COMPARISON OF TIME AND FREQUENCY DOMAIN MEASURES OF THE VOICE SOURCE

Christer Gobl, Irena Yanushevskaya, Andy Murphy, Ailbhe Ní Chasaide

Phonetics and Speech Laboratory, Centre for Language and Communication Studies,
Trinity College Dublin, Ireland
cegobl@tcd.ie, yanushei@tcd.ie, murpha61@tcd.ie, anichsid@tcd.ie

ABSTRACT

Voice source modulation is of fundamental importance in speech communication. Many parameters have been proposed to capture the characteristics of the voice, but it is not always clear how the different kinds of parameters relate to each other. Two widely used approaches to voice parameterisation involve either time or frequency domain measures. In this paper they are compared in the analysis of a small data set – an utterance produced with different affects. The results for the time and frequency parameterisation are correlated, and both are compared in turn to an alternative parameter, MDQ, which is based on the wavelet transform. The correlation of time and frequency measures was reasonably high for parameters that relate to the lower end of the source spectrum, but correlations for the parameters relating to the upper end of the spectrum ran contrary to expectations. MDQ measures correlated strongly with time domain measures, and reasonably strongly with frequency domain measures of the low end of the source spectrum.

Keywords: glottal parameters, voice source, time domain, frequency domain, MDQ, voice quality

1. INTRODUCTION

Voice quality carries multiple strands of information to the listener, some pertaining to the characteristics of the individual's voice, while others carry important aspects of the message's meaning [19]. The voice is the carrier of prosody, and it has been argued [21-23] that the modulation of different dimensions the voice (and not only f_0) is an essential part of linguistic prosody (e.g., prominence [28, 29, 20], accentuation [24], declination [25]) as well as crucial to paralinguistic prosody [11, 30].

Despite its fundamental importance, this area of speech communication is not well understood, and this is to a large part due to the methodological difficulties in voice analysis (for a discussion of some of these, see [7, 13, 14, 18]). One pressing concern is to find parameters that capture the important characteristics of the voice – needed not only for speech analysis, but also for speech generation, where one aspires to synthetic voices that can resemble more the nuanced prosody of human speech [8].

In the main, two rather different approaches tend to be adopted for voice source parameterisation, and these are based on either time or frequency domain measures. Time domain measures (detailed below) have the advantage that the relationship to speech production can more easily be inferred. On the other hand, the frequency domain is more readily linked to auditory perception. Being able to map between time and frequency measures would (i) extend our understanding of the source and (ii) open the way to more robust techniques for source analysis.

This paper compares time and frequency domain measures, which were obtained for a data set based on an utterance by one speaker portraying a number of different emotions. Furthermore, these measures are also compared to a rather different measure, the MDQ parameter. The MDQ parameter, is not based directly on either time or frequency measures, but draws rather on techniques used in image processing (see more below). It has been proposed as an alternative measure that can serve as an indicator of voice source differences along the tense-lax continuum. If so, this measure could be very useful in many applications as a more readily calculated proxy measure of voice quality.

2. THE VOICE SOURCE PARAMETERS

2.1. Time domain parameters

In addition to f_0, the following time domain voice source parameters were analysed (see also [13]).

R_g, the normalised glottal frequency, is a measure of the characteristic frequency of the glottal pulse (F_g), normalised to f_0. R_g mainly affects the relative amplitudes of the low end of the source spectrum.

R_k is a measure of glottal pulse symmetry, defined as the duration of the closing portion of the pulse relative to the duration of the opening portion. Thus, a lower R_k value means a more skewed pulse.

O_q, the open quotient, is a measure of the open phase of the glottal pulse as a proportion of the glottal period. O_q is determined here entirely by R_g and R_k according to $O_q = (1+R_k)/(2R_g)$. It thus excludes the return phase (captured instead by the R_a para-

meter). O_q mainly affects the amplitudes of the lower end of the source spectrum.

R_a is the normalised effective duration of the return phase, i.e. the interval for which the glottis remains open after the main excitation. R_a relates to the spectral slope: the higher the R_a value, the greater the spectral slope.

E_e, the excitation strength, is the negative amplitude of the differentiated glottal waveform at the time point of maximum change in the waveform derivative. It relates to the overall strength of the glottal excitation.

U_p, the peak glottal flow, is a measure of the maximum amplitude of the glottal flow pulse.

R_d is a global parameter that is proposed to capture some of the main features of the glottal pulse in one single measure [3, 4]. It is derived from f_0, E_e and U_p as follows: $(1/0.11)\times(f_0 \cdot U_p/E_e)$. For the similar NAQ parameter, see [1] and for other amplitude based measures, see [12].

2.2. Frequency domain parameters

The frequency domain parameters analysed were the four parameters of the frequency domain model proposed by Kreiman and colleagues [17, 6]:

H1-H2 is the relative amplitude levels of the two first harmonics of the source spectrum.

H2-H4 is the relative amplitude levels of the second and fourth harmonics of the source spectrum.

H4-2k is the relative amplitude levels of the fourth harmonic and the harmonic closest to 2 kHz.

2k-5k is the relative amplitude levels of two harmonics closest to 2 kHz and 5 kHz respectively.

2.3. Maxima Dispersion Quotient, MDQ

The Maxima Dispersion Quotient (MDQ), is a relatively new measure, which purports to capture differences in the tense-lax dimension of voice quality [15]. Based on wavelet filtering, it was found effective for edge detection in image processing. This property is exploited in the calculation of the MDQ parameter: if the glottal pulse excitation is more impulse-like (typical for tense voice) the maxima from the wavelet decomposition will appear close to the excitation. However, much less impulse-like excitation is involved in the production of lax or breathy voice, and the maxima from the outputs of the wavelet filtering will be more dispersed.

The MDQ measures are derived completely automatically: Initially, the SE-VQ [16] algorithm is used to detect the time points of the main glottal excitations. LPC is then applied to derive an estimate of the glottal source signal, which is analysed using a dyadic wavelet transform. Here we use six scaled versions of the wavelet function, which results in an octave band, zero-phase filter bank, with filter centre-frequencies from 125 Hz to 4 kHz.

For each glottal excitation detected, a search interval is defined. The locations of the six maxima are determined within this interval and the absolute durations relative to the excitation time point are measured. The mean of these durations is then divided by the fundamental period to obtain the MDQ value. For further details, see [15].

2.4. Expected trends

In the present comparison, the expected trends are that the strongest correlation will be found between those time domain parameters associated with the low end of the source spectrum (O_q and the source parameters that contribute to O_q, namely R_g, R_k) and the H1-H2 measure. Some similar, though perhaps weaker correlations might also be expected with the H2-H4 parameter.

As regards the higher end of the source spectrum, one would expect the strongest correlations to be between the time domain parameter R_a and the frequency domain parameter 2k-5k. A somewhat weaker correlation with the H4-2k parameter might also be expected.

As the MDQ value is determined by the sharpness of the glottal excitation, strong correlation would be expected with E_e, R_a and R_k, which most directly capture the shape the glottal excitation. A general positive correlation would be expected with frequency domain parameters.

3. ANALYSIS METHODS

3.1. Speech data

The speech data analysed involved part of the corpus used in [27, 14]. The recordings were of a single speaker repeating an all-voiced sentence, 'We were away a year ago', so as to portray differing affective states. These included angry, surprised, sad, bored and a neutral rendition. The different versions varied considerably in terms of their voice quality, as discussed in [27]. The dataset comprised 5 utterances totalling 649 glottal pulses. However, only data for 506 pulses are presented here: pulses for which the automatic analysis failed to return an estimate were excluded.

3.2. Voice source analysis

The time domain parameter data were obtained by carrying out manual interactive analysis using the software systems described in [10, 13]. This involved inverse filtering of each individual glottal pulse to derive an estimate of the differentiated

glottal flow signal (i.e. the effect of lip radiation was not cancelled) followed by source parameterisation using the LF model matching technique, again carried out for each individual pulse.

The VoiceSauce program [26] was used to automatically extract the frequency domain parameters H1-H2, H2-H4, H4-2k and 2k-5k.

MDQ data were obtained according to the process described in Section 2.3, using the MDQ algorithm of the GlóRí analysis system [2].

3.3. Regression and correlation analysis

Linear regression analysis and Pearson product-moment correlation (Pearson's *r*) were carried out to explore the relationship between (i) the time and frequency domain parameters and (ii) the MDQ measures and those obtained for the time and frequency domain parameters. Prior to the analysis, all parameter data were first smoothed by applying 5-pulse median filtering followed by 5-pulse moving average filtering.

4. RESULTS AND DISCUSSION

4.1. Time vs. frequency domain measures

Table 1 shows the correlations for the time and frequency domain parameters. In the first column, a bracketed (+) or (−) sign indicates the expected directionality of correlation. Fig. 1 shows the linear regression lines and R^2 values for a subset of these.

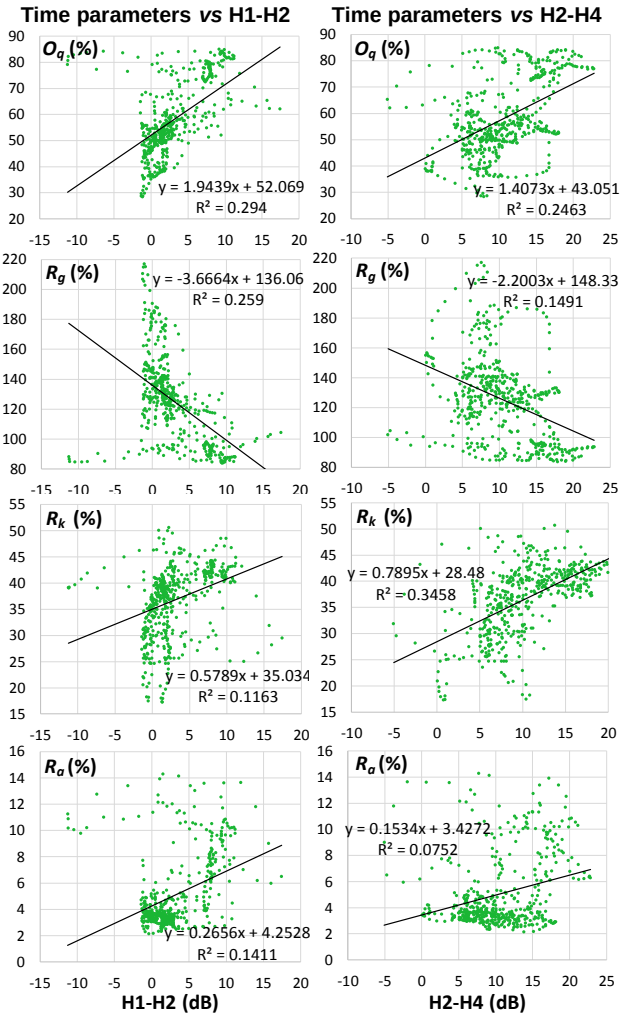
Overall, the correlations that emerged were surprisingly low, but nonetheless some clear trends emerge. At the low end of the source spectrum, one would expect H1-H2 and O_q to be the most strongly correlated (as traditionally assumed), and this was borne out. One would also expect the parameters R_g and R_k, which here together define O_q to also be correlated (negatively in the case of R_g) and this emerges, for R_g at least.

Table 1: Correlation (Pearson's *r*) between the time and frequency domain parameters. The (+) and (–) signs indicate the expected direction of the correlation. * indicates significance at p<0.01.

Parameters	H1-H2	H2-H4	H4-2k	2k-5k
O_q (+)	*0.54	*0.50	*0.29	*–0.40
R_g (–)	*–0.51	*–0.39	*–0.26	*0.34
R_k (+)	*0.34	*0.59	–0.03	–0.05
R_a (+)	*0.38	*0.46	*–0.27	*–0.54
E_e (–)	*–0.35	*–0.43	*–0.28	*0.32
U_p	0.02	*0.21	*–0.20	*0.12
R_d (+)	*0.50	*0.49	*0.27	*–0.39
$log(f_0)$	–0.12	*–0.65	–0.05	0.10

It was expected that the H2-H4 parameter might also be correlated with these same time domain parameters, and this was also borne out. The correlation of H2-H4 with R_k was in fact higher than for H1-H2 – something that was not expected, and suggests that the glottal pulse skew influences harmonics beyond the very lowest.

Figure 1: Data for time domain parameters O_q, R_g, R_k and R_a vs H1-H2 and H2-H4.



For parameters that relate to spectral tilt, particularly at the higher frequencies, the expectation was that R_a would be strongly (and positively) correlated with the 2k-5k parameter. However, a moderate negative correlation emerged. It is not entirely clear why this is so, but differences in the inverse filtering might in part explain this result. The data used here entailed considerable dynamic variation in the vocal tract filter, and this can present quite a challenge for inverse filtering. Another possible factor could be an interaction of the higher harmonics with the noise component – something that is not captured by the R_a parameter.

The global waveshape parameter R_d yielded correlations with the frequency domain parameters that are very like those for O_q. In fact, R_d and O_q were found to be very highly correlated to each other in this dataset (r = 0.94).

The correlation of f_0 with the frequency domain parameters is also shown in Table 1. One might have expected a correlation to emerge principally with the H4-2k parameter, as it compares an f_0 dependent harmonic to a (near) fixed frequency. In fact no correlation emerges here, but a relatively strong one emerges with the H2-H4 parameter. This suggests that an increasing f_0 affects the slope of this part of the source spectrum, something that is not immediately obvious from the source parameters.

4.2. Correlations with MDQ

Table 2 shows the correlation of the MDQ data with the time domain measures (left) and frequency domain measures (right). Fig. 2 also shows the regression analysis for a subset of these time and frequency domain parameters.

Table 2: Correlation (Pearson's *r*) between MDQ and the time and frequency domain parameters. (+) and (−) show expected direction of the correlation. * indicates significance at p<0.01.

MDQ			
Time parameters		*Freq. parameters*	
O_q (+)	*0.83	H1-H2	*0.62
R_g (−)	*−0.77	H2-H4	*0.44
R_k (+)	*0.67	H4-2k	0.09
R_a (+)	*0.62	2k-5k	*−0.19
E_e (−)	*−0.69		
U_p	−0.08		
R_d (+)	*0.82		

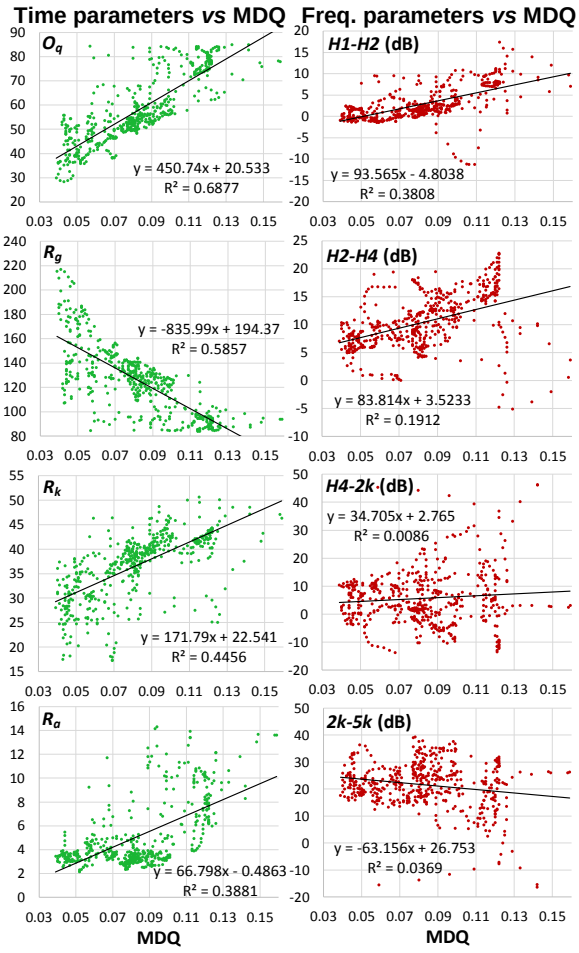
MDQ correlations with the time domain parameters are overall high (compare also analysis in [14]). Counter to expectations, the correlations for the parameters E_e (r = −0.69), R_a (r = 0.62) and R_k (r = 0.67), which were expected to be the highest (as they capture the sharpness of the excitation) were in fact somewhat lower than for the parameters which relate to the low end of the source spectrum, O_q (r = 0.83) and R_g (r = −0.77). Correlations with R_d emerged as being high, at r = 0.82.

The correlation of MDQ with the frequency domain measures is overall weaker. It is fairly high for H1-H2 (r = 0.62), weaker for H2-H4, but very low for H4-2k and for 2k-5k. We find again that the negative correlation with the 2k-5k parameter runs counter to expectations.

5. CONCLUSIONS

In this comparison of time and frequency domain measures of the voice source we found a correlation, though not a very strong one, between those parameters that pertain to the lower end of the source spectrum. Although in phonetics research, H1-H2 has traditionally been viewed as reflecting O_q variation, it is clear that other factors are involved. As shown in [9], the correlation with O_q may only hold within certain ranges of R_k values. This highlights the complexity of parameter interaction, and the need for caution in interpreting measures such as H1-H2 in production terms.

Figure 2: MDQ values compared to time (left) and frequency (right) domain measures.



The correlation of the parameters that relate to the upper end of the source spectrum ran counter to what was expected. The reasons for this anomaly are not clear, but it might be partially explained by differences in the analysis methods used. This is an area that will warrant further investigation, and modelling experiments may help complement the present approach, to elucidate the interaction of source parameters and the mapping between the time and frequency domains.

As regards the MDQ parameter, the high correlations with the time domain measures supports its use as a proxy measure of voice quality.

6. ACKNOWLEDGMENTS

This research was supported by funding from the Department of Culture, Heritage and the Gaeltacht, Government of Ireland (ABAIR project).

7. REFERENCES

[1] Alku, P., Bäckström, T., Vilkman E. 2002. Normalized amplitude quotient for parameterization of the glottal flow. *J. Acoust. Soc. Am.* 112, 701-710.

[2] Dalton, J., Kane, J., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. 2014. GlóRí - the Glottal Research Instrument. *Proc. 7th Int. Conf. on Speech Prosody*, Dublin, Ireland, 944-948.

[3] Fant, G. 1995. The LF-model revisited: transformations and frequency domain analysis. *STL-QPSR* 2-3, 119-156.

[4] Fant, G. 1997. The voice source in connected speech. *Speech Comm*. 22, 125-139.

[5] Fant, G., Liljencrants, J., Lin, Q. 1985. A four-parameter model of glottal flow. *STL-QPSR* 4, 1-13.

[6] Garellek, M., Samlan, R., Gerratt, B., Kreiman, J. 2016. Modeling the voice source in terms of spectral slopes. *J. Acoust. Soc. of Am.*139, 1404-1410.

[7] Gobl, C. 2003. The voice source in speech communication – production and perception experiments involving inverse filtering and synthesis. Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm.

[8] Gobl, C., Bennett, E., Ní Chasaide, A. 2002. Expressive synthesis: how crucial is voice quality? *Proc. of the IEEE Workshop on Speech Synthesis*, Santa Monica, California, USA.

[9] Gobl, C., Murphy, A., Yanushevskaya, I. and Ní Chasaide, A. 2018. On the relationship between glottal pulse shape and its spectrum: correlations of open quotient, pulse skew and peak flow with source harmonic amplitudes. *INTERSPEECH 2018*, Hyderabad, India, 222-226.

[10] Gobl, C., Ní Chasaide, A. 1999. Techniques for analysing the voice source. In: Hardcastle, W. J., Hewlett, N. (eds), *Coarticulation: Theory, Data and Techniques.* Cambridge: Cambridge University Press, 300-320.

[11] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189-212.

[12] Gobl, C., Ní Chasaide, A. 2003. Amplitude-based source parameters for measuring voice quality. *Proc. ISCA Tutorial and Research Workshop VOQUAL'03 on Voice Quality: Functions, Analysis and Synthesis*, Geneva, Switzerland, 151-156.

[13] Gobl, C., Ní Chasaide, A. 2010. Voice source variation and its communicative functions. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds), *The Handbook of Phonetic Sciences* (2nd edition). Oxford: Blackwell, 378-423.

[14] Gobl, C., Yanushevskaya, I., Ní Chasaide, A. 2015. The relationship between voice source parameters and the Maxima Dispersion Quotient (MDQ). *INTERSPEECH 2015*, Dresden, Germany, 2337-2341.

[15] Kane, J., Gobl, C. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 1170-1179.

[16] Kane, J., C. Gobl, C. 2013. Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Comm.* 55, 295-314.

[17] Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., Zhang, Z. 2014. Toward a unified theory of voice production and perception. *Loquens* 1, e009.

[18] Kreiman, J., Gerratt, B. R., Antoñanzas-Barroso, N. 2007. Measures of glottal source spectrum. *J. Speech and Hearing Research* 50, 595–610.

[19] Laver, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press. 3.

[20] Murphy, A., Yanushevskaya, I., Ní Chasaide, A., Gobl, C. 2018. Voice Source Contribution to Prominence Perception: Rd Implementation. *INTERSPEECH 2018*, Hyderabad, India, 217-221.

[21] Ní Chasaide, A. Gobl, C. 2004. Voice quality and f0 in prosody: towards a holistic account. *Proc. 2nd Int. Conf. on Speech Prosody*, Nara, Japan, 189-196.

[22] Ní Chasaide, A., Gobl, C. 2004. Decomposing linguistic and affective components of phonatory quality. *Proc. 8th Int. Con. on Spoken Language Processing, INTERSPEECH 2004*, Jeju Island, Korea, 901-904.

[23] Ní Chasaide, A., Yanushevskaya, I., Gobl, C. 2011. Voice source dynamics in intonation. *Proc. 17th ICPhS*, Hong Kong, China, 1470-1473.

[24] Ní Chasaide, A., Yanushevskaya, I., Kane, J., Gobl, C. 2013. The voice prominence hypothesis: the interplay of F0 and voice source features in accentuation. INTERSPEECH 2013, Lyon, France, 3527-3531.

[25] Ní Chasaide, A., Yanushevskaya, I., Gobl, C. 2015. Prosody of voice: declination, sentence mode and interaction with prominence. *Proc. 18th ICPhS*, Glasgow, UK.

[26] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. 2011. VoiceSauce: A program for voice analysis, *Proc. 17th ICPhS*, 1846-1849.

[27] Yanushevskaya, I., Gobl, C., Ní Chasaide, A. 2009. Voice parameter dynamics in portrayed emotions. *Proc. 6th Int. Workshop: Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA*, Florence, Italy, 21-24.

[28] Yanushevskaya, I., Gobl, C., Ní Chasaide, A. 2017. Cross-speaker variation in voice source correlates of focus and deaccentuation. *INTERSPEECH 2017*, Stockholm, Sweden, 1034-1038.

[29] Yanushevskaya, I., Murphy, A., Gobl, C. and Ní Chasaide, A. 2016. Perceptual salience of voice source parameters in signaling focal prominence. *INTERSPEECH 2016*, San Francisco, California, pp. 3161-3165.

[30] Yanushevskaya, I., Gobl, C., Ní Chasaide, A. 2018. Cross-language differences in how voice quality and f0 contours map to affect. *J. Acoust. Soc. of Am.* 144, 2730–2750.