FITTING A LF-MODEL TO INVERSE FILTER SIGNALS

Helmer Strik, Bert Cranen and Louis Boves

University of Nijmegen, Dept. of Language and Speech, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail: STRIK@LETT.KUN.NL

ABSTRACT

A method is presented for the automatic extraction of voice source parameters from speech. An automatic inverse filtering algorithm is used to obtain an estimate of the glottal flow signal. Subsequently, an LF-model [1] is fitted to the glottal flow signal. In the current article we will focus on the improvement of the automatic fit procedure.

To keep track of the performance of the fit procedure, a quantitative evaluation criterion is preferred. It is difficult to obtain such a criterion for natural speech. Therefore, we propose an evaluation method in which synthetic speech is used. We also conducted qualitative tests for disturbances that are often found in natural speech, i.e. source-filter interaction.

Keywords: inverse filtering, LF-model, fit, evaluation

1. INTRODUCTION

Now that a number of techniques for glottal inverse filtering (IF) are available, there is an increasing need for reliable and robust techniques that allow one to parameterize glottal flow signals. Parametric representations facilitate the interpretation and the practical use of the analysis results enormously. In the past, a large number of possible parametric representations of glottal flow have been proposed. Among them, the Liljencrants-Fant (LF) model seems to be very popular. In our work we have chosen the LF-model to parameterize glottal flow.

Waveform parameterization is an intricate process, the complexities and implications of which are not always fully appreciated. In terms of 'accuracy' the result of a parametric fit depends a.o. on the optimization algorithm and the cost function to be minimized. Moreover, the results are heavily influenced by the presence of random or systematic error in the signal to be fitted. Last but not least, if the model used to fit the signal differs substantially from the process that generated the signal, the concept of accuracy of the fit looses most of its meaning.

The purpose of this paper is to investigate to what extent formal automatic fitting of LF-parameters to inverse filtered signals is at all feasible. In doing that, we take the IF output for granted (see section 2.2). The performance of a fit procedure can be investigated in several ways. One way would be to study the residual error that remains after the optimal fit has been found. However, this error does not give information about the accuracy of individual model parameters in non-ideal conditions. The latter aspect can only be investigated using synthetic signals that are generated with a process that fits exactly to the parametric model. In this research we have opted for the second approach. We are convinced that this quantitative approach has allowed us to come to grips with at least some aspects of the problems, that would have remained inaccessible had we used subjective, visual evaluations of the fits.

2. EVALUATION

2.1. Evaluation material

We used two types of test signals. The first test signal was generated by a serial pole-zero synthesizer [3] that uses an LF-source [1]. We synthesized a single CVCV... utterance, where V varied over all 12 Dutch monophthong vowels and C over all liquids and glides. Both source and filter change continuously during the utterance. The test utterance contains 230 pitch periods. A sampling frequency of 10 kHz and an amplitude resolution of approximately 12 bits were used.

A second test signal was generated with a synthesizer that accounts for non-linear source-tract interaction. The area functions that were used during synthesis were derived from the LF-pulses of the first test signal, and the vocal tract was kept constant (vowel /a/). Inverse filtering of this test signal will yield a glottal waveform signal that contains an interaction ripple. We used this test to investigate the sensitivity of the fit procedure to a disturbance that is often found in natural speech.

2.2. Evaluation criterion

Both test signals were inverse filtered automatically with the technique described in [2]: an FB-tracker based on a dynamic programming algorithm is used to determine optimal values for formant frequencies and bandwidths, that are subsequently used to create the inverse filter.

For evaluation of the first type of test signals we propose to use the average percentual error (APE) as an evaluation criterion. The APE for each parameter P is calculated in the following way. For each pitch period we compute the percentual error (PE) by comparing the estimated parameter with the input parameter (PE = $|P_{est}$ -

$P_{inp}|/P_{inp}$). The APE is then obtained by averaging over all pitch periods. The reasons for using this evaluation criterion are:

- This evaluation criterion provides a measure of the difference between estimated and input values.

- Averaging the values of each parameter over all pitch periods has the desirable effect of providing a single evaluation criterion for each parameter.

- Taking the absolute value is necessary as positive and negative values could otherwise cancel each other out when the average is calculated. In this case the average error can be small even if the errors for individual pitch periods are large.

- Given the considerable amount of variation in magnitude, both within and between parameters, it is better to use a relative error. By weighting the errors relative to the magnitude of the parameters it is possible to average the errors within each parameter and to make meaningful comparisons between parameters.

Instead of the mean we could have used the median to calculate the evaluation criterion. In this case the effect of the outliers would have been reduced. However, as outliers form part of the results of the method, we did want to take them into account. Therefore, we opted for the mean. If possible, these outliers should be detected automatically. This information could then be used to obtain better estimates of the parameters, thus leading to smaller APE values.

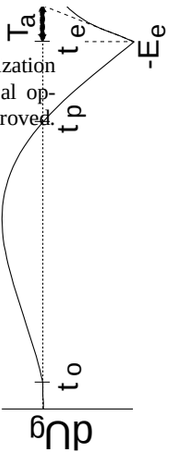
For the second type of test signals APE cannot be used to evaluate the accuracy of the LF-parameter estimation procedure. This is due to the fact that the relation between the original LF-pulses and the area functions used in synthesis is non-linear to begin with. Thus, we must resort to a less formal evaluation. In this case we studied the jitter in the parameters resulting from the various stages of the LF-model fit.

3. AUTOMATIC FIT PROCEDURE

Voice source parameters could be derived directly from the dU_g signal by means of heuristic rules. However, as dU_g usually is a noisy signal, direct estimates might yield unreliable values. Fitting a voice source model to the data seems to be a more robust method of obtaining voice source parameters.

We opted for a pitch synchronous fit procedure. For each pitch period the aim is to find an LF-pulse that optimally resembles the glottal pulse, i.e. the goal is to minimize an error. The optimization problem for the LF-model requires the use of a non-linear algorithm. To start such an algorithm, initial estimates of the parameters are needed.

3.1. Initial estimates



Initial estimates are very important for optimization procedures. The probability of finding the global optimum is enlarged if the initial estimates are improved. Thus, it will pay to optimize this stage of the fit procedure. In method 2 of [2] the initial estimates for a given pitch period were the resulting parameters of the previous pitch period. The idea behind this was that the LF-parameters change gradually; thus, the optimal parameters obtained for period N should be good starting values for period N+1. However, it appeared that this strategy often caused the optimizer to cling to wrong estimates. Therefore, we decided to make independent estimates of the initial values for all pitch periods.

Glottal flow signals often contain noise and ripple. In order to make the initial estimate more robust, the signal is first low-pass filtered. Low-pass filtering does change the shape of the pulse, and thus the value of the calculated parameters. The amount of the change is determined a priori, and an appropriate correction is made. The low-pass filter that gave the best results was a convolution with a 7-point Blackman window, which is guaranteed to have a ripple-free impulse response.

Initial estimates for the parameters t_o, t_p, t_e, and E_e (see Figure 1) are derived from the low-pass filtered time signal. t_e and E_e are place and magnitude respectively of the minimum of dU_g. This is the first event searched for each period. From t_e we search to the left (towards the beginning of the flow pulse) for t_p, which is taken as the first zero-crossing. t_o is the time-point to the left of t_p where the flow pulse first becomes lower than a small threshold.

As it has been claimed that it is difficult to get a good estimate of T_a in the time domain, we tried to derive an initial estimate in the frequency domain. Several methods and predictors were tested, of which the following method gave the best results. For the given pitch period only the signal of return phase is used, i.e. the signal from sample t_e upto sample t_c. All values of the samples of the

Figure 1. The LF-model and its parameters.

return phase are first divided by E_e (normalization). Next, an FFT algorithm is used to calculate the spectrum of the return phase. The magnitude of the maximum of this spectrum proved to be a good predictor of T_a. Probably, the explanation is that this magnitude closely resembles the DC-component of the return phase. An increase in T_a would result in a larger DC-component. Normalization (i.e. division by E_e) is necessary, because otherwise an increase in E_e would also lead to an increase in the DC-component.

The error in the initial estimate of t_e is small. For 218 of the 230 pitch periods (95%) the estimated value was correct. Therefore, t_e was kept constant in the optimization procedure. The errors that occur in estimating t_e are mainly due to formant ripple that had not been removed from the IF result.

The LF-parameters T_p and T_e are calculated from t_o, t_p, and t_e: $T_p = t_p - t_o$ and $T_e = t_e - t_o$. The APE's of the initial estimates are given in the second row of Table I.

3.2. Error criterion

One of the most important decisions that must be made when using non-linear optimization is the selection of an error criterion or cost function. After extensive experimentation we have settled for the RMS-error of a complete pulse in the time domain, i.e. a pulse starting at t_o and ending at t_c (see Fig. 1).

Although it might seem that a combined time-frequency domain error should be preferred, such a combination does not work well in practice. The problem is that the time domain part of the error changes independently of the frequency domain error contribution, and neither is bounded in any predictable way. Consequently, it is not possible to find reliable weights to combine the two contributions. In almost all cases where the frequency domain error dominates the problem the performance was bad. This explains our preference for a time domain error.

3.3. Low-pass filtering

Most dU_g signals contain high-frequency disturbances. Thus, it seems reasonable to fit LF-parameters to a low-pass filtered version of the flow. However, a detailed analysis of the results obtained in [2] revealed that the LF-parameters are extremely sensitive to low pass filtering. This is especially true for T_a, and to a lesser extent for T_p. Actually, mis-estimates due to LPF are the main cause of the relatively high APE for the best method in [2] reported as the first row in Table I. In order to cope with this problem, we filter both the flow pulse and the LF-pulse. Filtering is implemented by convolving both flow and LF-pulse with an 11 point Blackman window.

3.4. Optimization algorithms

We have experimented with a number of different non-linear optimization procedures. Best results are obtained using a three-step procedure as follows:

1. Initial estimates for the 5 parameters: E_e, t_o, t_p, t_e, T_a using the approach described above

2. Optimization using the simplex search method of Nelder and Mead [4]. This algorithm is said to be relatively insensitive to inaccuracies in the initial estimates.

3. Optimization with a steepest descent algorithm, using a subroutine described in [5]. Steepest descent is believed to converge to the optimal solution in a small number of iterations, provided the point of departure is not too far off.

Obviously, we count on the simplex algorithm for correcting occasional gross errors in the initial estimates, to enhance the chances that the steepest descent algorithm will find the global minimum.

Table I. APE of E_e, T_p, T_e, and T_a for method 2 of [2], for the 3 stages of the current method, and for a fit with the current method on $dU_{g,inp}$.

	E_e	T_p	T_e	T_a
[2]	6.9	12.2	7.2	33.2
initial estimation	6.8	5.8	4.5	18.0
simplex	6.5	4.7	3.7	15.4
steepest descent	5.8	4.3	3.4	13.9
fit on $dU_{g,inp}$	1.8	1.9	1.6	3.8

4. RESULTS

If IF were perfect, the result would be the glottal flow signal that is used as input during synthesis ($dU_{g,inp}$). The fit procedure was tested on $dU_{g,inp}$. The errors of the resulting parameters are listed in the last row of Table I. This is the baseline performance of the current procedure. It is clear that the error in the fit of noise-free dU_g signals remains at a non-negligible level. There are several possible explanations for this observation. First of all, T_e was tied to the sample position where dU_g is minimal during the initial estimation. But it is possible that the correct T_e falls between two samples [6, 7]. Quantisation noise can also explain part of the error. Increasing the sampling frequency and the number of bits that were used to code each sample would probably reduce the error.

The APE values obtained after each step of the fit procedure are given in Table I. Despite the fact that the initial estimates are already quite good, the non-linear

optimizers succeed in reducing the average error. However, the improvements in the APE after applying simplex and steepest descent algorithms are small. This raises the question whether the non-linear optimization routines are at all worth the huge amounts of CPU power they consume. To obtain a better insight in this matter we studied the APE for each of the parameters in each individual period at the end of the three stages in the fit procedure. It appears that the quality of the initial estimates is the single most important factor. If the initial estimate contains gross errors the optimization routines still succeed in improving the final results, but not always to such an extent that the final results are in the same range that is obtained for the periods where the initial estimates are 'correct'. Obviously, the simplex algorithm does not completely succeed in correcting these gross errors in the initial estimates.

It is of some interest to observe that the errors for the four parameters are correlated. High correlations are found between the errors of T_p and T_e, and between the errors of E_e and T_a. This tendency can, at least partly, be explained by the properties of the LF-model and the dU_g signals.

In most speakers the pulse onset is very gradual. Consequently, at pulse onset dU_g is close to zero, and even the second derivative of dU_g is relatively small. In such a situation the initial estimate of t_o is extremely sensitive to disturbances in the signal. The error in the estimate of t_o will contribute to the errors of both T_p and T_e. This probably explains the high correlation in the errors of T_p and T_e. Most likely, the correct estimation of opening time t_o will be a problem, no matter what voice source model is used.

The high correlation in the errors of E_e and T_a is probably due to the mutual interdependence of these parameters in the LF-model. Although a non-orthogonal parameter set may not harm in synthesis -the goal that virtually all proposers of parametric flow descriptors had in mind- it may be very harmful when the model waveform is used in a non-linear optimization procedure. An estimation of E_e that is too small will lead to an estimation of T_a that is too large, and an estimation of E_e that is too large will cause the estimation of T_a to be too small. Of course, a pulse descriptor with orthogonal parameters might be less vulnerable here.

5. SOURCE-FILTER INTERACTION

One very likely violation of the basic hypothesis that natural speech can be accurately modelled by a linear filter that is excited by an LF-signal is the often observed non-linear source-tract interaction. For the speech production system to be linear, the glottal impedance must be much larger than the acoustic load impedance. Transglottal pressure is the sum of the DC-pressure just below the glottis, and the AC-pressure variations in the trachea and the vocal tract due to the excitation of sub- and supraglottal resonances. More often than not, the AC-component is nonnegligible with respect to the DC-component. If that happens, and the glottal impedance is not high enough, the flow through the glottis contains a formant-like ripple that cannot always be removed by an inverse filter built on the assumption that a maximum of five vocal tract resonances must be cancelled. Consequently, the resulting IF signal will contain formant-like ripple, that is not accounted for by the LF-model. Such systematic discrepancies between the real-world signal and the model may upset the estimation procedure.

The jitter of the parameters resulting from the various stages of the LF-model fit were expected to be small, because the parameters in the underlying LF-signal varied only slowly and continuously. Although there were occasional gross errors in the initial estimates, overall the jitter was well within bounds. Despite the fact that the RMS-error kept decreasing during the simplex and steepest descent optimization, the jitter in the resulting parameter tracks appeared to increase. Apparently, the optimization procedure succeeds in reducing the RMS-error between the flow pulse and the fitted LF-model, but it can do so using quite different sets of parameters for adjacent periods. Note that this is not just a natural consequence of our decision to make independent initial estimates for each period. Although the optimization procedure appeared to have great difficulty in converging to the 'true' values of the LF-parameters in the linear synthesis case, it had less difficulty in diverging from the putative 'true' parameters -while still improving the RMS-error- when the input signal did not closely resemble an ideal LF-pulse. We are forced to conclude that uniformly weighted RMS-error in the time domain does not capture all aspects of waveform similarity relevant in this context.

6. CONCLUSIONS

Our experiments show that good approximations of LF-parameters can be obtained as long as the waveshape of inverse filtered glottal flow signals does not deviate too much from the ideal LF-template.

Also, it has become clear that even a small amount of low-pass filtering already affects the estimates of the LF-parameters considerably. Therefore, fitting of glottal flow with any parametric model should not be attempted without applying the same filter to the model waveform also. In order to unnecessarily complicate the estimation procedure one should use a low-pass filter with a ripple-free impulse response.

A lot of detailed research is still necessary to determine the sensitivity of the fit procedure to specific errors in inverse filtering.

Another field where additional research is badly needed is the formulation of cost functions for the estimation procedure. The conclusion seems inevitable that uniformly weighted RMS-error in the time domain -or in the frequency domain, for that matter- does not adequately conform to the intuitive ideas cherished by most speech researchers about important and less important aspects of the similarity of glottal flow waveforms.

REFERENCES

[1] Fant, G.; Liljencrants, J.; and Lin, Q. (1985) A four parameter model of glottal flow. STL-QPSR, Vol. 4, pp. 1-13.

[2] Strik, H.; Jansen, J.; and Boves, L. (1992) Comparing methods for automatic extraction of voice source parameters from continuous speech. Proceedings ICSLP-92, Banff, Vol. 1, pp. 121-124.

[3] Boves, L.; Kerkhoff, J.; and Loman, H. (1987) A new synthesis model for an allophone based text-to-speech system. Proceedings of the European Conference on Speech Technology, J. Laver and M.A. Jack (eds.), Edinburgh, Vol. II, pp. 385-388.

[4] Nelder, J.A., and Mead, R. (1964) A Simplex Method for Function Minimization. Computer Journal, Vol. 7, pp. 308-313.

[5] Daniels, R.W. (1978) An introduction to numerical methods and optimization techniques (chapter 9). Elsevier North-Holland, Inc., New York.

[6] Fujisaki, H.; and Ljunqvist, M. (1986) Proposal and evaluation of models for the glottal source waveform. Proceedings of the ICASSP, Tokyo, pp. 1605-1608.

[7] Milenkovic, P.H. (1993) Voice source model for continuous control of pitch period. J. Acoust. Soc. Am., Vol. 93 (2), pp. 1087-1096.