TOWARDS ORDINAL CLASSIFICATION OF VOICE QUALITY FEATURES WITH ACOUSTIC PARAMETERS

Felix Schaeffler^{1,2}*, Matthias Eichner*^{2}*, Janet Beck*^{1,2}
^{1}*CASL Research Centre, Queen Margaret University Edinburgh,* ^{2}*Fitvoice CIC,*
Musselburgh, Scotland, UK, fschaeffler@qmu.ac.uk

Abstract: The human voice is capable of fine-grained variation that results in listener attributions of various psychological, social and biological factors. The complexity of this process is reflected in the number and richness of terms that are used to describe human voices. In this paper we argue that any application that attempts a mapping of the acoustic voice signal onto voice descriptor labels would benefit from an intermediate auditory-phonetic level. As a point of departure we explore the relationships between acoustic parameters and some specific perceptual features derived from Vocal Profile Analysis (VPA), a phonetically motivated voice quality analysis scheme.

Perceptual analysis of voice samples from 133 speakers was carried out using VPA for three key phonation features (creakiness, whisperiness, harshness). We extracted eleven acoustic parameters from the samples and used stepwise linear regression to identify acoustic parameters with predictive value. Samples from female speakers were used to derive regression equations which were then used to predict VPA ratings of male voices. Results show significant predictors for all three phonation features and indicate that predictions for the three phonation types rely mainly on different parameters. If a tolerance of ± 1 scalar degree for the perceptual analysis is accepted, then classification accuracy lies at or above 90% for all three phonation features.

1 Introduction

The human voice shows an extraordinary amount of meaningful variability and an individual's voice quality and timbre typically causes listeners to make a range of affect and personality attributions (e.g. [1]). The semantic space associated with voice description is immense. Laver [2] observed that there are "hundreds of labels" (p. 62) that can be used to describe the sound of voices and suggests a semiotic typology for them, capturing the fact that voice terms sometimes refer to the auditory impression of a voice (consider terms like *thick*, *whining*, *gravelly, plummy*) and sometimes to attributions made towards the speaker or their mood based on the sound of a voice (consider terms like *young, cold, nervous*) or the effect the voice may have on the listener (e.g. *boring, soothing, frightening*).

In contrast to that, voice classification in fields like clinical voice analysis is often rather crude. For example, one of the most widely used systems for scoring disordered voices, the GRBAS scale [3], operates with just 5 dimensions. These dimensions benefit from relatively high inter-rater reliability [4] and may capture some features that are important for the assessment of disordered voices, but are hardly fine-grained enough to provide a basis for a comprehensive description of voice quality.

An alternative approach to voice quality description is the Vocal Profile Analysis (VPA) scheme [5][6]. This scheme aims at a phonetic description of voice quality, independent of health status of a voice. It relies on well-established phonetic parameters for the description of voice quality, goes beyond phonation features and takes into account physiological constraints to sound production at both glottal and supraglottal level. VPA is a perceptual scheme that relies on rating by trained experts. Ratings are based on ordinal seven point scales (zero to six) that indicate presence and strength of a certain voice quality feature. Scalar degrees one to three are reserved for moderate feature strengths, four to six for extreme (often pathological) strengths. As some features are antagonistic, they could also be described as a 15 point scale (minus seven to seven).

In the latest version of VPA, 14 voice quality feature sets are grouped into three domains: vocal tract features, overall muscular tension and phonation features. There are eight vocal tract feature categories, comprising labial, mandibular, lingual tip/blade, lingual body, velopharyngeal, pharyngeal and larynx height settings. There are two muscular tension categories: vocal tract tension and laryngeal tension. The three phonation feature categories address voicing type, laryngeal frication and laryngeal irregularity. The focus of the current paper is on these phonation features. Figure 1 shows the relevant section of the VPA scoring form. Detailed descriptions of the VPA approach to phonation type analysis can be found elsewhere [6], [7], but it is important to note that for VPA scoring of phonation a neutral baseline auditory quality is used (sometimes referred to as "modal voice"). This is associated with (a) full vocal fold adduction without any audible fricative airflow and (b) regular, periodic vocal fold vibration without any audible roughness. Few, if any, speakers have completely neutral phonation, so the term "neutral" is not synonymous with "normal".

Three non-neutral phonation types, falsetto, whisper and creak, may occur in isolation as alternatives to modal voice, in which case they are simply marked as present on the scoring form. Whisper refers to fricative glottal airflow with no vocal fold vibration and creak to low frequency "pulsed" phonation. Whisperiness and creakiness can also occur in combination with voice, as can harshness (associated with irregular vocal fold vibration) and breathiness (a lax phonation with high levels of airflow through the glottis). A very wide range of combinations is thus possible (e.g. harsh, whispery voice). In complex phonation types, scalar degree judgments are used to indicate perceptual balance. For example, in moderately whispery voice, where the fricative component is much less salient than the voice component, voice would be ticked and whisperiness would be marked in the 1-3 scalar degree range. If whisperiness is more salient than the voicing, a scalar degree score of 4-6 would be given. In general, scalar degree 1 means that the setting is just audible, and scalar degree 6 is the maximum degree of that setting that can be produced. Intermittent present of a setting can be marked with "i". The scoring shown in Figure 1 illustrates analysis of a voice that combines voice with extreme whisperiness, moderate harshness and moderate intermittent creakiness. Note that VPA follows phonetic conventions in differentiating between whisperiness (relatively tense phonation with high levels of fricative energy) and breathiness (characterized by very high levels of airflow and laryngeal laxness). This contrasts with systems like GRBAS which use the term "breathy" to describe both whisperiness and breathiness.

Besides voice quality, VPA also allows scores for prosodic features and temporal organization. These features are not discussed further here, but are important extensions for future studies.

C. PHONATION FEATURES									
		Present		Scalar Degree					
		Neutral	Non-neutral	Moderate			Extreme		
	SETTING			1	2	3	4	5	6
12. Voicing type	Voice		√						

	Falsetto								
	Creak		i						
	Creaky					i			
13. Laryngeal frication	Whisper		√						
	Whispery							√	
	Breathy								
14. Laryngeal irregularity	Harsh		√		√				
	Tremor								

Figure 1 - VPA scoring form: phonation features section with example scoring (see text).

VPA has been recommended for a variety of applications and good inter-rater agreement can be reached with training [8]. Detailed studies of the acoustic correlates of VPA scores are relatively rare (see e.g. [9]).

There have been various successful attempts to recognize and classify phonation types (see e.g. [10] for creaky voice) but a holistic acoustic evaluation that assigns VPA-type scores for all relevant voice quality parameters to a speaker or voice sample is currently not available. An automatic and sufficiently fine-grained phonetic description of voice quality would be useful in various contexts, be it voice classification and indexing, voice similarity and confusion judgments and selection of voices in various applications.

Phonetic voice quality classifications could also allow for more systematicity in the analysis of links between voice quality and affect, mood or personality attributions. The necessity of intermediate levels of representation in this context has been stressed in psychological models of emotion perception from speech [11]. In our view, a model that maps voice descriptions to an auditory-phonetic description of voice quality instead of mapping directly to the acoustic signal would have several advantages.

As discussed earlier, voice descriptor labels form a vast and rather unstructured semantic field, and this is a particular challenge for commercial activities that rely on voice labelling. While many labels will have potential to be mapped against acoustic features, it is unclear a) how heterogeneous the various attributions would be across different listener groups (for example, what constitutes *persuasive* or *boring* might differ between age groups or genders) and b) whether attributions remain stable over time or are subject to fluctuations due to changes in culture, fashion etc. It would be possible to attempt a direct mapping of the acoustic signal to descriptive labels, but inclusion of an intermediate layer of auditory-phonetic analysis such as VPA allows for better descriptive stability and transparency in our view. Ideally, the mapping of acoustics to the auditory phonetic layer would be invariant, and group differences as well as changes over time could be described in altered mappings from the auditory-phonetic layer to the voice descriptor labels.

As a prerequisite for such a model, a stable mapping between acoustic and VPA features would need to be established. In this paper we have analysed VPA scores for 133 voices, completed by a highly trained expert (the third author). We focused on three phonetic dimensions from the phonation feature section (creakiness, whisperiness, harshness) and eleven acoustic parameters (see Section 2).

We used stepwise linear regression to uncover acoustic parameters with predictive value and tested models derived from a corpus of women's voices on a comparable corpus of male voices.

2 Method

2.1 Samples

The voice samples for this study were taken from the KayPENTAX Disordered Voice Database [12]. This database contains data from about 700 speakers (classified as normal or pathological by diagnostic category). We derived a sub-selection of 330 samples that satisfied the following criteria:

- The audio data included a connected speech sample as well as a sustained vowel.
- English was the native language of the speaker
- The sample had a valid diagnostic label (including 'normal').

VPA ratings have been completed for 133 samples. We deliberately chose a mixture of healthy and pathological samples to achieve a wide spread of VPA scalar degree ratings.

The DVDB samples consisted of 12 second extracts of readings of the 'Rainbow Passage' [13], cut from the start of the passage. The 133 samples originated from 80 female speakers (60 "normal", 20 "pathological") and 53 male speakers (13 "normal", 40 "pathological").

2.2 Acoustic parameters

Table 1 provides an overview over all acoustic parameters used in the study. Parameters constituted either examples of conventional measures of voice functionality (like mean F0, jitter or shimmer) or have shown good performance in studies of acoustic voice quality parameters (e.g. CPPS, GNE or H1H2c).

Table 1 - Acoustic parameters used for multiple stepwise regression

Parameter	***Description***
MF0	Mean fundamental frequency calculated over the whole sample, using the Praat cross correlation procedure with standard (gender-specific) settings.
SDF0	Fundamental frequency standard deviation calculated over the whole sample, same approach as above.
ShdB	Shimmer (dB), period-to-period amplitude fluctuation, expressed by the logarithm (base 10) of he difference of consecutive periods, multiplied by 20. The parameter is called Shimmer (local, dB) in Praat (cf. Praat manual [14]). Derived from Praat "voice report" functionality.
RAP	Relative average perturbation (RAP) of the fundamental period, a jitter measure. See Buder [15], p. 138 for precise formula. (derived from Praat "voice report" functionality [14])
HNR	Harmonics to noise ratio as implemented in Praat and derived from Praat "voice report" functionality [14].
Slope	Slope of the LTA spectrum (suggested as part of the "Acoustic Voice Quality Index" (AVQI) parameters, see [16] for details and implementation)
Tilt	Tilt of regression line through the long-term average spectrum (suggested as part of the "Acoustic Voice Quality Index" (AVQI) parameters, see [16] for details and implementation)
GNE	Glottal noise excitation ratio – an alternative measure of harmonics to noise ratio, developed to reduce the influence of jitter and shimmer on HNR measures (see [17] [18]), partly implemented in Praat, with modifications by current authors.
H1H2	Difference between the amplitudes of the 1st and 2nd harmonic amplitude,

	implementation in Praat by current authors.
H1H2c	Difference between the amplitude of the 1st and 2nd harmonic, corrected for vocal tract influence (see [19]), implementation in Praat by current authors.
CPPS	Smoothed Cepstral Peak Prominence (CPPS), originally introduced as a dysphonia measure [20] and part of the AVQI parameters, see [16] for details and implementation)

2.3 Acoustic and statistical analysis

Acoustic analysis was performed with Praat 6.0.21 [14] in two processing steps. First, voiced segments were extracted following a procedure published by [16]. From the voiced segments, the 11 acoustic parameters were extracted with a Praat script with different settings for male and female F0 analysis.

For statistical analysis, stepwise linear regression was performed with SPSS for all 11 independent acoustic variables. Separate analyses were run for three dependent variables: VPA creakiness score, VPA whisperiness score and VPA harshness score. Variables were entered in the model if the probability of F was < .05, and removed from the model if probability of F was ≥ .10. Regression analyses were run for the female data and resulting regression equations were used to predict VPA scores for the male data by rounding real numbers to integers.

3 Results

3.1 Regression results (women)

Table 2 - Variables selected by stepwise linear regression for three phonation types and resulting regression equations

Dependent variables	**Independent variables**	**Adjusted R^2**	**Regression equation**
Creakiness	H1H2c	0.23	2.49-0.16 * H1H2c
	MF0	0.33	(MF0 excluded as a predictor due to male/female differences)
Whisperiness	GNE	0.59	7.09-4.81*GNE+0.05*H1H2-0.32*CPPS+0.16*HNR
	H1H2	0.65	
	Slope	0.70	
	CPPS	0.71	
	HNR	0.73	
	-Slope (removed in step 6)	0.74	
Harshness	CPPS	0.45	5.1-0.29*CPPS

Table 2 provides the results from the stepwise regression analysis of the female data. Whisperiness reached the highest amount of explained variation (74%), and GNE is the strongest predictor, which on its own explains almost 60% of the variation. This confirms the role of GNE as a good predictor of (perceived) added noise in the voice signal. For harshness, a single predictor was selected (CPPS), which again confirmed the conventional interpretation of this parameter, although it was originally suggested as a 'breathiness' indicator. Explained variation was moderate for this VPA parameter (45%). Creakiness resulted in the lowest explained variation (33%) but the strongest predictor was a parameter that has been linked to creakiness before [21].

3.2 Prediction of male VPA scores from female regression analysis

We applied the regression equations reported above to the male data and predicted the VPA scalar degrees for the three phonation features. Real numbers resulting from the equation were rounded to the nearest integer. Table 3 provides the frequencies for hits and misses in percent.

Table 3 - hits and misses for male data VPA score predictions in percent

Feature	hits	miss ± 1 scalar degree	miss ± 2 scalar degrees	miss ± 3 scalar degree
Creakiness	21%	53%	21%	6%
Whisperiness	34%	43%	23%	0%
Harshness	36%	51%	13%	0%

Table 4 - True positives (TP), false negatives (FN), false positives (FP), true negatives (TN) and derived classifier metrics and their weighted averages (wA).

Creakiness									
Rating	TP	FN	FP	TN		Precision	Recall/ Sensitivity	Specificity	Accuracy
0	0	0	7	46		0.00	0.00	0.87	0.87
1	12	2	2	37		0.86	0.86	0.95	0.92
2	12	8	0	33		1.00	0.60	1.00	0.85
3	9	4	1	39		0.90	0.69	0.98	0.91
4	6	0	2	45		0.75	1.00	0.96	0.96
5	0	0	2	51		0.00	0.00	0.96	0.96
Sum	39	14	14	251	wA	0.91	0.74	0.98	0.90
Whisperiness									
Rating	TP	FN	FP	TN		Precision	Recall/ Sensitivity	Specificity	Accuracy
0	1	3	4	45		0.20	0.25	0.92	0.87
1	7	0	4	42		0.64	1.00	0.91	0.92
2	4	4	4	41		0.50	0.50	0.91	0.85
3	18	4	0	31		1.00	0.82	1.00	0.92
4	9	1	0	43		1.00	0.90	1.00	0.98
5	1	0	0	52		1.00	1.00	1.00	1.00
6	1	0	0	52		1.00	1.00	1.00	1.00
Sum	41	12	12	306	wA	0.82	0.77	0.97	0.92
Harshness									
Rating	TP	FN	FP	TN		Precision	Recall/ Sensitivity	Specificity	Accuracy
0	4	3	0	46		1.00	0.57	1.00	0.94
1	16	1	1	35		0.94	0.94	0.97	0.96
2	10	0	5	38		0.67	1.00	0.88	0.91
3	7	1	1	44		0.88	0.88	0.98	0.96
4	3	2	0	48		1.00	0.60	1.00	0.96
5	6	0	0	47		1.00	1.00	1.00	1.00
Sum	46	7	7	258	wA	0.90	0.87	0.97	0.95

Table 3 shows that, while the frequency of absolute hits was not satisfactory, the majority of errors tended to be small. If variations of ± 1 scalar degree were tolerated, then hit rates for creakiness increase to 74%, for whisperiness 77% and for harshness 87%.

Table 4 provides detailed results for prediction of ratings for the male data using the model estimated from the female data where variations of ± 1 scalar degree were tolerated. Precision, recall, specificity and accuracy were calculated individually for each class. The average values were then calculated by weighted aggregation of the individual scores according to the prevalence of the corresponding label in the test data.

4 Discussion

While the relative small database of this study limits generalisability, results are generally promising. In all three cases, predictor variables achieve moderate to high levels of explained variance and confirm previous findings regarding the explanatory value of these parameters. It is also worth mentioning that predictions for whisperiness and harshness, while not completely independent, mainly rely on different parameters. GNE was not chosen as a predictor for harshness, and while CPPS was chosen as a predictor for both whisperiness and harshness, its contribution to whisperiness was quite small and it could probably be dropped from the model for whisperiness without major losses in accuracy.

Creakiness regression relied mainly on differences between the first and second harmonic and mean F0, which corresponds with previous accounts, as mentioned above. Creakiness will probably have the most complex acoustic signature of the three phonation features discussed here, and its overall impression might rely more on intra-sample variability than the other two features. A single mean calculated over a speech sample might not provide enough information for the assessment of perceived creakiness in the sample, and future approaches should take further measures (e.g. the standard deviation of the H1H2c measure or measures suggested in [10] into account). The situation is complicated by the fact that creakiness is often intermittent, occurring more frequently in certain prosodic contexts and towards the end of breath groups. For these reasons, a prosodic approach that weighs creakiness differently according to prosodic constituents (like prominent syllables or phrase edges) might add further nuance to the model.

While linear regression is a relatively simple model for prediction and classification and its application to ordinal response variables is not without problems, the prediction results for the male data can serve as a benchmark for more sophisticated approaches. If deviations of ± 1 scalar degree are tolerated, then classification accuracy lies at or above 90% for all three phonation features. Given that deviations of ± 1 scalar degree have been accepted in evaluations of reliability for a number of VPA studies [6], this might be acceptable.

The three features studied here are obviously only a small part of the voice quality settings that can be described with VPA, and future studies will look into the acoustic correlates of other sections of the VPA scheme and how VPA and acoustic features can be mapped onto voice descriptor labels.

5 Literature

[1] M. Latinus and P. Belin, ‘Human voice perception’, *Curr. Biol.*, vol. 21, no. 4, pp. R143–R145, Feb. 2011.

[2] J. Laver, ‘Labels for voices’, *J. Int. Phon. Assoc.*, vol. 4, no. 2, pp. 62–75, Dec. 1974.

[3] M. Hirano, *Clinical Examination of Voice*. Vienna: Springer-Verlag, 1981.

[4] K. Nemr *et al.*, 'GRBAS and Cape-V Scales: High Reliability and Consensus When Applied at Different Times', *J. Voice*, vol. 26, no. 6, pp. 812.e17-812.e22, Nov. 2012.

[5] J. Laver, S. Wirz, J. Mackenzie, and S. M. Hiller, 'A perceptual protocol for the analysis of vocal profiles', in *The Gift of Speech: Papers in the Analysis of Speech and Voice.*, J. Laver, Ed. Edinburgh: Edinburgh University Press, 1991, pp. 265–280.

[6] J. Mackenzie-Beck, 'Perceptual Analysis of Voice Quality: the Place of Vocal Profile Analysis', in *A Figure of Speech*, W. J. Hardcastle and J. Mackenzie-Beck, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2005, pp. 285–322.

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

[8] E. San 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.*, pp. 1–28, Jun. 2018.

[9] Z. A. de Camargo, P. G. Coutinho, S. Madureira, and L. C. Rusilo, 'Voice quality description from a phonetic perspective: Supralaryngeal and muscular tension settings', in *ICPhS*, 2015.

[10] T. Drugman, J. Kane, and C. Gobl, 'Data-driven detection and analysis of the patterns of creaky voice', *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1233–1253, Sep. 2014.

[11] T. Bänziger, G. Hosoya, and K. R. Scherer, 'Path Models of Vocal Emotion Communication', *PLOS ONE*, vol. 10, no. 9, p. e0136675, Sep. 2015.

[12] *Disordered Voice Database, model 4337.* developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Lab. KayPENTAX Corporation, 1994.

[13] G. Fairbanks, 'The rainbow passage', *Voice Articul. Drillbook*, vol. 2, 1960.

[14] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (version 6.0.21) [computer software]*. 2016.

[15] E. H. Buder, 'Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902-1990', in *Voice Quality Measurement*, R. D. Kent and M. J. Ball, Eds. San Diego, CA: Singular Publishing Group, 2000.

[16] Y. Maryn and D. Weenink, 'Objective Dysphonia Measures in the Program Praat: Smoothed Cepstral Peak Prominence and Acoustic Voice Quality Index', *J. Voice*, vol. 29, no. 1, pp. 35–43, Jan. 2015.

[17] D. Michaelis, T. Gramss, and H. W. Strube, 'Glottal-to-noise excitation ratio - a new measure for describing pathological voices', *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[18] J. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, 'The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders', *J. Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[19] M. Iseli and A. Alwan, 'An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation', in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, vol. 1, pp. I-669–72 vol.1.

[20] Y. D. Heman-Ackah *et al.*, 'Cepstral peak prominence: A more reliable measure of dysphonia', *Ann. Otol. Rhinol. Laryngol.*, vol. 112, no. 4, pp. 324–333, 2003.

[21] P. Keating, M. Garellek, and J. Kreiman, 'Acoustic properties of different kinds of creaky voice', *Proc. 18th ICPhS Glasg.*, 2015.