CONTRIBUTIONS OF ACOUSTIC MEASURES TO THE CLASSIFICATION OF LARYNGEAL VOICE QUALITY IN CONTINUOUS ENGLISH SPEECH

Chenzi Xu[1], Paul Foulkes[1], Philip Harrison[1], Vincent Hughes[1], Jessica Wormald[1]

[1]Department of Language and Linguistic Science, University of York, UK
{chenzi.xu|paul.foulkes|philip.harrison|vincent.hughes|jessica.wormald}@york.ac.uk

ABSTRACT

Laryngeal voice qualities (e.g. breathy and creaky voice), variable within and across speakers, often pose a challenge in data collection. Their acoustic correlates are still inadequately understood. This study revisits the acoustics of laryngeal voice qualities in high-quality recordings of continuous British English speech produced by experienced phoneticians. Through principal component analysis and multinomial logistic regression with l_1 regularisation, this study identifies contributions of a variety of acoustic measures to the classification of laryngeal voice qualities and provides a multidimensional acoustic profile for breathy, creaky, and modal voice. Classification rates as high as 90% were achieved using the first 5 principal components. The most salient acoustic correlates for creaky voice are, compared to other categories, higher mean H2*, lower mean f0 and HNR below 500 Hz, and for breathy voice, higher mean H1* and spectral tilt measures such as H1*–A1* and H1*–H2*.

Keywords: Voice quality, phonation types, acoustics

1. INTRODUCTION

Variation in laryngeal voice quality, such as breathiness and creakiness, is abundant in speech. Voice quality may be used contrastively as in Jalapa Mazatec [1], enhance perceptual distinctiveness of other phonetic categories such as lexical tones as in White Hmong [2], or function as prosodic configurations such as phrase-final creak in various languages [3]. There is also evidence of sociolinguistic differences in voice quality (e.g. [4]), as well as speaker-specific variation which may be relevant to forensic speech science [5, 6].

Laryngeal voice quality has received increasing attention from the perspectives of articulation, acoustics, and perception in the past two decades. Phoneticians have been dedicated to searching for acoustic features that are capable of distinguishing modal, creaky and breathy voice (e.g. [7]). The two basic articulatory dimensions to laryngeal voice quality are the degree of constriction of vocal folds and aspiration noise [8], which can be captured by spectral tilt and harmonic to noise ratio (HNR) measures respectively as proposed in the psychoacoustic model described in [9]. For spectral tilt there are various proposed measures, of which the most common one is H1–H2 or the corrected H1*–H2*[10], the difference in amplitude between the first and second harmonics, amongst other higher-frequency slopes in source spectrum such as H2–H4, H4–H2kHz, and H2kHz–H5kHz. An alternative set of measures of spectral tilt include H1–A1, H1–A2, H1–A3 varying in their harmonic bandwidth depending on the formant frequencies, or in other words, vowel quality. These may also correlate with the source spectral tilt measures. In addition to spectral tilt and HNR measures, other proposed acoustic correlates of laryngeal voice quality in the literature include corrected amplitude of individual harmonics such as H1, H2, and H4, cepstral peak prominence (CPP) [11], f0, formant frequencies and bandwidths [12].

Although it has been found that both the source and filter characteristics provide important cues for phonation contrasts [9, 13, 14], the influence of the filter on laryngeal voice quality is often ignored. The correlations among the selected acoustic measures are also often neglected, which impacts the interpretability of statistical models.

This paper revisits the acoustics of laryngeal voice quality via an interpretable classification algorithm whereby combinations of acoustic features are used to predict laryngeal voice quality. We examine the contributions of a wide range of the proposed acoustic measures (including features associated with the filter) in distinguishing laryngeal voice quality. While many studies measured voice quality from sustained vowels (e.g. [15]), this study utilises continuous speech, which better represents the dynamic attributes of voice in speech.

2. DATA

This paper reports on a subset of the available materials from a bespoke corpus, collected as part of an on-going project, containing high quality recordings of experienced phoneticians in various vocal conditions.

2.1. Participants

Data from four male adult speakers were included in this study (P1-P4). They are all experienced professional phoneticians.

2.2. Procedure

Each participant read the first two paragraphs of *The Rainbow Passage* in twenty-four vocal conditions, varying laryngeal and supralaryngeal voice quality settings, as well as accent guises and other vocal changes (e.g. holding a pen between their teeth). Each vocal condition was repeated three times non-consecutively within a recording session. All participants took part in three sessions which were at least a week apart. Thus, each vocal condition has 3*3 = 9 repetitions, and the recording of each repetition was saved as a PCM WAV file. The repeated design better captures intra-speaker variation. This paper reports only on the modal, breathy, and creaky voice conditions.

2.2. Recordings

The recording sessions were conducted in an anechoic chamber at the University of York. Each participant wore a DPA omnidirectional headset microphone on the right side of their face. The recording was made on a single channel at a sampling rate of 48 kHz and 24-bit quantization using a Zoom F8n recorder.

3. METHOD

This study employs multinomial logistic regression to predict the phonation types with a wide range of voice source and vocal tract features. The preprocessing, data exploration and analysis were implemented in MATLAB, R, and Python, and the scripts are available as supplemental materials[1].

3.1. Acoustic Measurements

Using VoiceSauce [16] in MATLAB, 29 acoustic features (see Table 1) were extracted every 25ms frame with 10ms frame shift for all audio files. All spectral magnitude measures were corrected using f0 and formants. Both f0 and formants were estimated using the Snack algorithm [17]. The f0 extraction range set from 40 Hz to 300 Hz, and four formants were tracked using the default setting with LPC order being 12 and pre-emphasis set to 0.96.

Only measurements from voiced portions (f0>0) were included. Visual inspection of f0 tracking with spectrogram and waveform revealed that the f0 of each speaker did not exceed 230 Hz. Thus, measures of frames whose f0 was larger than 230Hz were considered tracking artefacts and were excluded.

Each audio file is on average 29.04s in duration (SD = 3.77s) and evenly divided into six breath-group-size [18] intervals of about 4.84s each. The mean duration of voicing portion of the intervals is 2.4s (SD = 0.75s). For the analysis, the mean and standard deviation of measures of each acoustic feature over each interval were used. The division of intervals enriched the dataset, instead of using means and standard deviations for acoustic measures calculated across the entire sample of speech.

There are, hence, 162 sets (9 repetitions of 3 voice qualities with 6 intervals per repetition) of measures for each speaker, and each set contains 58 measures (means and standard deviations of 29 features).

Measure	Explanation
Spectral slope measures (dB)	
H1*, H2*, H4*, H2k*	Amplitude of the first, second, fourth harmonic, and the harmonic nearest 2,000 Hz
H1*–H2*, H2*–H4*	Difference in amplitude between the first and second harmonics, second and fourth harmonics
H4*–H2k*	Difference in amplitude between the fourth harmonic and the harmonic nearest 2,000 Hz
H2k*–H5k*	Difference in amplitude between the harmonic nearest 2,000 Hz and the harmonic nearest 5,000 Hz
A1*, A2*, A3*	Amplitude of the first, second, and third formant
H1*–A1*, H1*–A2*, H1*–A3*	Difference in amplitude between the first harmonic and the harmonic closest to the first, second, and third formant
Energy/Amplitude related measures	
HNR05, HNR15, HNR25, HNR35	Harmonic-to-Noise Ratio for 0-500 Hz, 0-1500 Hz, 0-2500 Hz, and 0-3500 Hz
CPP	Cepstral Peak Prominence
Energy	Root Mean Square (RMS) energy
Vocal tract related measures	
F1, F2, F3, F4	First through fourth formant
B1, B2, B3, B4	First through fourth bandwidth
Glottal measures	
f0	Fundamental frequency (Hz)

Table 1: Acoustic measures extracted from VoiceSauce.

3.2. Multinomial logistic regression

To learn the relationship between phonation types and acoustic features, we fitted multinomial logistic regression models with l_1 regularisation, using the

LogisticRegression() function and SAGA solver [19] in Scikit-Learn Python library [20]. The l_1 regularisation shrinks non-significant coefficients to exactly 0 through penalising the absolute values of magnitude of coefficients during optimization, thereby reducing unnecessary acoustic measures in the final model. The predicted variable was one-hot encoded such that $\{$*breathy* = 0, *creaky* = 1, *modal* = 2$\}$. All acoustic features were z-score normalised, since features measured at different scales might not contribute equally to model fitting and create a bias.

3.3. Cross-validation

Cross-validation is employed to evaluate model performance, given the small-scale dataset. This prevents model overfitting by partitioning a subset of data to validate the model prediction accuracy and utilise all the data both for training and for testing.

In our implementation, two sets of cross-validation were built. In set S, data from one **speaker** was singled out as a test set while the remaining data was used for training, and this process was repeated for every speaker. In set R, data from one **repetition** for all four speakers was held as a test set while the remaining eight repetitions were used for training, and this process was repeated for every repetition.

Two metrics were used to measure the classifier performance:(1) accuracy, the percentage of correctly predicted phonation types in the test data, and (2) F1-score, whether the model predictions are balanced across three predicted categories. An F1-score close to 1 generally means better classification.The average results of all iterations in cross-validation are reported.

3.4. Dimensionality reduction

Regression models with all 58 standardised acoustic features (base) as predictors were initially built. To simulate the chance level predictions, baseline models were also set up by randomly sampling predictions from a discrete uniform distribution of [0,2] and comparing them to true labels. The classification results are presented in Table 2.

	Base models		Random baseline	
Average	Set R	Set S	Set R	Set S
Accuracy	0.98	0.79	0.32	0.34
F1 score	0.98	0.78	0.25	0.27

Table 2: Classification results of full models and random baseline models.

The base models greatly outperformed the random predictions, indicating that there are statistical regularities between phonation types and these acoustic features. Set R has higher accuracy than Set S, suggesting that interspeaker variation is larger than intraspeaker variation in laryngeal voice quality.

The base models, however, are not fully interpretable, as many of the features are highly correlated. Hence, we applied principal component analysis (PCA) to the acoustic features prior to logistic regression models, thereby removing multicollinearity between predictors, as PCA transforms a set of correlated variables into a smaller number of orthogonal variables. We built a pipeline chaining PCA and logistic regression to search for optimal parameters of PCA.

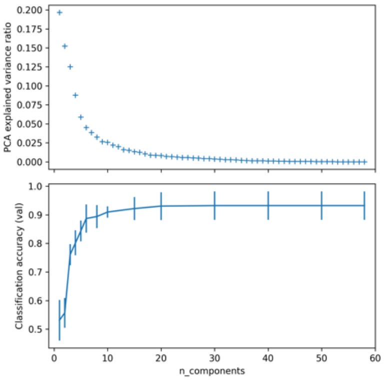

Figure 1: The explained variance and mean classification accuracy across principal components (set R).

We found that with the first five principal components, the classification accuracy consistently reaches over 80%, both for the R and S sets. In Figure 1, the inclusion of the third component greatly increased the accuracy from below 60 % to close to 80%. Hence, we transformed the data with the first five principal components and then used them as predictors in logistic regression (PCA models).

4. RESULTS

4.1. Classification

The PCA models achieved good overall classification results as shown in Table 3, although the first five components captured only about 60% variability in the acoustic data. The confusion matrix[2] from the classification indicates that breathy voice was distinguished from creaky voice with 100% accuracy (and vice versa). Here, we will focus on Set S models tested on unseen speakers.

	PCA models	
	Set R	Set S
Avg. Accuracy	0.9	0.78
Avg. F1 score	0.9	0.78
Avg. Explained variance	0.62	0.63

Table 3: Classification results of PCA models.

The regression model coefficients for each principal component (PC) are listed in Table 4. PCs with larger absolute coefficients indicate their larger contribution to prediction. We can see that all five PCs contribute to the classification of laryngeal voice quality, with PC1 and PC3 generally the most important variables.

Set S	PC1	PC2	PC3	PC4	PC5
Breathy	-0.96	-0.58	1.43	0	0.90
Creaky	2.52	0	-2.86	0.93	-1.82
Modal	0	0.29	0	-0.13	0

Table 4: Model coefficients for one of the models in Set S. Insignificant coefficients were shrunk to 0 by l_1 regularization.

In order to better understand how principal components predict each phonation type, we visualise the test data in the space of fitted PCs. Figure 2 shows the distribution of test set data (speaker P4; Professor Francis Nolan) in a two-dimensional space of PCs derived from the training data. The inspection of PCA transformed test data in Set S suggests that the first PC is crucial to separate creaky voice (green dots) from the other categories, while PC3 distinguishes breathy voice (orange dots) from the others.

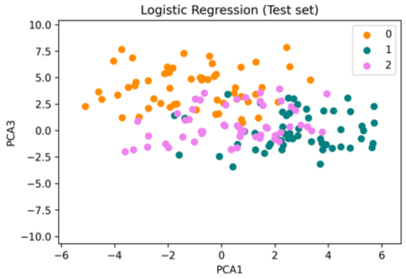

Figure 2: Two-dimensional projection of the test set data using PCs from training set. 0, 1, 2 represents breathy, creaky, and modal voice respectively.

4.2. Acoustic measures contributing to the PCA

The results of section 4.1 can be made interpretable by examining the factor loadings, i.e. the correlation between PCs and acoustic features. Table 5 presents the key factor loadings for the first and third PCs from the data of the first three speakers, used as training data for classification of voice qualities for the test speaker P4 (Figure 2). Full factor loadings can be accessed in the supplemental materials[3].

In Table 5, PC1 negatively correlates with standard deviations of amplitude of harmonics, f0 and HNR mean measures, especially HNR below 500 Hz, and positively correlates with the mean measures of H2*, H4*–H2k*. The first component mainly captures f0 and amplitude or energy related measures, both in inharmonic and harmonic source, distinguishing creaky voice from other categories. PC3 correlates positively with spectral tilt means across all frequency bands H1*–A3*, H1*–A2*, H1*–A1*, H1*–H2*, the mean amplitude of H1*, and the mean and standard deviation measures of first formant bandwidth, and negatively with HNR mean measures.

In Figure 2, creaky voice tends to have higher PC1 values, indicating, as might be expected, lower mean f0 and HNR measures, especially HNR below 500Hz, compared to others. That breathy voice tends to have higher PC3 suggests that breathy voice tends to have higher mean H1*, spectral tilt mean measures such as H1*–A1* and H1*–H2*, mean/SD of first formant bandwidth, and lower mean HNR measures, especially HNR below 2500 and 3500Hz.

Set S	PC1	PC3
Positive	H2* (0.14), H4*–H2K* (0.15) (mean)	H1*, H1*–H2*, H1*–A1*, H1*–A3*, B1 (mean); B1 (sd)
Negative	f0, HNR05 (mean); H1*, H4*, A1*, A3*, H2*, A2*, H2K* (sd)	HNR35, HNR25, HNR15 (mean)

Table 5: Key acoustic measures for PCs (Training set: P1-P3; selected: absolute loadings >0.2 unless specified).

5. DISCUSSION

Laryngeal voice quality is manifested in multiple acoustic features, both source and filter measures. PCA reveals underlying patterns in the wide range of acoustic measures and provides a multidimensional description of the acoustics of phonation types. For instance, PC1 in the sample training set correlates positively with mean H2* and negatively with mean HNR <500 Hz and f0, assembling features potentially profiling creaky voice, while PC3 correlates positively with mean H1*, spectral tilt measures and B1, and negatively with HNR measures, which related to characterising breathy voice. The orthogonal nature of PCs might also provide insights into person-specific patterns, when test data of different speakers was visualised on the PC space learnt from the same reference set.

6. CONCLUSION

The study revisits the acoustics of laryngeal voice quality via a classification approach using PCA and logistic regression. Laryngeal voice quality was well differentiated using the selected acoustic measures. The average classification accuracy was 78% (Set S) and 90% (Set R); well above chance. Consistent with [13], lower mean HNR measures characterise breathy and creaky voice and our findings further suggests HNR measures at different frequency bands play a role in differentiating breathy voice from creaky voice. Other salient acoustic features include spectral measures H1*, H2*, H1*–H2*, and f0, B1.

7. REFERENCES

[1] M. Garellek and P. Keating, 'The acoustic consequences of phonation and tone interactions in Jalapa Mazatec', *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 185–205, Aug. 2011, doi: 10.1017/S0025100311000193.

[2] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, 'Voice quality and tone identification in White Hmong', *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1078–1089, Feb. 2013, doi: 10.1121/1.4773259.

[3] M. Garellek, 'Perception of glottalization and phrase-final creak', *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 822–831, Feb. 2015, doi: 10.1121/1.4906155.

[4] J. Stuart-Smith, 'Glasgow: Accent and voice quality', in *Foulkes & Docherty 1999*, 1999, pp. 201–222.

[5] E. S. Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, 'The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals', *J. Int. Phon. Assoc.*, vol. 49, no. 3, pp. 353–380, Dec. 2019, doi: 10.1017/S0025100318000130.

[6] N. Francis, 'Forensic Speaker Identification and the Phonetic Description of Voice Quality', in *A figure of speech: A Festschrift for John Laver*, W. J. Hardcastle and J. M. Beck, Eds., New York: Routledge, 2005, pp. 385–411. doi: 10.4324/9781410611888-28.

[7] S. Seyfarth and M. Garellek, 'Plosive voicing acoustics and voice quality in Yerevan Armenian', *J. Phon.*, vol. 71, pp. 425–450, Nov. 2018, doi: 10.1016/j.wocn.2018.09.001.

[8] M. Garellek, 'The phonetics of voice 1', in *The Routledge Handbook of Phonetics*, W. F. Katz and P. F. Assmann, Eds., 1st ed.Abingdon, Oxon ; New York, NY : Routledge, 2019. | Series: Routledge handbooks in linguistics: Routledge, 2019, pp. 75–106. doi: 10.4324/9780429056253-5.

[9] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, 'Toward a unified theory of voice production and perception', *Loquens Span. J. Speech Sci.*, vol. 1, no. 1, p. e009, Jan. 2014, doi: 10.3989/loquens.2014.009.

[10] Y. Chai and M. Garellek, 'On H1–H2 as an acoustic measure of linguistic phonation type', *J. Acoust. Soc. Am.*, vol. 152, no. 3, pp. 1856–1870, Sep. 2022, doi: 10.1121/10.0014175.

[11] J. Hillenbrand and R. A. Houde, 'Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech', *J. Speech Lang. Hear. Res.*, vol. 39, no. 2, pp. 311–321, Apr. 1996, doi: 10.1044/jshr.3902.311.

[12] M. Gordon and P. Ladefoged, 'Phonation types: a cross-linguistic overview', *J. Phon.*, vol. 29, no. 4, pp. 383–406, Oct. 2001, doi: 10.1006/jpho.2001.0147.

[13] M. Garellek, 'Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality', *J. Phon.*, vol. 94, p. 101155, Sep. 2022, doi: 10.1016/j.wocn.2022.101155.

[14] I. R. Titze, 'Nonlinear source–filter coupling in phonation: Theory', *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, May 2008, doi: 10.1121/1.2832337.

[15] B. R. Gerratt, J. Kreiman, and M. Garellek, 'Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech', *J. Speech Lang. Hear. Res.*, vol. 59, no. 5, pp. 994–1001, Oct. 2016, doi: 10.1044/2016_JSLHR-S-15-0307.

[16] Y.-L. Shue, P. Keating, and C. Vicenik, 'VOICESAUCE: A program for voice analysis.', *J. Acoust. Soc. Am.*, vol. 126, no. 4, p. 2221, 2009, doi: 10.1121/1.3248865.

[17] K. Sjölander, 'The Snack Sound Toolkit'. KTH Stockholm, Sweden, 2004. Accessed: Dec. 30, 2022. [Online]. Available: https://www.speech.kth.se/snack/

[18] Y.-T. Wang, J. R. Green, I. S. B. Nip, R. D. Kent, and J. F. Kent, 'Breath Group Analysis for Reading and Spontaneous Speech in Healthy Adults', *Folia Phoniatr. Logop.*, vol. 62, no. 6, pp. 297–302, Oct. 2010, doi: 10.1159/000316976.

[19] A. Defazio, F. Bach, and S. Lacoste-Julien, 'SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives', *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[20] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[1] The supplemental materials are available here: https://github.com/uoy-research/pasr-output/tree/main/icphs_23_voicequality.

[2] Confusion matrices of models are presented in the supplemental materials.

[3] Section 4.2 mainly reported on one of the Set S models with the highest test score to demonstrate how to interpret the principal components. Due to limited space, results of other models can be accessible in the supplemental materials.