectors along the x-axis, another along the y-axis, and the last with a distance very close to the origin (and thus not well-described by PC1 or PC2). This is not what we find.

The variable correlation plots show that, across all three languages, all HNRs were highly positively correlated with one another and dominated PC1. Moreover, H1-H2 and f0 were most important for PC2 in Hmong and Mandarin and their corresponding vectors were not well-correlated with HNRs—in fact, they're orthogonal to HNRs in Mandarin. H1-H2 was also strongly positively correlated with f0 in Hmong and Mandarin. However, in Cantonese, f0 was positively well-correlated with HNRs, as indicated by the relatively small angle between them, and both f0 and H1-H2 did not strongly dominate PC2, but contributed to all first three PCs.

In sum, according to the exploratory PCA analysis over all T21 tokens (including many non-creaky ones), the statistical independence of the acoustic correlates of the three proposed creaky voice quality dimensions varied across the three languages. PC1 defined the noise dimension (as indexed by HNRs) and was independent from the spectral tilt (H1-H2) and f0 dimensions across all three languages; however, spectral tilt and f0 were strongly, positively correlated and distributed across PCs. The strong positive correlation between spectral tilt and f0 is consistent with [27]'s perceptual results of the integration of the two in pitch perception and the idea of low spectral tilt serving as an enhancement for low f0. Interestingly, SHR was almost orthogonal and thus largely uncorrelated with the HNR

noise measures, although SHR can also be considered a noise measure. This indicates that both types of noise measures may be useful in describing the phonetic space of often-creaky low tones.

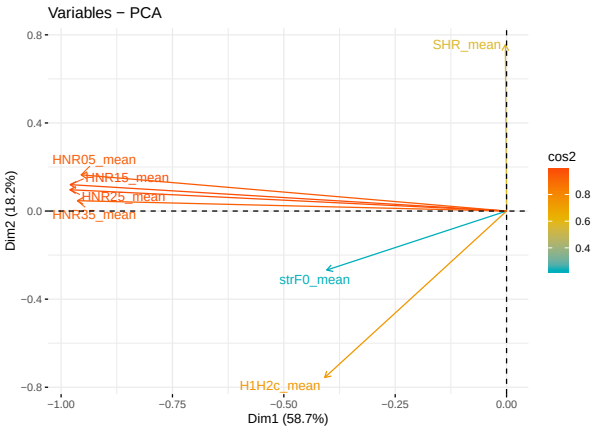


Figure 1: *Variable correlation plot:* $(PC1, PC2)$*, Cantonese*

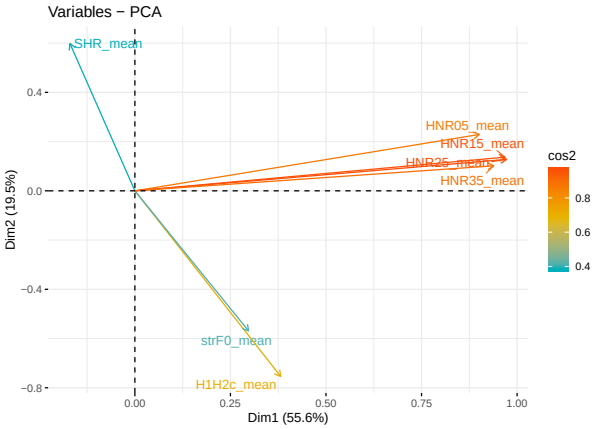


Figure 2: *Variable correlation plot:* $(PC1, PC2)$*, Hmong*

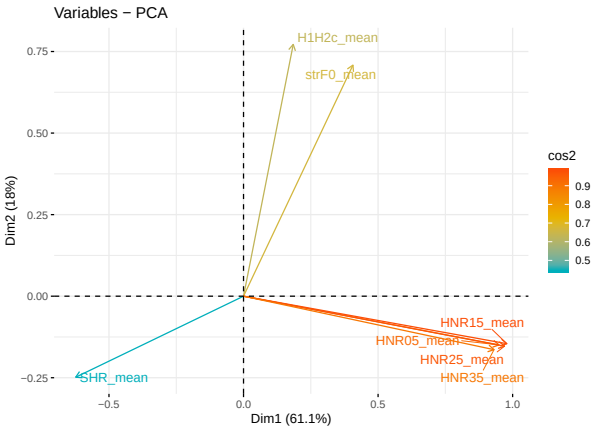


Figure 3: *Variable correlation plot:* $(PC1, PC2)$*, Mandarin*

The biplots in Figures 4, 5, and 6 illustrate the distribution of individual T21 tokens over the acoustic space defined by $(PC1, PC2)$ for each language, together with the projection of the voice quality variables onto the first two principal components shown previously. Each token is colored by its estimated likelihood of being creaky, with the tokens most likely to be creaky colored on the red side of the color spectrum. All three plots show that T21 tokens estimated to have a high likelihood of creak (the reddest dots) cluster in the same region: where HNRs are low (high noise). We could draw a line aligned to the PC1 axis dominated by HNR as an excellent decision boundary between the creaky and non-creaky tokens. The clustering of creaky tokens does not appear to be well-explained by any other variable than HNR in any language, as evidenced by the sprawl of creaky tokens across the y-axis. Thus, at least according to the creak detector of [10], creakier T21 tokens in the White Hmong, Cantonese, and Beijing Mandarin samples analyzed are distinguished from non-creaky T21 tokens primarily by high noise.

In [10], the automatic creak detector was validated against manual determination for its performance on qualitatively different kinds of creak: Pattern A: "highly irregular temporal characteristics" (not necessarily low f0, but high noise, some multiple pulsing), Pattern B: "fairly regular temporal characteristics with strong secondary excitation peaks" (low f0, constricted, not necessarily high noise, but with "double-beats"), and Pattern C: "fairly regular temporal characteristics without strong secondary excitation peaks" (low f0, constricted, not necessarily high noise), i.e., vocal fry. The automatic creak detector did the worst on detecting Pattern C. Therefore, it is possible that the detector was biased towards identifying high noise tokens as most likely to be creaky, accounting in part for the clustering of creaky tokens observed.

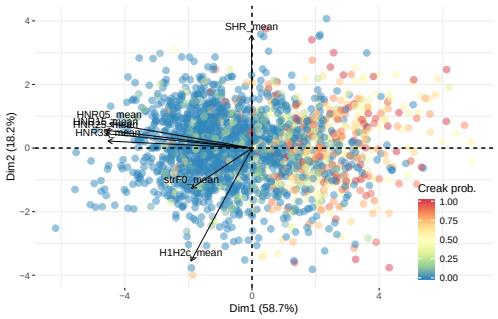


Figure 4: *Biplot of Cantonese T21 tokens over PCs 1, 2*

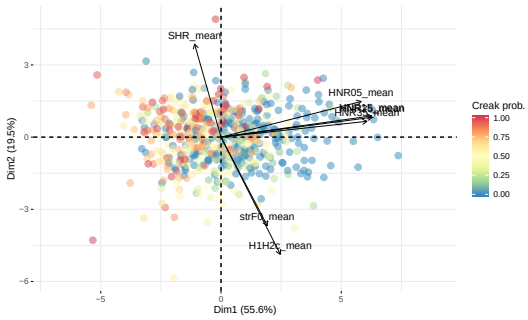


Figure 5: *Biplot of Hmong T21 tokens over PCs 1, 2*

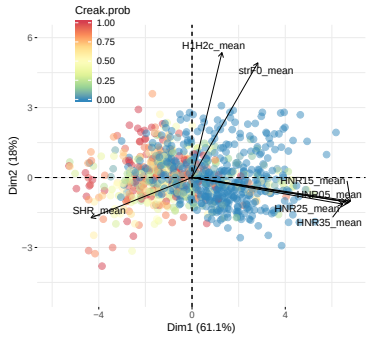


Figure 6: *Biplot of Mandarin T21 tokens over PCs 1, 2*

### 3.2. PCA space for T21 tones with an estimated >50% likelihood of being creaky

The same exploratory analysis was conducted for the subset of T21 tones estimated as having a >50% likelihood of being creaky in each language: 385 tokens (18% of total T21 tokens) for Cantonese, 258 tokens (42%) for Hmong, and 270 (30%) for Beijing Mandarin. The first three PCs accounted for 83-91% of the total variance over T21 tokens: for PC1-PC3, respectively, Cantonese: 49.5%, 26.4% and 16.1%, Hmong: 41.9%, 29.3%, 11.8%, Mandarin: 46.6%, 25.7%, 16.3%. Unlike the correlation plots for all T21 tokens, the variable correlation plots for the creaky subset (Figures 7-9) show much more mixing of HNRs, f0, and H1-H2 within PCs. In particular, HNRs, f0, and H1-H2 all contribute comparably to PC1 and PC2 in Cantonese, rather than PC1 being dominated by HNRs. In Mandarin, SHR contributed most to PC1, and H1H2 to PC2, while f0 and HNRs were spread across multiple PCs. Hmong showed the least change from the space over all T21 tokens: HNRs still dominated PC1 and f0 and H1-H2 were still strongly correlated and dominated PC2. Additionally, when the subsets of creaky T21 tokens were plotted over the first three PCs, no clusters of points emerged, e.g., a cluster of tokens with acoustic properties expected for multiple pulsing and/or a cluster with properties expected of vocal fry, etc.

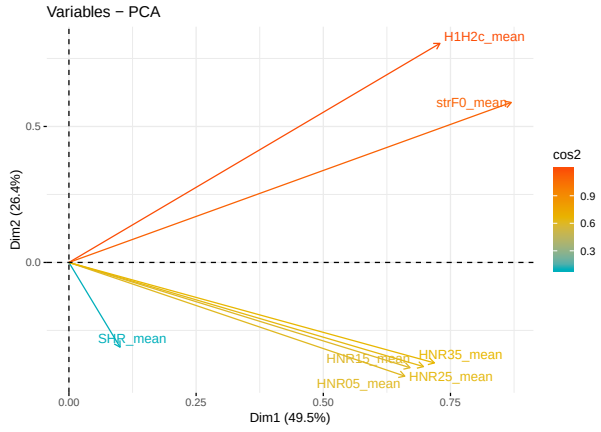


Figure 7: *Variable correlation plot: Cantonese creaky T21s*

## 4. Conclusions

Within the three language samples of low falling tokens, we did not observe evidence of different kinds of creak types, e.g., vocal fry vs. tense vs. prototypical creak vs. multiply pulsed

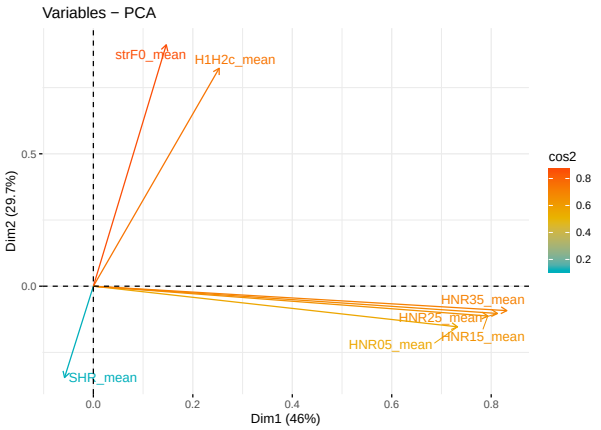


Figure 8: *Variable correlation plot: Hmong creaky T21s*

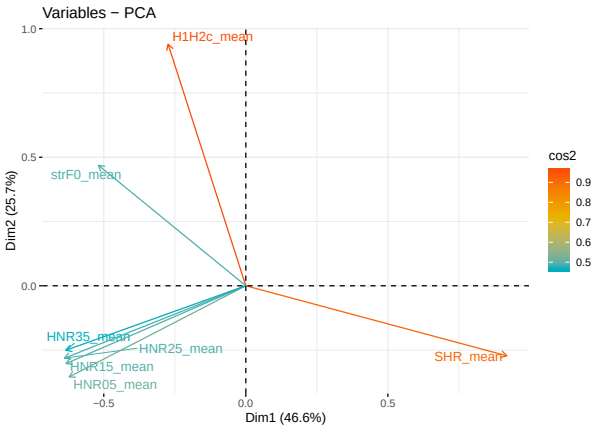


Figure 9: *Variable correlation plot: Mandarin creaky T21s*

creak in the distribution of tokens over PCA space or in the definition of principal components over all tokens over within the detected creaky subset. Rather, harmonics-to-noise ratios as a class were statistically independent from f0 and spectral tilt and accounted for the most variability in the data, and tokens were quite evenly distributed over PCA space rather than clustered. This indicates that although creaky voice quality may have subtypes that are categorically distinct from one another, productions of low falling tokens that are often creaky can fall on a gradient over the phonetic space. Notably, the likely creaky tokens, among all T21 tokens, were not confined in low f0/low spectral tilt regions, which we might expect if low spectral tilt and low f0 work together to enhance a low pitch percept. The importance of HNRs in both characterizing the variability over the low falling tokens and in identifying tokens likely to be creaky is a welcome result, as HNRs are less sensitive than many other voice quality measures to the difficulties of estimating f0, which is often ill-defined in creak. However, these results come with the caveat that they are of course dependent on the definition of the creak detector used. Thus, while the exploratory strategies employed here can be generalized regardless of the chosen method of creak detection, the results cannot. To avoid circularity, the results here must be supplemented with further validation both by human detection of creakiness and additional automatic creak detectors such as [28].

# 5. References


[1] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, no. 4, pp. 365–381, Oct. 2001. [Online]. Available: http://www.sciencedirect.com/science/article/B6WKT-457CJ0P-J/2/a40000ac94fd4ba475653abd2ec493db

[2] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of ICPhS 2015*, 2015.

[3] M. Garellek, "The phonetics of voice," in *The Routledge Handbook of Phonetics*, W. Katz and P. Assmann, Eds. Routledge, To appear.

[4] J. Kuang, "Covariation between voice quality and pitch: revisiting the case of Mandarin creaky voice," *Journal of the Acoustical Society of America*, vol. 142, pp. 1693–1706, 2017.

[5] K. M. Yu and H. W. Lam, "The role of creaky voice in Cantonese tonal perception," *Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1320–1333, 2014.

[6] C. M. Esposito, "An acoustic and electroglottographic study of white hmong tone and phonation," *Journal of Phonetics*, vol. 40, no. 3, pp. 466–476, May 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0095447012000174

[7] Y. Huang, "Low f0 as a creak attribute in Mandarin tone perception," 2019, conference presentation.

[8] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1078–1089, 2013. [Online]. Available: http://link.aip.org/link/?JAS/133/1078/1

[9] Y. Zhang and J. Kirby, "Weighting of f0 mean, f0 change and phonation cues in tone perception: The case of Cantonese tone 4/ tone 6," 2018, manuscript.

[10] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, pp. 1233–1253, 2014.

[11] K. M. Yu, "The learnability of lexical tones from the speech signal," Ph.D. dissertation, University of California Los Angeles, Los Angeles, CA, 2011.

[12] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, pp. 61–83, 1997.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.32) [computer program]." 2010, `http://www.praat.org`.

[14] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: a program for voice analysis," *Proceedings of ICPhS XVI*, 2011.

[15] H. Kawahara, "STRAIGHT-TEMPO: a universal tool to manipulate linguistic and para-linguistic speech information," in *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on*, vol. 2, 1997, pp. 1620–1625 vol.2.

[16] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proceedings of ICASSP*, Montreal, Canada, May 2004, pp. 669–672.

[17] G. de Krom, "A cepstrum-based technique for determing a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language & Hearing Research*, vol. 36, pp. 254–266, 1993.

[18] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 1, 2002, pp. I–333–I–336 vol.1.

[19] J. Kuang, "The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin," *Journal of the Acoustical Society of America*, vol. 143, no. EL509, 2018.

[20] Y. Chai, "The source of creak in Mandarin utterances," in *Proceedings of ICPhS 2019*, 2019.

[21] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47–56, 2008.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.

[23] H. Wickham, R. François, L. Henry, and K. Müller, *dplyr: A Grammar of Data Manipulation*, r package version 0.8.3 ed., 2019.

[24] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer, 2009.

[25] A. Kassambara and F. Mundt, *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, r package version 1.0.5 ed., 2017.

[26] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101

[27] J. Kuang and M. Liberman, "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," *Frontiers in Psychology*, vol. 9, p. 2147, 2018.

[28] J. Villegas, K. P. Markov, J. Perkins, and S. J. Lee, "Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2019.