

AUTOMATIC PARAMETERISATION OF THE GLOTTAL WAVEFORM COMBINING TIME AND FREQUENCY DOMAIN MEASURES

John C. Kane, Christer Gobl

Phonetics and Speech Laboratory, Centre for Language and Communications Studies
Trinity College Dublin

Abstract: This paper describes a new technique for automatically parameterising the inverse filtered speech waveform by exploiting frequency domain measures and amplitude measures in the time domain. The technique is motivated by the difficulties posed by time domain analysis and by the consequent risks of inconsistencies on the part of both researchers and time based algorithms. The results demonstrate that the system can obtain accurate measurements on synthetic source signals. Analysis was also carried out on short utterances of three male speakers producing tense, modal and breathy voice qualities. Perception tests which involved comparing different resynthesised utterances provide evidence that the new technique is at least as good as our manual method for modal and tense voices. For breathy voice qualities, however, the system needs further development to include aspects like the noise component to provide a more breathy percept.

Index Terms: voice source, parameterisation, LF model

I. Introduction

Despite the attention voice source analysis has received researchers are still seeking to make improvements in terms of accuracy and robustness of parameterisation. Many applications require very accurate and consistent characterisation of the voice source. Recently researchers have started exploring the possibility of including a more sophisticated source model in HMM based speech synthesis in an attempt to reduce ‘buzziness’ [1, 2]. This is an example of one application which requires high accuracy as well as consistency throughout the analysis. For the purpose of analysing subtle changes in pathological voices accurate parameterisation is also required. Parameter measurement by automatic algorithms, however, tends not to be robust enough particularly across different voice qualities.

Typically the parameterisation of the voice source first requires some type of inverse filtering. This source-filter decomposition is an attempt to remove the effect of vocal tract filtering on the voice source. This is essentially the reverse of the speech production process, as described in [3]. It is done by getting an estimate of the transfer function of the speaker’s vocal tract. The speech signal is then filtered using the inverse of this transfer function which produces an estimate of the speaker’s voice source. Automatic inverse filtering systems exist (e.g., [4] or those described in [5]), but from our experience there is high risk of incomplete cancellation of formant oscillations. As the purpose of the paper is to test a parameterisation system we require good estimates of the source signal and, hence, have opted to inverse filter small amounts of speech data manually (as described in [6]).

Once a speech signal has been inverse filtered it can then be parameterised. This can be done by marking certain timepoints in glottal waveform or by fitting a model to the pulse. The most documented voice source model, and the one to be used in this study, is Liljencrants-Fant (LF) model [7]. The LF model is a four parameter model of differentiated glottal flow (see Fig. 1). The shape of the model can be described using the parameters R_a , R_k and R_g . The differentiated glottal flow is essentially the residual after inverse filtering as the effect of lip radiation has not been removed. The model is thought to be able to characterise a wide range of phonation types, however as with any model a certain amount of error will exist.

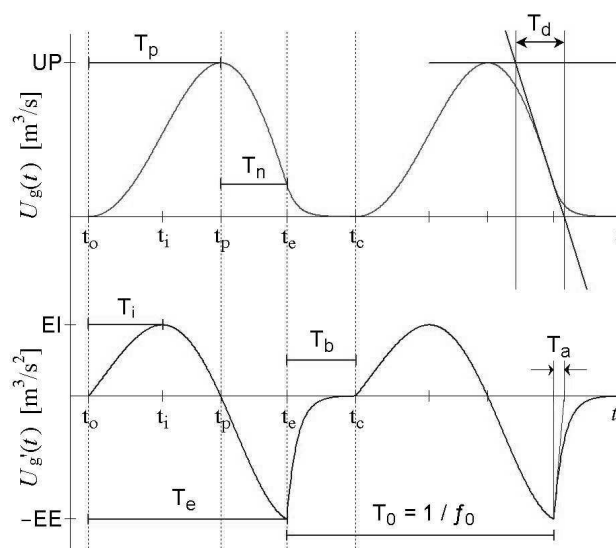


Figure 1: Examples of LF pulses (bottom) and corresponding glottal pulses (top) (taken from [8])

When parameterising the source signal most methods involve marking specific timepoints. The precise location of these timepoints can at times be quite unclear and can lead to errors as well as inconsistencies. These difficulties are heightened in the cases of non-modal phonations, e.g., in breathy voices, where a timepoint, for instance the point of glottal opening, can involve very subjective measuring.

A further difficulty with analysing and synthesising breathy voice is that the source signal contains both a periodic voice component and an aperiodic noise component [9]. The LF model parameters are used to characterise only the periodic as-

pect of the voice source. If the signal is not decomposed into periodic and noise components parameter measurements may also include the influence of the aspiration noise and, hence, may not effectively characterise the periodic component. In this study no noise analysis is carried out. We wish to include a noise analysis and synthesis system, perhaps similar to that in [10], in future algorithms.

Frequency domain analysis is thought to have a better mapping to the perception of speech than time domain analysis. In the time domain even minor errors in model fitting can have major perceptual effects. The power spectrum also bypasses any phase distortions which can upset time domain parameterisation. A complete frequency domain approach would, hence, lessen the need for high fidelity recordings which preserve phase linearity, and would allow for the analysis of a far greater range of speech data. This is a current direction of our research and work on a full frequency domain parameterisation system is underway. For the present study only the return phase parameter R_a will be measured from the source spectrum.

The remaining source parameters can be calculated from amplitude based measures in the time domain. Such measures are said to be more robust than marking time instances, especially in automatic systems [11]. It is hoped that this novel combination of frequency domain measures and time based amplitude measures can avoid some of the pitfalls of purely time point measurements and provide a robust and automatic analysis of the inverse filtered signal.

II. Method

This section outlines the different methods applied in our system in order to arrive at a full set of LF model parameters. As the study is mainly concerned with parameterisation we have analysed small amounts of speech data that have been carefully inverse filtered.

A. Inverse Filtering

We have opted for a manual inverse filtering approach for this study, as in [6]. The software first uses a linear predictive coding (LPC) method for estimating formant frequencies and bandwidths. The user then fine-tunes formant frequencies and bandwidths manually by utilising time and frequency domain displays to ensure complete formant cancellation. This fine-tuning is done for each pulse and the final output is an estimation of the differentiated glottal waveform.

B. Amplitude-based measurements

The first amplitude measurement is the negative peak of the differentiated glottal waveform, E_e . This parameter is simply measured by the new automatic system as the maximum negative amplitude of each glottal pulse. The next parameter, which is again straightforward to measure, is the peak positive amplitude of the differentiated glottal pulse, E_i , see Fig. 1 (bottom).

The measurement of the maximum amplitude of the undifferentiated flow is complicated by the occurrence of zero line drift. The LF model is designed to have equal area above and below the zero axis which means when you integrate, the signal sits neatly on the zero axis, see Fig. 1, top. Real speech, however, does not maintain this exact property and as a result the integrated waveform drifts off the zero axis, see Fig. 2 (top).

To adjust for this our system marks the major negative points for each pulse. A line is drawn from the origin then to each of the negative peaks, see the dashed line in Fig. 2 (top). Then at

each sampling point the distance between the dashed line and the zero axis is added to the signal at that sampling point which results in the signal being lifted onto the zero axis, see Fig. 2 (bottom). The system can now easily measure the maximum amplitude of each pulse, our U_p value.

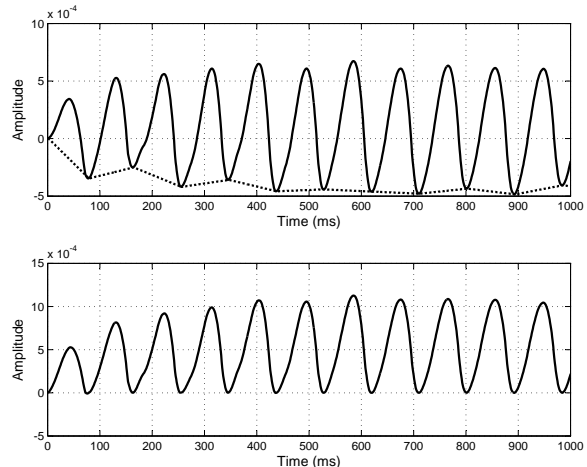


Figure 2: *The integrated source waveform (top) and the same source waveform adjusted for zero-drift (bottom)*

C. Calculating R_k and R_g from amplitude measures

Aside from f_0 and E_e the LF model can be described using three further shaping parameters. We use the parameters R_a , R_k and R_g which can be calculated from time instance measurements, as in equation 1. The positions of these time instances can be seen in Fig. 1 (bottom).

$$R_a = \frac{T_a}{T_0} \quad R_k = \frac{T_n}{T_p} \quad R_g = \frac{T_0}{2T_p} \quad (1)$$

The parameters R_k and R_g can also be estimated using our three amplitude measures (E_e , E_i and U_p). [12] gives a detailed description of how voice source parameters can be arrived at using amplitude measurements. Equation 2 shows how one can obtain an amplitude representation for R_k and R_g , as described in [12].

$$R_{ka} = \left(\frac{2}{\pi}\right) \left(\frac{E_i}{E_e}\right) \quad R_{ga} = \frac{\left(\frac{1}{\pi}\right)\left(\frac{E_i}{U_p}\right)}{f_0} \quad (2)$$

D. Frequency domain analysis

With R_k and R_g already estimated one parameter remains to configure the LF model. The parameter R_a characterises the return phase of the source waveform and it is, perhaps, considered to be the most important LF parameter [13]. However, getting accurate estimations of R_a is thought to be a challenging task [14]. Our approach derives a value for R_a from the frequency domain and the process is presented graphically in Fig. 3.

We define a set of possible R_a values, e.g., 1% to 20% (the values here refer to the percentage the return phase is of the pulse duration). The lowest is the minimum possible R_a value and the highest is the maximum possible R_a value and 50 linearly spaced values in between. The system takes a section of one pulse length from the signal and gets the spectrum (the dark

line in Fig. 3). We then generate 50 LF pulses for each of the R_a values but with all other parameter values remaining fixed. Spectra are then taken of each of the pulses (the grey lines in Fig. 3). The system then uses a Euclidean distance measure to choose the LF configuration with the closest match to the source signal. The R_a value used in this configuration is chosen as the optimal value. The system then proceeds to the next pulse and the process is repeated. This continues until the end of the signal is reached.

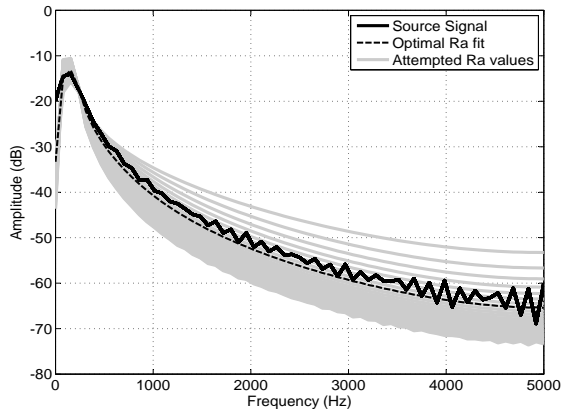


Figure 3: *Spectrum of a single pulse from the voice source signal (dark line). The grey lines show the range of LF configurations produced and the dashed line shows closest matching configuration.*

III. Evaluation

The evaluation process for the system is two-pronged. The first testing stage is on synthetic source pulses. The reason for this is that the correct source parameter values are known and the actual size of the error can be measured. The second stage involves analysing real source signals and for this the correct parameter values are not known. To provide an evaluation for real speech we have opted to apply the current system and the manual parameterisation method, as described in [6]. The extracted parameters for each method were then used to make resynthesised utterances. The resynthesised utterances were then used in perception tests where participants chose the stimulus which sounded most like the original speech utterance. This preliminary evaluation should provide evidence for whether this system provides good parameterisation for real speech. A more rigorous evaluation is required and is planned for future work.

A. Recordings

Three male English speakers were recorded producing an [a] vowel in tense, modal and breathy phonation modes. These categories of voice quality are those as described by [15] and participants listened to samples of each phonation type and practised them several times before recording. Each utterance was around half a second long and was recorded using a Pearl CC30 condenser microphone in a semi-anechoic room. Speech segments were digitised at 44.1 kHz, then downsampled to 10 kHz and high pass filtered at 40 Hz. The choice of microphone and filter ensured phase linearity was maintained. Using the method described in [6] utterances were manually inverse filtered to minimise errors.

B. Parameterisation

Each utterance was parameterised both using the new system and a manual parameterisation method, described in [6]. The manual method involves the user manually fitting the LF model to each source pulse by varying one amplitude based and four time based markers in an effort to achieve the optimal match. Although the fitting is done in the time domain, a frequency domain display is also available to the user to facilitate a more thorough fit.

C. Synthetic source analysis

Nine synthetic source signals were generated using static parameter settings. The signals were made by constructing an LF model with particular parameter settings and concatenating 10 identical pulses. In every signal E_e was set to 1. The values of R_a , R_k , R_g and f_0 were chosen so as to have a range of source signals corresponding to modal, tense and breathy voice qualities. Previous analysis of these voice qualities aided the choice of value for the above parameters. The nine signals were analysed using only the new parameterisation system.

D. Perception Tests

Perception tests were used as a method of comparing the parameterisation of the new automatic method with the manual method for real speech. 18 volunteers participated in a two-part perception test in a quiet room using high quality loudspeakers.

In test one participants were presented with 45 groups of three stimuli. In each group the first two stimuli were resynthesised versions of the original utterance, one from parameters obtained from the automatic method and one from the manual method. Resynthesis was carried out using a cascade formant synthesiser. The two synthetic signals in each group used the same f_0 , EE and formant frequencies and bandwidths. They differed only in the R_a , R_k and R_g parameters which were extracted from the two methods. The third stimulus was the original speech sound and participants had to choose which of the synthesised sounds they deemed closest to the original. The assumption here being that a closer sounding resynthesis demonstrates more accurate voice source parameterisation.

Part 2 of the test was a standard ABX discrimination task where the participants were presented with the two different resynthesised utterances and then a third sound which was a randomly chosen one of the previous two. Participants had to choose which sound had been repeated. Again there were 45 groups of stimuli. This test was chosen to demonstrate whether the differences in the parameter measurements by both methods produced two differentiable sounds.

In both parts of the perception test the order of the resynthesised stimuli in each group, as well as the order of the 45 groups, were randomised.

IV. Results and discussion

Table 1 summarises the testing of the parameterisation system on the nine synthetic source signals. The mean and standard deviation of the error (i.e. the difference between the actual source parameter values and those extracted by the system) as well as the range of values used are presented. Encouragingly, the error size for all three parameters is reasonably low. This is evidence that the amplitude based representations of R_k and R_g can indeed provide good estimates of those parameters for a wide range of settings. These results also demonstrate that the

novel method of estimating R_a is effective, at least for synthetic source pulses.

Table 1: Summary of parameterisation error for R_a , R_k and R_g . The range of values used to generate the signals and the mean and standard deviation of the differences between the extracted values and the actual parameter values are shown

| | R_a | R_k | R_g |
|------------------------|---------|---------|----------|
| Range | 1.8%-7% | 28%-39% | 79%-137% |
| Mean Error | 0.9% | 3.9% | 6.1% |
| St Dev of Error | 0.7% | 3.1% | 0.4% |

The results of the perception tests are presented in Table 2. Test 1, the first row, shows the percentage of instances participants chose the automatic method to be closer to the original (i.e. over 50% shows preference for the automatic method while under 50% shows preference for the manual method). Overall, at 50.3%, we can see that participants showed no preference for either method in terms of closeness to the original. The second row contains the results from test 2 which show the percentage of instances participants correctly identified the repeated synthesised stimuli.

Table 2: Test 1: the percentage of instances participants believed the synthesised stimuli from the automatic method to sound closer than the manual method to the original speech utterance. Test 2: the percentage of correctly identified synthesised stimuli

| Test | Modal | Tense | Breathy | Overall |
|----------|-------|-------|---------|---------|
| 1 | 49.8% | 61.2% | 40% | 50.3% |
| 2 | 54.5% | 66.7% | 72.9% | 65.1% |

For modal voice qualities 54.5% for test 2 suggests that participants were largely unable to discriminate between the two resynthesised utterances. Test 1 results, 49.8%, show that participants believed neither synthesised utterances to be closer to the original. For tense voice qualities participants slightly favoured resynthesised versions which used the new system's parameter values (61.2% of participants showing preference for the automatic method) and they were reasonably able to discriminate the two synthesised sounds.

Breathy voice qualities, as expected, proved more difficult than the other voice qualities in both the inverse filtering and parameterisation stages. It was found that participants showed slight preference for the manual method, with 40% stating that they preferred the automatic method. Participants also demonstrated a reasonable ability to differentiate the two sounds, at 72.9%. It should be noted that the synthesised versions of breathy voice qualities overall were of poorer quality than the modal and tense versions. This suggests that the LF model alone does not provide enough source information to convey a breathy percept.

V. Conclusion

Overall evidence from the evaluation appears to be encouraging for the new system described here. Analysis of synthetic signals confirms that R_k and R_g can be estimated with good accuracy from amplitude measurements alone. The analysis also

demonstrates the effectiveness of the new method of R_a estimation in the frequency domain. We hope to include this method in our forthcoming all-frequency domain parameterisation system.

Results from the perception tests suggest that the automatic system is at least as effective as the manual method for modal to tense voice qualities. This inference, however, comes solely from the fact that resynthesised utterances were judged to sound closer to the original speech utterances and does not rigorously demonstrate accuracy of parameterisation.

For breathy voice qualities we hope that by implementing analysis and synthesis of the noise component, perhaps similar to that in [10], we can provide a more perceptually breathy sound. The issue of finding further methods of demonstrating accuracy of parameterisation for breathy voice qualities and for voiced speech in general requires further attention.

VI. 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 also like to thank Irena Yanushevskaya for her useful comments and help with the inverse filtering.

VII. References

- [1] Cabral, J. P., Renals, S., Richmond, K. and Yamagishi, J., "Glottal spectral separation for parametric speech synthesis", Proc. of Interspeech, 2008.
- [2] Raitio, T., "Hidden markov model based Finnish text-to-speech system utilizing glottal inverse filtering", Masters Thesis, 2008.
- [3] Fant, G., *The acoustic theory of speech production*, Mouton, Hauge (2nd Edition 1970).
- [4] Airas, M., "Methods and studies of laryngeal voice quality analysis in speech production", Ph.D. Thesis, 2008.
- [5] Pfitzinger, H. R., "Influence of differences between inverse filtering techniques on the residual signal of speech", DAGA München, 2005.
- [6] Gobl, C. and Ní Chasaide, A., "Techniques for analysing the voice source" in *Coarticulation: Theory, Data and Techniques* edited by Hardcastle, W. and Hewlett, N., pp 300-320, Cambridge University Press, 1999.
- [7] Fant, G. and Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 26(4):1-13, 1985.
- [8] Gobl, C., "The voice source in speech communication - production and perception experiments involving inverse filtering and synthesis.", Ph.D. thesis, KTH, 2003.
- [9] Mehta, D. and Quatieri, T., "Synthesis, analysis, and pitch modification of the breathy vowel", IEEE Workshop on applications of signal processing to audio and acoustics, 2005.
- [10] Gobl, C., "Modelling aspiration noise during phonation using the LF voice source model", Interspeech, Pittsburgh, 2006.
- [11] Alku, P., Bäckström, T. and Vilkmán, E., "Normalized amplitude quotient for parameterization of the glottal flow", Journal of the acoustical society of America, 112, pp. 701-710, 2002.
- [12] Gobl, C. and Ní Chasaide, A., "Amplitude-based source parameters for measuring voice quality", VOQUAL'03, 2003.
- [13] Fant, G. and Gustafson, K., "LF-frequency domain analysis", STL-QPSR, 2, 1996.
- [14] Strik, H., "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses", Journal of the Acoustical Society of America, 103(5), pp. 2659-2669, 1998.
- [15] Laver, J., *The phonetic description of voice quality*, Cambridge University Press, 1980.