Analysing breathy voice in forensic speaker comparison
Using acoustics to confirm perception

Katharina Klug[1], Christin Kirchhübel[2], Paul Foulkes[1] and Peter French[1,3]

[1] Department of Language and Linguistic Science, University of York, UK;
[2] Soundscape Voice Evidence, Lancaster, UK; [3] J P French Associates, York, UK
{kk667 | paul.foulkes | peter.french}@york.ac.uk, ck@soundscapevoice.com

ABSTRACT

This study explores the interplay between perception and acoustics, focussing on breathy voice. Using perceptual analysis, four forensic speech analysts rated 22 spontaneous speech samples with regard to whether they were breathy or non-breathy. The voices rated to be the most extreme on the breathy/ non-breathy continuum were then analysed acoustically. Spectral slope and additive noise characteristics were obtained from vowels and sonorant consonants using VoiceSauce. Significant correlations were found between the perception of breathiness and three acoustic measures, namely the intensity difference between the lowest two harmonics, the intensity difference between the lowest harmonic and the harmonic closest to the first formant, and cepstral peak prominence. Our results confirm that the findings from previous studies in relation to non-spontaneous speech are also applicable to spontaneous speech samples. Further, there appears to be no detriment when broadening the sample to include sonorant consonants as well as vowels.

Keywords: voice quality, breathy voice, forensic speaker comparison

1. INTRODUCTION

Voice quality (VQ) can be a highly individual marker of a speaker's voice, due both to its anatomical and habitual origin [26]. It is widely considered an important variable for forensic speech science [16], a field that seeks to identify features with power to discriminate between individuals.

Owing to its multidimensional nature [4], VQ is generally analysed perceptually. However, there are ongoing efforts to utilise acoustic analysis in order to confirm perceptual VQ judgements.

Perceptual judgements are inherently subjective, as standards and thresholds may vary from rater to rater [25]. Also, individual raters' standards often lack stability. For example, the range of voices presented in one rating session can cause a drift in VQ judgement when re-rating the same voice [14]. Nonetheless, the perceptual approach does have advantages over acoustic approaches. Perceptual analysis enables a holistic assessment of a speaker's overall VQ, including aspects of respiration, phonation and articulation [4]. In contrast, acoustic measurements can only provide information on very specific VQ aspects, e.g. the open glottal quotient. Furthermore, the human ear is capable of capturing fine-grained VQ differences even under less than optimal conditions [23]. Typically, the recordings which forensic speech experts face are of poor technical quality and contain only partially analysable speech. This may limit or even prevent accurate acoustic measurements. A reduction in low-frequency energy, for example, which is common for telephone-transmitted speech, often affects the frequency of the first formant [6]. Therefore further research is needed into how stable various acoustic parameters are within speakers, and across different transmission channels and speaking styles.

Given that both perceptual and acoustic approaches have strengths and weaknesses - it is preferable to combine the strengths of both approaches to assess the multidimensionality of VQ. This would position VQ assessments in line with other variables commonly assessed in forensic casework using a combined auditory-acoustic approach (e.g. vowels and consonants).

This study focuses on breathy voice, to assess the extent to which it is possible to combine auditory-perceptual and acoustic approaches.

2. ACOUSTICS OF BREATHY VOICE

2.1. Spectral slope parameters

Breathiness arises from incomplete or non-simultaneous glottal closure resulting in a higher open quotient. This, in turn, yields a strong first harmonic amplitude (H1) [22] and a steep spectral slope [18].

Previous studies analysed spectral slope using harmonic-based and formant-based measurements. Harmonic-based measurements obtain amplitude differences between (1) the first and the second harmonic (H1−H2), (2) the second and fourth harmonic (H2−H4), or (3) the fourth harmonic and

the harmonic closest to 2 kHz (H4−H2K). Formant-based measurements calculate the difference in amplitude between H1 and the harmonics closest to the first three formants (A1, A2, A3; i.e. H1−An).

2.2. Additive noise parameters

In breathy voice high frequency aspiration noise is generated via a persistent glottal gap, causing a decrease in additive noise measurements [18]. Respiration noise is commonly measured by obtaining harmonic-to-noise ratio (HNR) and cepstral peak prominence (CPP). HNR displays the amplitude difference between harmonic and noise energy [8] and decreases in breathy voice [e.g. 12]. HNR can be measured for various frequency bands, typically 0−500/1500/2500 Hz (labelled HNR05/15/25).

CPP is a measure of cepstral peak amplitude relative to the overall amplitude. As a measure of periodicity it is helpful in detecting less periodic signals, e.g. mid and high frequency ranges in breathy voice due to aspiration noise [18]. Lower CPP values correlate with breathy phonation due to the low-intensity higher harmonics [e.g. 34].

3. CORPUS DATA

3.1. Previous studies

Studies of the modal/breathy distinction have investigated either contrastive phonation types of various languages [10, 13, 20] or pathological voices [e.g. 1]. Non-pathological voices and non-contrastive uses have been neglected. Furthermore, most studies have based their analysis on sustained vowels [e.g. 1, 18], vowels from isolated words [e.g. 10, 13, 34], read speech [e.g. 41], or synthetically manipulated stimuli [e.g. 11, 22]. Studies of spontaneous speech are rare [31]. However, [36] reports significant differences in perceptual judgements due to shorter vowel duration and assimilation processes found in spontaneous speech.

3.2. Current study

A selection of non-pathological breathy voice samples was compiled using six corpora of spontaneous conversation from male speakers of British English [15, 17, 21, 28, 29, 35]. The samples were all provided at 44.1 kHz frequency and 16-bit resolution sampling. The first author chose 22 voices based on auditory-perceptual analysis, aiming to reflect a natural mixture along the breathy/non-breathy continuum. Approximately three minutes of speech was extracted from each sample. Using Audacity (version 2.1.2) [2] the maximum intensity level was equalised across samples (max. amplitude -1.0 dB, remove DC offset, center on 0.0 vertically).

4. METHODOLOGY

4.1. Auditory-perceptual investigation

A survey was conducted to generate perceptual ratings for breathiness. Four listeners were engaged, all experts in forensic speech analysis, involved in training and research on VQ. All regularly use the same analysis scheme – a modified Vocal Profile Analysis (VPA) [27] – to rate VQ in forensic casework.

Using the survey tool Qualtrics, the participants were provided with the 22 voices in random order. For each sample they were asked: *'Would you mark breathiness as a dominant feature of this speaker's voice?'* Three answer choices were given together with a comment box: *'(1) Yes, (2) No, it is present but not dominant, (3) No, it is absent'*. The survey took 30-40 minutes. The listeners were allowed to listen to the samples as often as they liked and could leave and resume the survey at any point. They used closed-cup headphones in a quiet environment.

The voices which were rated by all four listeners to be the most extreme on both ends of the breathy/non-breathy continuum were chosen for acoustic analysis. 8 voices qualified: 5 were rated as dominantly breathy and 3 as non-breathy. The between-rater consistency for these 8 voices was established. Cohen's Kappa (κ) was calculated for each rater pair using RStudio (Version 1.1.463) [30]. Table 1 shows that all pairs of raters reached at least 'moderate' agreement. Two pairs (1-2, 2-4) reached 'substantial' agreement, and one pair (1-3) obtained 'almost perfect' agreement (κ=0.86).

Rater-pair	Kappa	z	p-value	Agreement
1-2	0.65	2.83	0.005	substantial
1-3	0.86	2.66	0.008	almost perfect
1-4	0.50	2.19	0.029	moderate
2-3	0.56	2.83	0.005	moderate
2-4	0.75	2.19	0.029	substantial
3-4	0.43	2.19	0.029	moderate

Table 1: Between-rater agreement in the perception survey (Cohen's Kappa for rater-pairs; weights: equal; subjects: 8; raters: 2)

4.2. Acoustic investigation

Acoustic analysis was carried out on the selected voices. Generally, acoustic analysis of phonation is based on vocalic segments only [13, 18], as vowels by nature contain source-specific information. However, forensic recordings might be of very short duration and therefore

can be restricted in terms of analysable speech available. Therefore, the data used in the present study included all sonorants (vowels, glides [j, w], liquids [l, r] and nasals [m, n, ŋ]), as they all contain glottal source characteristics. Sonorants were manually segmented using oscillographic, spectrographic and perceptual-impressionistic information and labelled on a segment-by-segment basis using Praat textgrids [5]. When comparing breathy voices with non-breathy voices, initial visual examination suggests that sonorant consonants and vowels behave similarly in terms of central tendency. Accordingly, in the following acoustic investigation vowels and sonorant consonants were combined.

4.2.1. Measurement procedure

VoiceSauce (version 1.31) [32] was used to take measurements from labelled segments. Default settings were applied: 0.96 pre-emphasis, 25 ms window length, measurements at 1 ms frame shift. The lower F0 range was adjusted to 40 Hz to capture potential intermittent low frequency creak components. The maximum measureable F0 limit was set to 300 Hz. To prevent formants from boosting nearby harmonic amplitudes, the formant-corrected harmonic amplitude measurements [19] implemented in VoiceSauce were obtained and marked by an asterisk (e.g. H1*−H2*). All measurements were averaged across each labelled sonorant. Thus there were 680-1149 averaged measurements per speaker.

4.2.2. Hypotheses

Table 2 summarises the acoustic measurements taken. We predicted the voices rated as dominantly breathy to show steeper spectral slope and lower additive noise. Furthermore, we predicted H1*−H2* to be most useful, as it is a rough indicator of open quotient [33].

Measure	Parameter	Predicted Effect	Prev. Studies
Spectral slope	H1*−H2*	non-breathy < breathy	22
	H2*−H4*		11, 24
	H4*−H2K*		24
	H1*−An*		13, 34
Additive noise	CPP	non-breathy > breathy	18
	HNR		12

Table 2: Predicted effects for spectral slope and additive noise measurements in breathy voice.

4.2.3. Data analysis

We generated boxplots for each acoustic parameter using RStudio (Version 1.1.473) [30], and we used the lme4 package [3] to perform linear mixed effects analyses on the relationship between perceptual ratings and acoustic parameters taken from all sonorants. VQ classification (breathy/non-breathy) was entered into each model as a fixed factor. Speaker-specific variation was accounted for by including by-speaker random slopes. The alpha level was set at $p<0.05$.

5. RESULTS

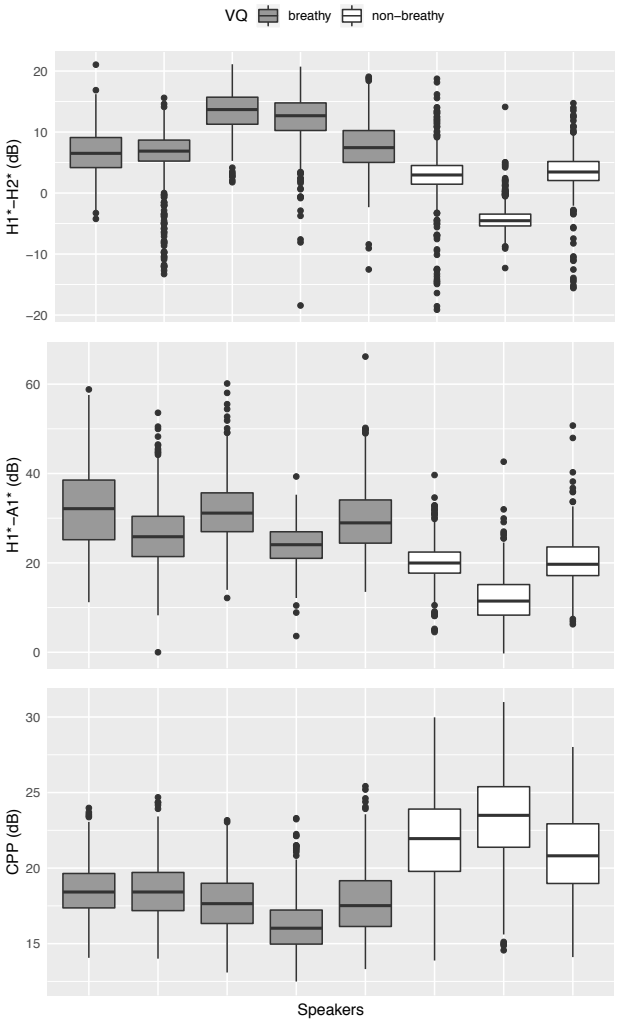

Figure 1: Boxplots for acoustic parameters revealing the clearest differences comparing breathy with non-breathy voices (H1*−H2*, H1*−A1*, CPP).

Figure 1 illustrates the results from the acoustic analysis of all sonorants. Overall, clear differences can be seen between the speakers rated to be dominantly breathy (in grey) and those rated to be non-breathy (in white). Indeed, the distributions for each speaker reveal very little overlap between breathy and non-breathy VQ for H1*−H2*, H1*−A1* and, in particular, CPP.

Table 3 shows the results from the linear mixed effects modelling for all sonorants for each acoustic parameter. Confirming the impressions gained from Figure 1, significant differences were found in

H1*−H2* ($p<0.05$), H1*−A1* ($p<0.05$) and CPP ($p<0.01$). Results close to significance were obtained for HNR05 ($p=0.097$) and HNR15 ($p=0.077$). The other measures did not show significant differences between breathy and non-breathy ratings based on all sonorants.

Model	Estimate	SE	t	df	Pr(>|t|)
H1*−H2*					
(Intercept)	9.20	1.47	7.27	4.00	0.003
non-breathy	-8.51	2.84	-3.00	3.50	**0.047***
H2*−H4*					
(Intercept)	8.44	1.27	7.73	4.00	0.003
non-breathy	-0.01	1.72	-0.01	5.94	0.994
H4*−H2K*					
(Intercept)	5.27	0.97	5.42	3.98	0.007
non-breathy	-1.17	2.10	-0.55	3.11	0.719
H1*−A1*					
(Intercept)	28.78	1.70	17.92	4.00	0.000
non-breathy	-11.15	3.21	-3.47	3.35	**0.034***
H1*−A2*					
(Intercept)	31.70	1.40	22.73	4.00	0.000
non-breathy	-12.04	4.75	-2.53	2.39	0.107
H1*−A3*					
(Intercept)	25.19	1.34	18.82	4.01	0.000
non-breathy	-9.73	5.13	-1.89	2.30	0.182
HNR05					
(Intercept)	8.03	2.47	3.25	4.00	0.031
non-breathy	9.73	4.44	2.19	3.81	0.097
HNR15					
(Intercept)	17.73	1.94	9.13	4.00	0.001
non-breathy	7.57	2.93	2.24	4.88	0.077
HNR25					
(Intercept)	21.94	1.70	12.93	4.00	0.000
non-breathy	5.28	3.01	1.75	3.88	0.157
CPP					
(Intercept)	17.78	0.41	43.27	4.00	0.000
non-breathy	4.19	0.79	5.31	3.52	**0.009****

Table 3: Estimate, standard error estimates (SE), t statistics, Satterthwaite approximated degrees of freedom (df) and predicting VQ classification (Pr(>|t|)) for each model (acoustic parameter). All models included by-speaker random slopes.

6. DISCUSSION

The present study confirms the capability of two low frequency spectral slope parameters (H1*−H2*, H1*−A1*) and one additive noise parameter (CPP) to distinguish auditory-impressionistic judgements of dominantly breathy voices from non-breathy voices. These results are in line with previous studies [1, 10, 13, 18, 20, 34], but, extended the findings from elicited speech to spontaneous speech samples. Mid-to-high frequency spectral slope parameters have previously been found to support the perception of breathiness (H2*−H4* [11, 24], H4*−H2K* [24], H1*−A2* [13] and H1*−A3* [34]). This was not the case here.

Given the more complex nature of spontaneous speech and the weaker intensity of sonorant consonants the results of the present study are promising.

7. CONCLUSION AND FUTURE WORK

Our results indicate that the perception of breathy VQ in spontaneous speech is captured mainly in a steep spectral slope of low frequency ranges (H1*−H2*, H1*−A1*) and in a low cepstral peak prominence (CPP). This outcome lays open the potential to formally adopt the combined auditory-acoustic approach for the assessment of VQ when breathiness is involved.

The auditory-perceptual approach is still the 'gold standard' in VQ analysis [31], which we do not want to challenge. However, our results demonstrate that there is potential for perceptual analysis to be corroborated by acoustic measurements. This would most likely have a positive effect on within-rater and between-rater consistency. Including sonorant consonants increases the practicability of this analysis in the forensic setting as speech samples are often short.

It remains to be examined how the measurements investigated here perform in recordings of poorer quality. Work in progress will test the effect of a mobile-landline telephone filter on acoustic measurements to assess the robustness under forensically realistic conditions.

8. ACKNOWLEDGEMENTS

We thank our two additional expert listeners Richard Rhodes and Jessica Wormald. Further thanks go to Amanda Cardoso and Márton Sóskuthy for assistance with statistical analysis.

9. REFERENCES

[1] Alpan, A., Schoentgen, J., Maryn, Y., Grenez, F., Murphy, P. 2009. Cepstral analysis of vocal dysperiodicities in disordered connected speech. *Proc. 10th Interspeech* Brighton, 959–972.

[2] Audacity Team 2017. Audacity (R): Free Audio Editor and Recorder [Computer application]. Version 2.1.2. released 25/11/2017. URL: http://www.audacityteam.org

[3] Bates, D., Maechler, M., Bolker, B. 2011. lme4. R package version 0.999375-38.

[4] Beck, J. M. 2005. Perceptual analysis of voice quality: the place of vocal profile analysis. In: Hardcastle, W. J., Mackenzie Beck, J. (eds), *A figure of speech. A festschrift for John Laver*. New York: Routledge, 285–322.

[5] Boersma, P., Weenink, D. 2017. Praat [computer program]. Version 7.0.22 (released 15/11/2017) URL: http://www.praat.org/

[6] Byrne, C., Foulkes, P. 2004. The 'mobile phone effect' on vowel formants. *International Journal of Speech Language and the Law* 11.1, 83–102.

[8] de Krom, G. 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *J. Speech Lang. Hear. Res.* 36.2, 254–266.

[10] Fraile, R., Godino-Llorente, J. I. 2014. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control* 14, 42–54.

[11] Garellek, M., Keating, P., Esposito, C. M., Kreiman, J. 2013. Voice quality and tone identification in White Hmong. *Journal of the Acoustical Society of America* 133, 1078–1089.

[12] Garellek, M. 2012. The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics* 31, 152–161.

[13] Garellek, M., Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *JIPA* 41.2, 185–205.

[14] Gerratt, B., Kreiman, J., Antonanzas-Barroso, N., Berke, G. S. 1993. Comparing internal and external standards in voice quality judgments. *J. Speech Lang. Hear. Res.* 36, 14–20.

[15] Gold, E., Ross, S., Earnshaw, K. 2018. The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proc. 19th Interspeech* Hyderabad, 2748–2752.

[16] Gold, E., French, P. 2011. International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18.2, 293–307.

[17] Haddican, W., Foulkes, P. 2017. *A comparative study of language change in Northern Englishes*. [Data Collection]. Colchester, Essex: ESRC. URL: http://reshare.ukdataservice.ac.uk/851013/

[18] Hillenbrand, J., Cleveland, R. A., Erickson, R. L. 1994. Acoustic correlates of breathy voice quality. *J. Speech Lang. Hear. Res.* 37, 769–778.

[19] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America* 121.4, 2283–2295.

[20] Keating, P., Esposito, C., Garellek, M., Khan, S., Kuang, J. 2011. Phonation contrasts across languages. *Proc. 17th ICPhS* Hong Kong, 1046–1049.

[21] Kirchhübel, C. 2013. *The acoustic and temporal characteristics of deceptive speech,* Doctoral dissertation, University of York.

[22] Klatt, D. H., Klatt, L. C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87, 820–857.

[23] Köster, O., Köster, J. P. 2004. The auditory-perceptual evaluation of voice quality in forensic speaker recognition. *The Phonetician* 89, 9–37.

[24] Kreiman, J., Garellek, M., Esposito, C. M. 2011. Perceptual importance of the voice source spectrum from H2 to 2 kHz. *Journal of the Acoustical Society of America* 130, 2570.

[25] Kreiman, J., Vanlancker-Sidtis, D. Gerratt, B. R. 2005. Perception of voice quality. In: Pisoni, D. B., Remez, R. E. (eds), *Handbook of speech perception*. Oxford: Blackwell, 338–362.

[26] Laver, J., Wirz, S., Mackenzie, J., Hiller, S. 1981. A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 13, 139–155.

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

[28] Llamas, C., Watt, D., French, J. P. 2016–19. The use and utility of localised speech forms in determining identity: forensic and sociophonetic perspectives. ESRC ES/M010883/1

[29] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16.1, 31–58.

[30] RStudio Team 2016. RStudio: Integrated development for R. RStudio, Inc., Boston, MA URL: http://www.rstudio.com/

[31] San Segundo, E., Foulkes, P. French, P., Harrison, P., Hughes, V., Kavanagh, C. 2018. The use of the vocal profile analysis for speaker characterization: Methodological proposals. *JIPA* [online].

[32] Shue, Y.-L., Keating , P., Vicenik, C., Yu, K. 2011. VoiceSauce: A program for voice analysis. *Proc. 18th ICPhS* Hong Kong, 1846–1849.

[33] Stevens, K. N., Hanson, H. M. 1995. Classification of glottal vibration from acoustic measurements. In: Fujimura, O., Hirano, M. (eds), *Vocal fold physiology: Voice quality control*. San Diego: Singular, 148–180.

[34] Wayland, R., Jongman, A. 2003. Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics* 31.2, 181–201.

[35] Wormald, J. 2016. *Regional variation in Panjabi-English,* Doctoral dissertation, University of York.

[36] Zraick, R. I., Wendel, K., Smith-Olinde, L. 2005. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice* 19.4, 574–581.