

A spectral LF model based approach to voice source parameterisation

John Kane¹, Mark Kane², Christer Gobl¹

¹Phonetics Lab, Centre for language and communication studies, Trinity College, Dublin

²Muster Research Group, Computer Science and Informatics Building, UCD, Dublin.

kanejo@tcd.ie, Mark.Kane@ucdconnect.ie, cegobl@tcd.ie

Abstract

This paper presents a new method of extracting LF model based parameters using a spectral model matching approach. Strategies are described for overcoming some of the known difficulties of this type of approach, in particular high frequency noise. The new method performed well compared to a typical time based method particularly in terms of robustness against distortions introduced by the recording system and in terms of the ability of parameters extracted in this manner to differentiate three discrete voice qualities. Results from this study are very promising for the new method and offer a way of extracting a set of non-redundant spectral parameters that may be very useful in both recognition and synthesis systems.

Index Terms: LF model, voice source, parameterisation, robustness, classification.

1. Introduction

Robust extraction of parameters describing the nature of the vocal fold vibration remains desirable for many speech technology applications. An example of the successful usage of information describing the glottal waveform in parametric speech synthesis can be seen in [1]. Despite this, the potential of voice source information in synthesis and recognition systems has yet to be fully exploited.

The standard technique for obtaining an estimate of the glottal waveform from the speech waveform is to use anti-resonance filters to cancel the effect of the vocal tract. A time domain model such as the LF model [2] can then be fitted to each pulse in the glottal waveform. This allows the extraction of a set of non-redundant parameters which can also be useful in controlling the sound source in parametric speech synthesis.

A difficulty with time domain model fitting is that small amounts of noise can upset time point estimation and seriously affect the robustness of the parameters [3]. Furthermore even very high quality recordings frequently suffer phase distortion, introduced by the recording system. Some voice researchers compensate for this, e.g., [4], however, they are by far in the minority. Although phase distortion is unlikely to affect the perception of recorded signals it is a necessary requirement of time based characterisation of the glottal flow as differences in phase cause variation to the shape of the waveform. This issue provides a serious block to the consistency and reproducibility of voice analysis recorded speech segments.

One obvious way around phase non-linearity is to take measurements from the amplitude spectrum of the voice source signal. However, most spectral parameters tend to describe just the very low frequencies and have been shown to be highly correlated and hence suffer from redundancy [4]. For this reason it would be desirable to fit a model to the glottal flow spectrum. This approach has been attempted in previous studies [5] where

noise and inappropriate error functions were suggested as reasons for not producing robust parameterisation.

The current study details a spectral parameterisation method for extracting LF model-based parameters which attempts to deal with the issues highlighted in [5]. The paper describes an excerpt of results from a full comprehensive robustness testing process applied to the method.

2. Methods and Materials

2.1. The LF model and its spectrum

The LF model has been the most commonly used model of differentiated glottal flow for more than two decades. It is a four parameter model, with a further parameter implicitly set from these parameter settings to ensure that the area above and below the zero line is equal. A typical set of parameters is one amplitude parameter E_e and three time points t_e , t_p and t_a . We use a transformed set of the time parameters (the R-parameters) stated below.

$$Rg = \frac{1}{2 \times t_p \times f_0}, \quad Rk = \frac{t_e - t_p}{t_p}, \quad Ra = t_a \times f_0 \quad (1)$$

[6] gives a detailed description of the spectrum of the LF model and the spectral consequences of specific parameter variations. One observation was that Ra variation mainly affects the higher frequencies. Although this observation is correct, the LF model's area balance principle implies that increases in Ra (i.e. increasing the pulses return phase) increases the area below the zero line and results in the maximum amplitude of the pulse also being increased (see panels A and B in Fig. 1). This in turn impacts on the lower frequencies of the LF model's spectrum. In fact none of the R-parameters affect a specific contained area of the spectrum.

2.2. LF based spectral parameterisation

In [5] the author suggested that problems in fitting the LF model spectrum to the glottal flow spectrum was due, in part, to noise in the high frequencies. As Ra is the main parameter for describing the higher frequencies it will clearly be affected by this. Furthermore due to changes in area balance this will also affect the other two R-parameters (Rk and Rg), meaning high frequency noise affects the robustness of all three R-parameters. Also, when using optimisation algorithms with a standard sum of squares error function to fit the LF model spectrum to the glottal flow spectrum it will predominantly produce unacceptable fits even when inputting good initial values. The following is a description of how our Frequency Domain Matching system (shortened to FreDoM for this study) deals with these difficulties.

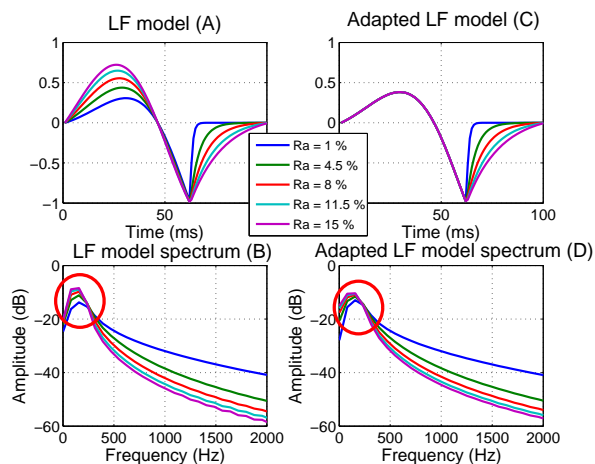


Figure 1: Panel A and B shows the effect of R_a variation (with other parameters constant) on the generated LF pulses and their spectra. Panel C and D shows the same effect but with the area balance parameter held constant.

The first step is to derive an estimation of the glottal waveform. In this study both automatic and manual methods are used. For the automatic method glottal closure instants (GCIs) are first computed using an implementation of the algorithm outlined in [7]. Speech signals are then inverse filtered using an implementation of the Iterative Adaptive Inverse Filtering (IAIF) method described in [8]. Manual inverse filtering is done using the method described in full in [9].

For the next step a codebook of a large number of LF model parameter variations and their corresponding first harmonic minus second harmonic ($H1^* - H2^*$) was developed. It should be noted at this point that unless otherwise stated all spectral measurements are made using a 256-point Hamming window centred on a GCI. When considering the LF model spectrum we are referring to a single LF pulse with identical pulses concatenated and again the window centred on a GCI. This stage measures $H1^* - H2^*$ from the glottal waveform spectrum and searches the codebook to find the closest $H1^* - H2^*$ value and extracts the corresponding LF model parameter values.

The fitting stage involves the use of two optimisation parts both of which use the well-known Nelder and Mead algorithm [10]. The first optimisation part attempts to fit an LF model spectrum to the glottal waveform spectrum using the initial values obtained from the codebook look-up. The algorithm tries to minimise the error between the first six harmonics of the two spectra by varying all the LF model parameters (with f_0 set). Differences between the model's $H1$ and $H2$ and the glottal waveform's $H1$ and $H2$ were doubled in the error function in order to prioritise their matching.

The next part uses the optimisation algorithm to vary just R_a in order to minimise the error between the two spectra in the higher frequencies. The LF model is modified for this procedure by keeping the area balance parameter constant. This essentially means using the area balance parameter from the already fitted set and allowing variation of R_a without affecting the area balance, see panels C and D in Fig. 1. This has the effect of allowing variation of the higher frequencies of the model with little impact on the already fitted lower frequencies.

2.3. Recordings

The speech segments used in this study come from the recordings used in [4] and were kindly provided to us by the authors. The original data were comprised of 11 native Finnish speakers, 6 of which were female, aged between 18 and 48 years. In this study one female speaker has been excluded. Speech was recorded with a unidirectional Sennheiser electret microphone together with a preamplifier and a digital audio recorder. Phase distortion was compensated for by obtaining the impulse response of the recording device, using a maximum length sequences (MLS) method [11], and convolving the recorded signals with the impulse response signal time reversed.

The speakers uttered eight Finnish vowels /a e i o u y æ ø/ in breathy, neutral and pressed phonation and each vowel was repeated three times. Prior to recording participants were trained to produce these voice qualities. While conducting the recordings participants were monitored to ensure consistent renditions of the particular voice qualities. After receiving the data we had a voice quality expert listen to all the speech segments who marked 23 of the total 720 vowels as not being representative of the particular voice quality. These vowels were excluded from the analysis in this study.

2.4. Evaluation

The job of evaluating voice source parameterisation methods is a very difficult task. An *a priori* knowledge of voice source parameter values does not exist so accuracy or error measurements are not possible when analysing human utterances. The approach to evaluation used in this study was used (1) to specifically assess the robustness of the extracted parameters and (2) to test their ability to differentiate 3 voice qualities.

The procedure for testing robustness is based on the method used for assessing the popular normalised amplitude quotient (NAQ) parameter [3]. 30 test signals were created by selecting an /æ/ vowel in breathy, neutral and pressed phonation modes from each of the 10 speakers in the recordings. Each signal was carefully inverse filtered using the manual fine-tuning method (described in [9]) and one pulse from each was selected. Each test signal was created by concatenating 10 identical pulses.

To test the new method's ability to cope with difficult conditions three simulations were applied to each of the test signals. The first two involved applying additive Gaussian (zero mean) noise with differing signal to noise ratio (SNR): 45 dB SNR (moderate noise level) and 30 dB SNR (high noise level). These simulations were chosen to test the robustness of parameterisation in two levels of noisy conditions. The third simulation involved convolving original test signals with the impulse response of the recording equipment (previously used to compensate for phase distortion). The purpose of this was to test both systems ability to deal with the distortions imposed on recorded signals by one particular recording setup.

In total 120 test signals were automatically parameterised using the new spectral approach described in this paper and a standard time based LF model parameterisation tool (an implementation of the algorithm described in [12] and is commonly used in other voice source analysis tools [13, 14]). Parameters values were analysed using two measures; relative change (RC) and Wilk's coefficient of variation (CV). RC was used to quantify the size of the effect of the three simulations on the parameterisation. CV was used to measure the amount of pulse-to-pulse variation of parameter values. The equations used for deriving these statistics are the same as those stated in [3]. Outliers, which were considered to be values that were over 2.5

standard deviations from the mean, were removed.

The second evaluation procedure involved testing the ability of the extracted parameters to differentiate the three voice qualities. The first part of this involving testing individual parameters. Linear regression analysis was used which involved putting the voice quality labels as the independent variables and the parameter values as the dependent variables. R^2 values were extracted which would demonstrate levels of explained variance. The second part of this evaluation stage was conducted to test the ability of the combination of extracted parameters to differentiate the voice qualities. Linear discriminant analysis (LDA) was used for this. A set was compiled with parameter values and the corresponding voice quality label for each vowel, for the new method and the time based one. Each set was randomly partitioned into 10 sets all with equal proportions of voice qualities. One of the ten sets was held as a testing set with the other 9 being used for training the classifier and classification accuracy was outputted in the form of a confusion matrix. This was repeated another 9 times each time with a different set being held out as the test set and all the confusion matrices, for each method, were summed.

3. Results

Due to space limitations this section presents only a summary of the results obtained from the evaluation procedure. A full account of the results will be reported in due course.

RC and CV scores used in the robustness testing are presented in Figures 2 (signals with moderate noise levels), 3 (signals with high noise levels) and 4 (signals convolved with the impulse response of the recording system). CV scores are consistently lower for the FreDoM method compared to the time based method. The same is seen for RC scores for signals convolved with the recording system's impulse response (Fig. 4). For signals with moderate noise levels scores were more mixed with the time based system producing lower scores for pressed signals. For breathy and neutral signals RC scores were quite similar across the two systems. This is matched quite closely in RC scores for signals with high noise levels.

R^2 scores are presented in Table 1. The FreDoM parameterisation method produced higher scores for each parameter for all the data and for gender separated analysis. Classification scores can be seen in Table 2. The new method produced higher classification scores of all three voice qualities. For neutral phonation, however, scores from both methods a rather low.

Table 1: R^2 values for each parameter, using both methods of parameterisation, with the voice qualities as the independent variables. Values are given in percentage

	All		Male		Female	
	Freq	Time	Freq	Time	Freq	Time
Rg	22.2	7.8	31.6	11.7	16.7	4.9
Rk	22.8	7.4	27.8	9.9	19.1	6.1
Ra	4.9	0.04	4.3	2.3	6.4	0.04

4. Discussion & conclusion

A new frequency domain method of extracting LF model-based parameters was presented here along with a section of the results of a larger evaluation study. The parameters extracted from the new method, shortened to FreDoM for this study, displayed comprehensively better results in terms of explained variance than a typical time domain method. R^2 scores for Ra, however,

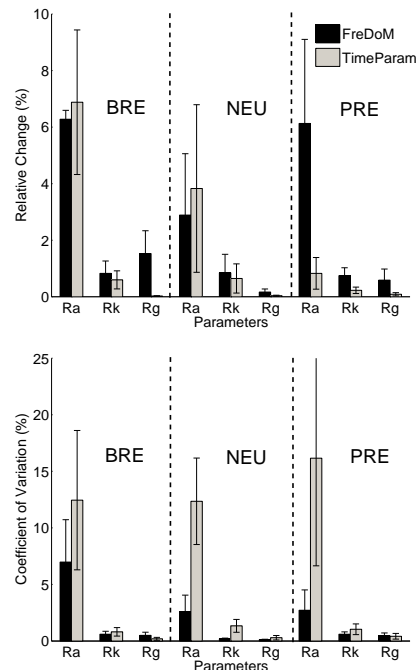


Figure 2: Top, Relative change (%) and bottom, Coefficient of variation (%) for the three R-parameteres in breathy, neutral and pressed signals with moderate noise levels (45 dB SNR) added, using the new FreDoM method and a time domain system ($n = 10$). Error is expressed as \pm SEM.

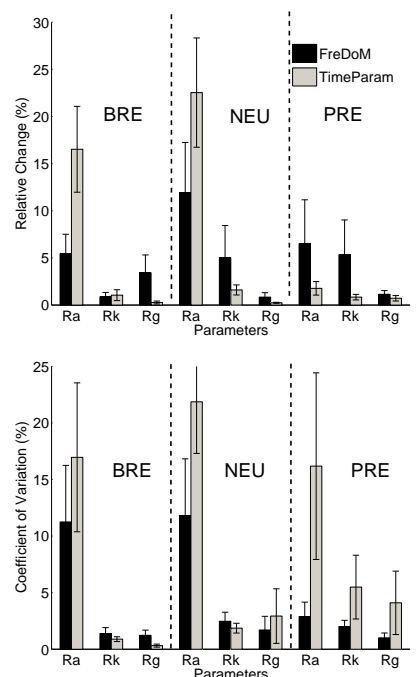


Figure 3: Top, Relative change (%) and bottom, Coefficient of variation (%) for the three R-parameteres in breathy, neutral and pressed signals with high noise levels (30 dB SNR) added, using the new FreDoM method and a time domain system ($n = 10$). Error is expressed as \pm SEM.

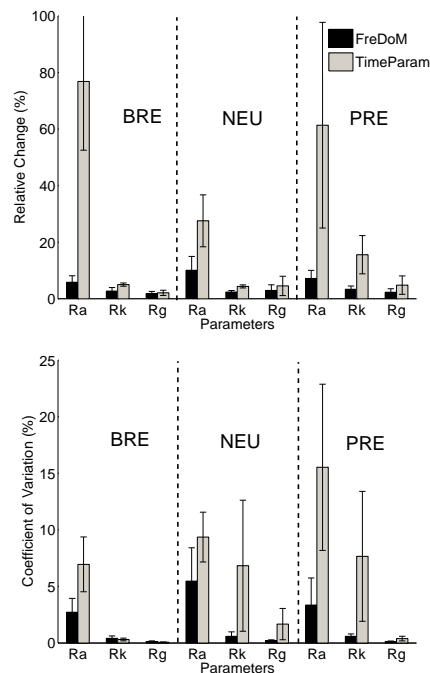


Figure 4: *Top, Relative change (%) and bottom, Coefficient of variation (%) for the three R-parameters in breathy, neutral and pressed signals convolved with the impulse response of the recording system, using the new FreDoM method and a time domain system ($n = 10$). Error is expressed as \pm SEM.*

Table 2: *Confusion matrix of classification scores (%) of the three voice qualities using the two systems.*

	Spec			Time		
	Bre	Neu	Pre	Bre	Neu	Pre
Bre	79	20	1	76	22	2
Neu	32	47	21	42	43	15
Pre	6	24	70	8	28	64

were very low which may suggest that its values vary considerably from speaker to speaker. The poor performance of the R-parameters extracted using a time based method here was also reflected in a previous study using the same dataset [4].

The classification scores suggest that the FreDoM system is generally better at classifying these three voice qualities. The low scores for neutral phonation may be due to the large variability in the organic voice quality of each of the speakers which may have come through most using their neutral phonation. If the classification was speaker dependent this score may be higher and indeed for gender specific classification (not presented here due to space limitations) scores for neutral phonation in particular were considerably higher.

The most striking result from the robustness testing was that in almost every instance the FreDoM method was less sensitive to distortions on the signal introduced by the recording system. This is likely to be in part due to avoidance of phase non-linearity issues when taking measurements from the amplitude spectrum. This result suggests that if the transfer function of the recording system has not been compensated for then the

FreDoM method may be more suitable for producing more consistent results.

The other impressions from the robustness testing were that while the FreDoM method produced clearly lower pulse-to-pulse variation with the two levels of noise, the results in terms of change in parameter values with each of the simulations (shown by relative change) were more inconsistent. Future work will require development of further methods for dealing with these noise levels.

Overall results are very promising for the new FreDoM parameterisation method and future research will involve expanding the set of voice qualities in the analysis, parameterisation of more dynamic running speech.

5. 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). We would like to warmly thank Dr. Matti Airas and Prof. Paavo Alku for providing us with the dataset used in this study and we would also like to thank Irena Yanushevskaya for her useful comments.

6. References

- [1] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering", Proceedings of Interspeech, 2008.
- [2] Fant, G., Liljencrants, J. and Lin, Q., "A four parameter model of glottal flow" STL-QPSR, 26(4):1-13, 1985.
- [3] Alku, A. and Bäckström, T., "Normalized amplitude quotient for parameterization of the glottal flow" J. Acoust. Soc. Am., 112(2):701-710, 2002.
- [4] Airas, M. and Alku, P., "Comparison of multiple voice source parameters in different phonation types", Proceedings of Interspeech 2007, 1410-1413, 2007.
- [5] Strik, H., "Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses" J. Acoust. Soc. Am., 103(5):2659-2669, 1998.
- [6] Fant, G., "The LF model revisited. Transformations and frequency domain analysis" STL-QPSR, 2-3:119-154, 1995.
- [7] Drugman, T. and Dutoit, T., "Glottal closure and opening instant detection from speech signals", Proceedings of Interspeech 2009.
- [8] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering" Speech communication, 11:109-118, 1992.
- [9] Gobl, C. and Ní Chasaide, A., "Techniques for analysing laryngeal articulation (section: B)", In Coarticulation: Theory, data and techniques (Hardcastle, W. J. and Hewlett, N., Eds.), 300-321, 1999.
- [10] Nelder, J. A. and Mead, R., "A simplex method for function minimization" The computer journal, 7(4):308-313, 1965.
- [11] Rife, D. D. and Vanderkooy, J., "Transfer-function measurement with maximum-length sequences" J Audio Eng, 37:102-113, 1989.
- [12] Strik, H., Cranen, B. and Boves, L., "Fitting a LF-model to inverse filter signals" Proceedings of EUROSPEECH-93, Berlin, 1:103-106, 1993.
- [13] Airas, M., "TKK Aparat: An environment for voice inverse filtering and parameterization" Logopedics Phoniatrics Vocology, 33:49-64, 2008.
- [14] Kreiman, J., Geratt, B. R. and Antoñanzas-Barroso, N., "Analysis and synthesis of pathological voice quality" Software Manual, 2007.