

Identifying regions of non-modal phonation using features of the wavelet transform

John Kane, Christer Gobl

Phonetics Laboratory, Centre for Language and Communication Studies,
Trinity College, Dublin

Abstract

The present study proposes a new parameter for identifying breathy to tense voice qualities in a given speech segment using measurements from the wavelet transform. Techniques that can deliver robust information on the voice quality of a speech segment are desirable as they can help tune analysis strategies as well as provide automatic voice quality annotation in large corpora. The method described here involves wavelet-based decomposition of the speech signal into octave bands and then fitting a regression line to the maximum amplitudes at the different scales. The slope coefficient is then evaluated in terms of its ability to differentiate voice qualities compared to other parameters in the literature. The new parameter (named here *Peak Slope*) was shown to have robustness to babble noise added with signal to noise ratios as low as 10 dB. Furthermore, the proposed parameter was shown to provide better differentiation of breathy to tense voice qualities in both vowels and running speech.

Index Terms: Voice quality, glottal source, wavelets

1. Introduction

The majority of approaches to speech processing involve using fixed window lengths and algorithms which embed assumptions about the type of signal that is expected. For instance for f_0 detection and spectral representations (e.g., MFCCs, LPCs) window sizes are typically not longer than 32 ms. As two pulse periods are typically required for periodicity or harmonicity information this limits possible f_0 values to being above 62.5 Hz. Generally speakers would produce f_0 values above this cut-off. However in the case of creaky phonation f_0 values can be as low 10 Hz, making the above window size unsuitable.

It is a current direction of our research to attempt to identify various regions in the speech waveform which can then be used to focus analysis strategies according to that region type. In the case of creaky phonation a method like that described in [1] could be used for identifying creaky segments and this could allow for f_0 or glottal closure instant (GCI) detection methods that allow for very long glottal pulse periods.

In breathy speech segments glottal closure instants may be much smoother than the sharp closures normally seen in modal segments. Correctly identifying breathy regions could allow for the deployment of an algorithm that is suited to analysing this type of occurrence (e.g., the GCI detection algorithm in [2]).

Furthermore, automatic identification of regions of non-modal phonation is useful for expressive unit selection synthesis. In the case of short utterances (e.g., “Yeh”), which can have very different meanings depending on how the utterance was produced, knowledge of the voice quality used can help facilitate the retrieval of appropriate speech units from large corpora.

The characterisation of voice qualities typically requires source-filter decomposition. The various approaches to this inverse filtering problem, however, tend to produce significant errors, particularly when filtering running speech. Also, it would be advantageous to have some *a priori* knowledge of the voice quality mode in order to tailor inverse filtering strategies.

For these reasons a measure of voice quality that does not require inverse filtering is clearly desired. The current study looks to utilise features of the wavelet transform for this very purpose. The use of wavelets has become popular in speech processing particularly in f_0 and GCI detection (see for example [2]-[4]). Wavelet based approaches in speech processing have also been shown to be robust against noisy conditions in voiced/unvoiced detection and other areas [5]. We hope to demonstrate that features of the wavelet transform can be used for robust voice quality identification without the need for inverse filtering.

2. Proposed method

The method of detecting voice qualities proposed here is in part motivated by the observations in [2]. In a previous study [3] the wavelet transform was applied at two or three smallest scales (relating to higher frequencies). By taking the modulus maxima, glottal closure instants (GCIs) could be well detected in voiced speech where there was sharp glottal closure. It was noted in [2], however, that for smoother glottal closures (for example in breathy voice or when voicing is offsetting) only considering the smallest scales was unsuitable. Hence, the relative importance of the scale of the modulus maxima is different for breathy phonation compared to modal phonation. We wanted to investigate if an acoustic feature could be designed to characterise this phenomenon and examine whether such a feature would be suitable for differentiating voice qualities on a breathy to tense scale.

We decided to use Eq. (1) as the mother wavelet (equation comes from [4]), where $f_s = 16$ kHz, $f_n = \frac{f_s}{2}$ and $\tau = \frac{1}{2f_n}$

$$g(t) = -\cos(2\pi f_n t) \cdot \exp\left(-\frac{t^2}{2\tau^2}\right) \quad (1)$$

Each wavelet of the analysed signal is obtained by convolving the speech signal, $x(t)$ with $g(\frac{t}{s_i})$, where $s_i = 2^i$ and $i = 0, 1, 2, \dots, 5$. This resulted in an octave band filter bank, with filters having centre frequencies of: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz.

For a given speech segment the above decomposition is performed and the modulus maximum at each scale is measured. A straight regression line is then fitted to these peak amplitudes, see Fig. 1. In Fig. 1 it can be seen that the slope of the fitted lines differs for the three voice qualities. Hence, the proposed

parameter is the slope coefficient of the regression line (henceforth referred to as *Peak Slope*).

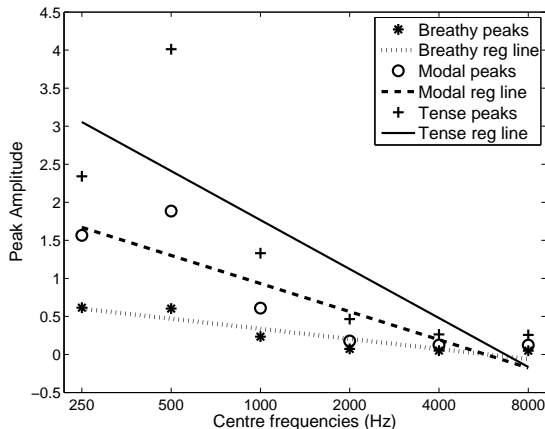


Figure 1: Wavelet peak amplitudes with regression lines for an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities.

3. Evaluation

3.1. Speech data

Two sets of speech data were used for evaluating the proposed method in the current study. The first was a dataset of Finnish vowels (/a e i o u y æ ø/) produced in three voice qualities (originally labeled: breathy, normal and pressed) by six female and 5 male speakers. These were the same recordings as used in [6] and totalled 792 vowel segments. Recording was done using a unidirectional Sennheiser electret microphone with a preamp (LD MPA10e Dual Channel Microphone Preamplifier) and a digital audio recorder (iRiver iHP-140). Audio was digitised at 44.1 kHz and downsampled to 16 kHz for the present study. Phase distortion imposed by the recording system was removed by getting the impulse response of the system and convolving recorded signals with the impulse response time-reversed.

Due to the ambiguous label of ‘normal’ as a voice quality we sought to re-label the dataset using Laver’s labelling framework [7] in order to have three independent sets of voice qualities (i.e. breathy, modal and tense). We conducted perception tests with three judges, all experienced in voice quality research. Judges rated sound files on a five point scale from breathy to tense. Vowel samples were excluded from the current study if the standard deviation of ratings was more than one (disagreement on voice quality label) or if the mean rating was 0.75 away from the corresponding voice quality label it was originally given. This strict criteria ensured little or no crossover in terms of the perception of the voice quality sets. This process resulted in the set of 792 vowels being reduced to 478 for the present work. Cohen’s Kappa (κ) increased from before ($\kappa = 0.526$) exclusions to after ($\kappa = 0.717$), demonstrating largely increased inter-rater agreement.

The second dataset contained 10 sonorant-only sentences spoken by three male speakers in breathy, modal and tense voice qualities (i.e. 90 sentences in total). All three speakers were experienced in speech research and repeated the utterance until it was deemed that the voice quality had been maintained for the entire sentence. Sentences were recorded in a semi-

anechoic recording studio using high quality recording equipment (a B&K 4191 free-field microphone and a B&K 7749 pre-amplifier). Audio was digitised at 44.1 kHz (using a Lynx-two sound card) and then downsampled to 16 kHz. Sentences were manually segmented and annotated at the phoneme level.

3.2. Experiments

Initially we wanted to test the ability of the *Peak Slope* parameter at differentiating breathy, modal and tense voice qualities in simple utterances. We also wanted to test the robustness of the parameter to background noise. To this end we parameterised each segment in the vowel dataset first without noise, then with babble noise added at 25 dB, 15 dB, 10 dB and 5 dB signal-to-noise ratio (SNR). A long waveform of babble noise was obtained from <http://spib.rice.edu/spib/data/signals/noise/babble.html> and was added to each clean vowel segment by randomly selecting start points in the long waveform.

For comparison we selected voice quality parameters that have been shown to be useful at discriminating breathy to tense voice qualities (see e.g., [6]) and also parameters which have been used in applied work on voice quality. As a result we selected the normalised amplitude quotient (NAQ) [8] and the difference between the first two harmonics of the narrowband glottal source spectrum in dB (H1-H2) [9]. These two parameters were previously shown to perform well at differentiating these voice qualities[6].

We also opted to include the H1*-H2*, using the inverse filtering strategy described in [9]. The harmonic amplitudes in H1*-H2* are obtained by subtracting Eq. (2) from both harmonics (where f is the frequency of the given harmonic).

$$20 \log_{10} \frac{F1^2}{F1^2 - f^2} \quad (2)$$

A further parameter suggested in [9] is H1*-A3* which is used to measure spectral slope. H1* is again derived using the method above. A3 refers to the maximum harmonic in the third formant peak. In order to neutralise the effects of the first two formants Eq. (3) is added to A3 (where $\tilde{F}1$ and $\tilde{F}2$ are the formant frequencies of a proposed neutral vowel), giving A3*. $\tilde{F}1$ and $\tilde{F}2$ are set as 555 Hz and 1665 Hz as in [9].

$$\frac{\left[1 - \left(\frac{F3}{\tilde{F}1}\right)^2\right] \left[1 - \left(\frac{F3}{\tilde{F}2}\right)^2\right]}{\left[1 - \left(\frac{F3}{\tilde{F}1}\right)^2\right] \left[1 - \left(\frac{F3}{\tilde{F}2}\right)^2\right]} \quad (3)$$

Some researchers also use H1-A3 without any formant based compensation [10]. As this measure is also used in applied work on running speech we also consider it in the present study.

Peak slope values were calculated by measuring the maximum peaks at each scale for the middle part of the utterance. NAQ and H1-H2 were measured pulse-by-pulse after automatic inverse filtering using the iterative adaptive approach [11]. Throughout the experimentation the window length for measuring spectral parameters was three pulse lengths centred on a GCI (obtained using the method in [13]) and using a Hamming window. The parameter values, as well as H1*-H2*, H1*-A3* and H1-A3, were then averaged for each vowel segment.

We first conducted an experiment using a leave-one-speaker-out approach to test if parameter values could be generalised to unseen speakers. This was done by establishing thresholds on the basis of individual parameter values for all speakers

but one. These thresholds were then used in order to test identification accuracy on the held out set. This was repeated for all 11 speakers and results were averaged. Defining thresholds was done using a multivariate optimisation algorithm [12] to maximise identification accuracy in the training set by varying two threshold values. We then used parameter values for the entire vowel dataset as a training set for setting thresholds and tested this on parameter values extracted from the same dataset but with different noise levels added before parameterisation.

The second set of experiments involved analysing the sentence dataset using the parameters. Parameter values were measured for each phoneme segment in each sentence. The distributions of the parameter values across the three voice qualities was then examined. We used thresholds, for each parameter, which were established on the basis of the distribution of values from the entire ‘clean’ vowel dataset and measured voice quality identification accuracy using these thresholds. This would show, to certain extent, how stable each of the parameters are when applied to running speech.

Peak slope values were measured in the sentence dataset by first performing the wavelet-based decomposition of the sentence signal and then measuring maxima on the different scales within each phoneme boundary. NAQ and H1-H2 were obtained by first doing automatic inverse filtering [11] and then analysing the pulses within phoneme boundaries and averaging them for each phoneme. The remaining parameters were measured using the same method but without inverse filtering.

In both datasets distributions of parameter values were analysed using Spearman’s Rank Coefficient (ρ) with the voice quality label being the independent variable and the parameter values being the dependent variable. This was used as a measure of the ability of the given parameter to differentiate the three voice qualities.

Finally, it has been suggested in previous studies that voice source parameters are affected by supraglottal settings in voiced consonants [14]. To determine whether the new parameter, or indeed the other parameters, are affected by such settings we carried out within-speaker two-way ANOVAs (with voice quality label and vowel/non-vowel as factors) on each of the parameter values, taken from the sentence dataset.

4. Results

Initial analysis of the vowel dataset showed *Peak Slope* ($\rho = -0.85$), NAQ ($\rho = -0.72$), H1-H2 ($\rho = -0.67$) and H1-A3 ($\rho = -0.52$) to be considerably better at differentiating voice qualities than H1*-H2* ($\rho = -0.36$) and H1*-A3* ($\rho = -0.11$) and hence they were removed from further analysis.

Distributions of the remaining four parameter values across the three voice qualities are shown in Fig. 2. Derived thresholds are marked with dashed lines. Fig. 3 shows the identification accuracies of the leave-one-speaker-out tests on the undistorted vowels and accuracy scores on the vowels with the different noise levels added and using thresholds set on the full vowel dataset without noise. The *Peak Slope* parameter produces higher identification accuracy than the other three in the leave-one-speaker-out tests. It also displays robustness down to SNR = 10 dB, with higher accuracy at this noise level compared to the other parameters with the lower noise levels. At SNR = 5 dB the accuracy is dramatically affected with the other three parameters showing comparable trends. NAQ, as expected, also displays robustness up to SNR = 10 dB and H1-H2 up to SNR = 15 dB. H1-A3 produces clearly lower accuracy levels and is heavily affected by noise from SNR = 15 dB.

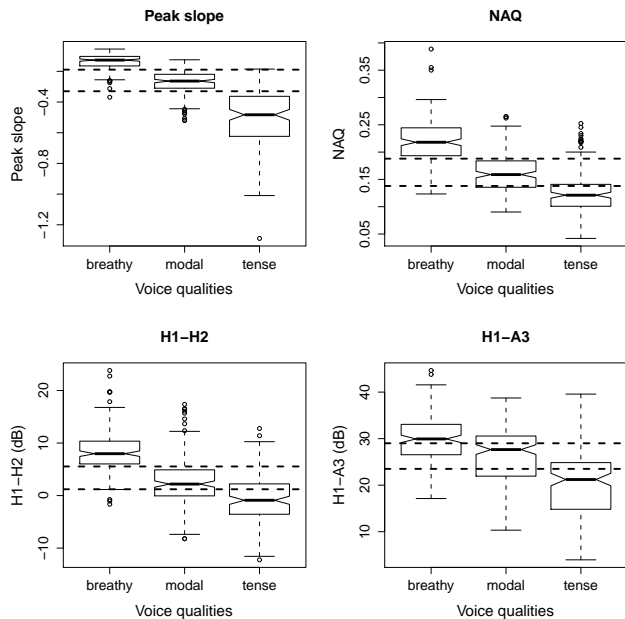


Figure 2: Distributions of the four parameters for breathy, modal and tense voice qualities from the **vowel** dataset. Proposed thresholds for voice quality identification (in the vowels with noise added and the in the sentence dataset) are shown as a dashed line.

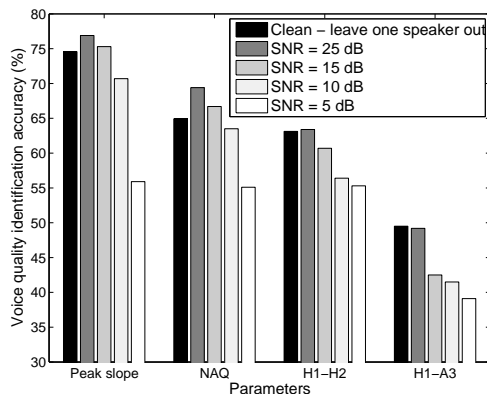


Figure 3: Voice quality identification accuracy (%) in the vowel dataset with different levels of babble noise added, for the four parameters.

Results from the sentence analysis show that in terms of the overall distribution of values the new parameter provides clearest differentiation of the three voice qualities (see Fig. 4). *Peak Slope* gave a higher ρ value (-0.67) than NAQ ($\rho = -0.61$), H1-H2 ($\rho = -0.64$) and H1-A3 ($\rho = -0.59$). Also, using the thresholds set from analysis of the ‘clean’ vowels identification rates were again highest for *Peak Slope* (67 %) compared with NAQ and H1-H2 (49 %) and H1-A3 (46 %).

ANOVAs as expected revealed significant differences across voice quality labels for *Peak Slope* [$F = 92.1644$, $df = 2$, $p < 0.001$], as well as for NAQ [$F = 27.5120$, $df = 2$, $p < 0.001$], H1-H2 [$F = 53.6246$, $df = 2$, $p < 0.001$] and H1-A3 [$F = 82.3829$, $df = 2$, $p < 0.001$]. It was also observed, however,

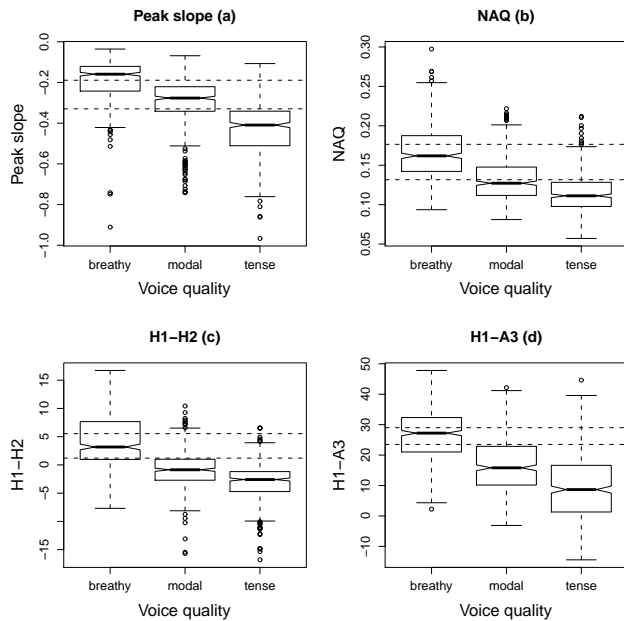


Figure 4: Distributions of the four parameters for breathy, modal and tense voice qualities for each phoneme in the **sen-tence** dataset. Thresholds obtained from the vowel dataset are marked here with dashed lines.

that the factor vowel/non-vowel did not have a significant effect on *Peak Slope* [$F = 3.4642$, $df = 1$, $p = 0.06$], *NAQ* [$F = 0.2745$, $df = 1$, $p = 0.60$], *H1-H2* [$F = 0.0116$, $df = 1$, $p = 0.91$] or *H1-A3* [$F = 0.7628$, $df = 1$, $p = 0.38$] values.

5. Discussion & conclusion

The experiments conducted in the present study demonstrate the usefulness of the proposed parameter, *Peak Slope*, at identifying regions of different voice quality without the use of inverse filtering. In line with the findings in previous studies (e.g., [6, 8]) the *NAQ* parameter performs well at differentiating between the voice qualities in the vowel dataset and shows robustness to added babble noise. However, the *Peak Slope* also displayed robustness to babble noise up to $SNR = 10$ dB and provided better differentiation of the voice qualities than *NAQ*, as well as *H1-H2* and *H1-A3*. All four parameters were heavily affected by the babble noise added at $SNR = 5$ dB suggesting that in their current state these measures are probably unsuitable for voice quality analysis in conditions of this noise level (and for this noise type). We found this trend to be true also for additive Gaussian noise (data not included in the present study).

The new parameter *Peak Slope* again performed better than the other three parameters in the analysis of running speech using thresholds set from analysis of the vowel data. This was due to the *Peak Slope* value ranges (across the three voice qualities) being more consistent between the two dataset. These findings in combination with the findings from the leave-one-speaker-out analysis in the vowel dataset suggest that the new parameter generalises well, compared to the parameters, across different speech data types. Furthermore, as the *Peak Slope* values in the sentence data were not significantly affected by the speech segment being a vowel or a non-vowel it may also be suitable for analysis of voiced consonants.

There are two striking advantages of the *Peak Slope* pa-

rameter. The first is that it is completely standalone, i.e. no other algorithms (e.g., f_0 , GCI detection, inverse filtering) are needed in order to obtain the values. This is beneficial in the case of background noise or ‘difficult’ speech segments which can hamper those algorithms and thus affect the voice quality parameter values. The second major advantage is that the *Peak Slope* parameter was developed without assumptions which would affect the decision on windowing, which can be quite complicated particularly when analysing non-modal voice quality segments. Here only a very simple rectangular window with phoneme boundaries was required. In particular the parameter may be useful in the analysis of conversational speech which would be likely to contain expressive utterances produced with a range of voice qualities. The *Peak Slope* parameter is further suited to this type of analysis as it has been shown to be both robust to noise and suited to the analysis of running speech.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07 / CE / I 1142) as part of the Centre for Next Generation Localisation (www.cngl.ie).

7. References

- [1] Ishi, C. “A method for automatic detection of vocal fry”, *IEEE Transactions on Audio, Speech and Language Processing*, 16 (1), 2008
- [2] Sturmel, N., d’Alessandro, C., Rigaud, F., “Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform”, *Proceedings of ICASSP*, 4517–4520, 2009.
- [3] Kadambe, S., Bourdreaux-Batels, G., “Application of the wavelet transform for pitch detection of speech signals”, *IEEE trans. on IT*, 32(2), 917–924, 1992.
- [4] Rahmouni, A., Bouzid, A., Ellouze, N., “Wavelet decomposition of voiced speech and mathematical morphology analysis for glottal closure instants detection”, *Proceedings of EUSIPCO*, 2002.
- [5] Van Pham, Tuan. “Wavelet Analysis For Robust Speech Processing and Applications”, Ph.D. Thesis, 2007.
- [6] Airas, M. and Alku, P., “Comparison of multiple voice source parameters in different phonation types”, *Proceedings of Interspeech 2007*, 1410–1413, 2007.
- [7] Laver, J. “The Phonetic Description of Voice Quality”, Cambridge University Press, 1980.
- [8] Alku, A. and Bäckström, T., “Normalized amplitude quotient for parameterization of the glottal flow” *J. Acoust. Soc. Am.*, 112(2):701–710, 2002.
- [9] Hanson, H., “Glottal characteristics of female speakers: acoustic correlates.” *J. Acoust. Soc. Am.*, 101(1):466–481, 1997.
- [10] Campbell, N., “Listening between the lines; a study of paralinguistic information carried by tone-of-voice” *Proc International Symposium on Tonal Aspects of Languages, TAL2004*, Beijing, China, 13–16, 2004.
- [11] Alku, P., “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering” *Speech communication*, 11:109–118, 1992.
- [12] Nelder, J. A. and Mead, R., “A simplex method for function minimization” *The computer journal*, 7(4):308–313, 1965.
- [13] Goncharoff, V., Gries, P., “An Algorithm for Accurately Marking Pitch Pulses in Speech Signals” *Proceedings of SIP*, Las Vegas, 1998.
- [14] Gobl, C., Ní Chasaide, A., Monahan, P., “Intrinsic voice source characteristics of selected consonants” *Proceedings of ICPhS*, Stockholm, 74–77, 1995.