Comparing methods for automatic extraction of voice source parameters from continuous speech

Helmer Strik, Joop Jansen and Louis Boves

University of Nijmegen, Dept. of Language and Speech
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

ABSTRACT

Two methods are presented for automatic calculation of the voice source parameters from continuous speech. Both methods are used to calculate the voice source parameters for natural speech. However, for natural speech no objective test procedure seems available. Therefore, both methods were also tested on synthetic speech.

INTRODUCTION

Modern text-to-speech systems produce speech that is intelligible, but not quite natural. A substantial improvement of the naturalness of synthetic speech can probably be achieved by the use of a properly controlled voice source model. To derive rules for voice source parameters large amounts of data are required. Extracting voice source parameters by hand is time consuming, subjective and thus probably not reproducible. Automatic extraction of voice source parameters from continuous speech is also far from trivial. Our aim is to develop methods for automatic extraction of voice source parameters from continuous speech.In this article we propose and test two generic automatic methods.

In continuous speech the glottal parameters may change from period to period. Thus, the procedure must be pitch-synchronous. In our case, estimates of the frequency response of the vocal tract are based on an analysis of the closed glottis interval only. It is well known that analysis over such short intervals can yield wildly fluctuating results if, for instance, the analysis window is shifted or extended by just one or two samples. Because the causes behind these fluctuations are not understood, it is not possible to determine the optimal window location and length by simple automatic procedures. Therefore, we estimate the vocal tract transfer function for a number of window positions and lengths for each pitch period. The two methods differ in the way in which these multiple estimates are used. Method 1 computes an inverse filtered flow waveform for each individual estimate, and then computes the median value for the parameters describing the glottal waveforms. Method 2 uses the multiple estimates to obtain a single optimal estimate of the vocal tract transfer function, that is subsequently used to obtain a single optimal glottal flow waveform.

Both methods are used to derive voice source parameters from natural speech. But for natural speech it is difficult to evaluate the performance objectively, because the true glottal flow waveforms are not known. The only possible evaluation is a subjective one, viz. checking that the voice source parameters behave the way they are expected to behave. As long as there is no objective test procedure for natural speech, an objective evaluation can only be done with synthetic speech. In this article we will first test if both methods give plausible results for natural speech. Then we will perform an objective, quantitative test of both methods with synthetic speech.

MATERIAL

Natural speech

To study voice source characteristics data were obtained for four male subjects. For the current article only the data of one subject were used. The speech signal was transduced by a condensor microphone (B&K 4134) placed about 10 cm in front of the mouth, pre-amplified at the microphone (B&K 1619), and amplified by a measuring amplifier (B&K 2607). The speech signal was A/D converted off-line at a 10 kHz sampling rate.

Synthetic speech

The synthesis system that is used to generate the synthetic speech is a serial pole-zero synthesizer [1] that uses the DEC-Talk source [2]. The main reason for using the DEC-Talk source is its computational simplicity. All synthesis signals have a sampling frequency of 10 kHz.

METHOD

Inverse filtering is often used to obtain an estimate of the glottal flow signal. At the base of this method is the assumption that the voice source and the vocal tract filter do not interact. It is known that this assumption is not valid [3], but it is a useful approximation of the human speech system. The approximation is best during the closed glottis intervals, because then there is least interaction between the sub- and supraglottal cavities. Therefore, the analysis window should preferably be confined to the closed glottis interval. In [4] we showed that Closed glottis interval Covariance Linear Predictive (CC-LP) analysis is as powerful as more sophisticated techniques, like robust pole-zero analysis.

General method

A block diagram of the general method is shown in Fig. 1. First the vocal tract filter is estimated. This is done by converting the results of the CC-LP analysis to M Formant-Bandwidth pairs. This module is therefore called "FB-EST" (see Fig. 1). The problem of finding the correct set of analysis parameters is treated in the next section.

The estimated filter is inverted, and the inverse filter is used to filter the audio signal (module IF in Fig. 1). The resulting inverse filtered signal, INV, is a first estimate of the differentiated glottal volume flow. As the inverse filtered signal often contains high-frequency noise, it is low-pass filtered (module LPF in Fig. 1c). The resulting signal is a new, usually better, estimate of the differentiated glottal volume flow, dU_g. The glottal volume flow, U_g, is obtained by integration of dU_g (module INT in Fig. 1).

Voice source parameters, like open quotient and skewing, could be derived directly from dU_g and U_g. However, since these signals

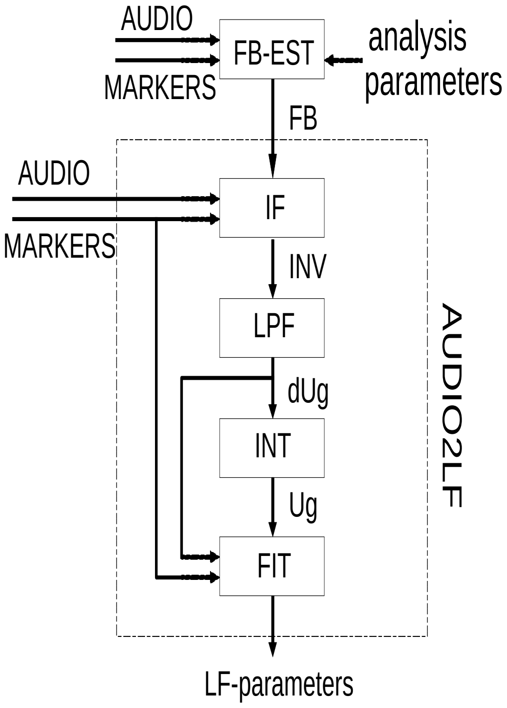

Figure 1 Block diagram of the general method.

often are noisy direct measurements yield unreliable values. Fitting a voice source model to the data probably is a more robust method for obtaining voice source parameters. In our system, the fit is done simultaneously on dU_g and U_g [5].

The LF-model was used as voice source model because it seems useful for synthesis, and because it has already been studied in great detail [6]. The model and its parameters are shown in Fig. 2. The LF-model is a four parameter model. There are different combinations of the LF-parameters that uniquely define a flow pulse. The four parameters that are used for the generation of flow pulses during the fit procedure are U_0, T_p, T_e, and T_a. Both during analysis and synthesis a fifth parameter is necessary to position the LF-pulses, viz. T_0.

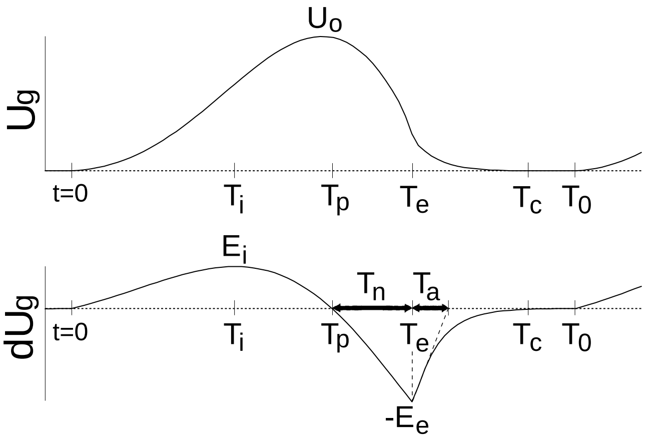

Figure 2 Glottal flow (U_g) and glottal flow derivative (dU_g) with the parameters of the LF-model.

The general method can be split in two parts. In the first part the formants and bandwidths of the vocal tract filter are estimated (FB-EST), and in the second part this filter is used to derive LF-parameters from the audio signal. This second module is therefore called AUDIO2LF (see Fig. 1).

Analysis parameters

For CC-LP analysis a number of choices have to be made like position and width of the closed glottis interval, order of the analysis, and pre-emphasis factor. Usually, no combination of choices is optimal for the whole utterance. However, for the natural speech that was used in this study a 12th order LPC analysis with a pre-emphasis factor of 1.00 (+6 dB/oct) worked satisfactorily for almost all pitch periods. Thus, window position and window length are left as parameters that can be varied.

Generally, the moments of glottal closure are easier to detect than the moments of opening. It is possible to identify the locations of the main excitation from the audio signal [7]. However, we prefer to determine the moments of glottal closure from the electroglottogram [5]. The signal with the closure markers is called "MARKERS" (see Fig. 1). The window position will be determined relative to a closure marker (this parameter is called window shift). Five window shifts (of -2, -1, 0, 1, 2 samples) were used. For synthetic speech the close markers are found by using the same operations (differentiation, low-pass filtering, and peak-picking) for the source signal.

For natural speech it is difficult to obtain an accurate estimate of the instant of glottal opening. Therefore, LP analysis is done for five different, fixed window lengths. If the length of the window is too large, then the last part will be in the open phase. This will perturb the estimates of the formants and especially of the bandwidths. On the other hand, if the length of the window is too short then LP analysis will not be able to make an accurate estimate. As a compromise we used window lengths of 33, 34, 35, 36, and 37 samples.

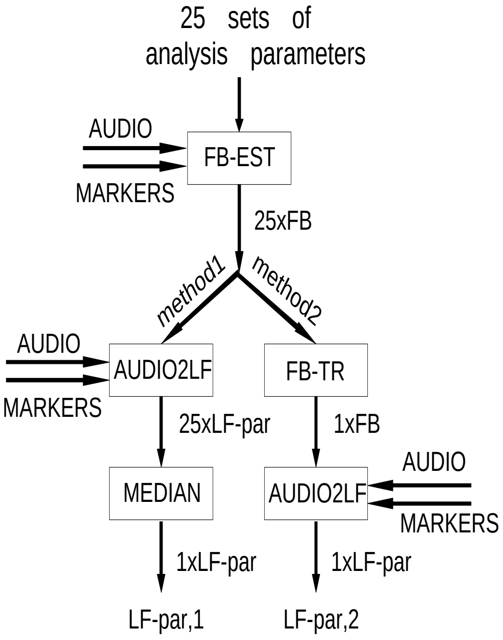

Figure 3 Block diagram with both methods.

In both methods the FB-pairs are estimated for the 25 combinations of analysis parameters (see Fig. 3). The 25 estimations of the vocal tract filter are used by both methods to obtain a single set of voice source parameters.

Method 1

In the first method the module AUDIO2LF is used 25 times (see Fig. 3). The result is 25 sets of LF-parameters. Finally median values are calculated for the LF-parameters (module MEDIAN in Fig. 3), which results in one set of average LF-parameters. This method is described in more detail in [5].

Method 2

In method 2 a Formant-Bandwidth TRacker (module FB-TR in Fig. 3) is used to obtain an optimal set of FB-pairs. In this method the module AUDIO2LF is applied once, where it uses the optimal filter to obtain one set of LF-parameters.

FB-Tracker

The goal of the module FB-TR is to find the first 4 FB-pairs. Generally, the first 4 FB-pairs are among the first 5 poles that are modelled by the LP analysis. Therefore, all 5 possible combinations of 4 out of 5 poles are generated. This is done for the 25 estimates of the vocal tract filter, and the result is a lattice of 4 FB-pairs with a depth of 125. A Viterbi algorithm is used to find the optimal path in this lattice.

The optimal path is the path with minimal total cost. For calculation of the cost, a transition and a local cost function is used. The transition cost function is the Euclidean distance between the values of the current frame and the values of the previous frame. The local cost function is the Euclidean distance between the values of the current frame and reference values. The reference values are obtained by a correlation LP analysis, using an analysis window of 256 samples.

For higher formants the variation in both frequency and bandwidth values is higher than for lower formants. For FB-tracking the effect would be that the FB-values of higher formants are more important in determining the optimal path. Therefore, all variables are converted to standard normal variables by first subtracting the mean value, and subsequently dividing by the standard deviation.

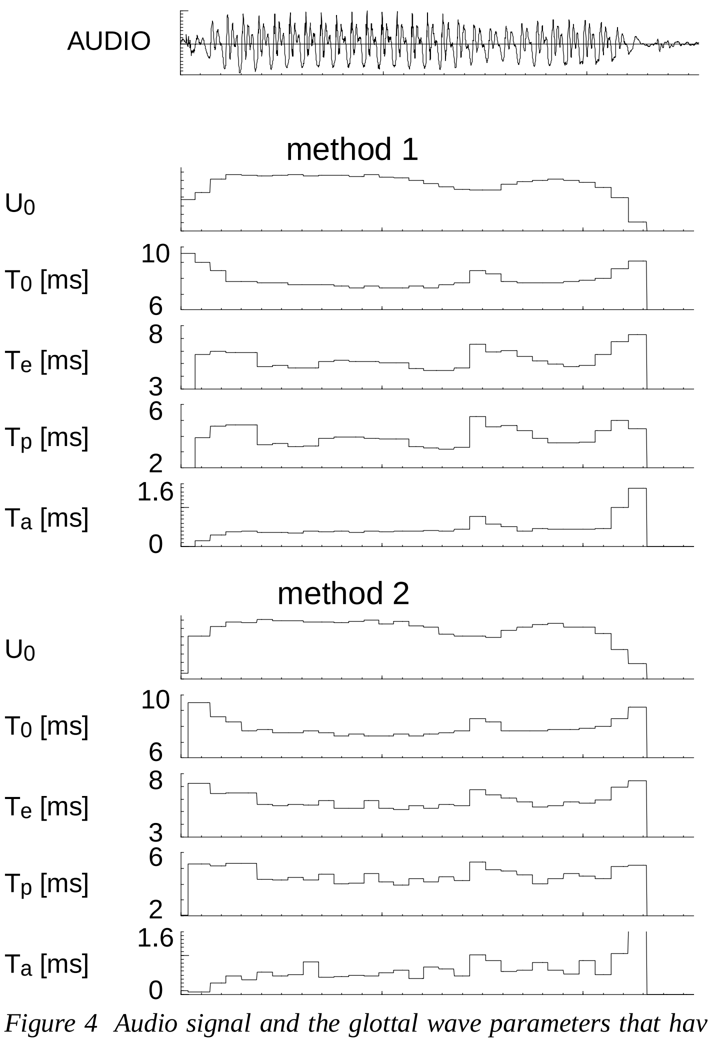

Figure 4 Audio signal and the glottal wave parameters that have been calculated by both methods.

The total cost function has four contributions, viz. the transition and local costs of the formants and the transition and local costs of the bandwidths. Each of these four contributions can be given a weight. The weights that have been used are 4, 2, 1, and 0, respectively. The reference bandwidths that are obtained by correlation LP analysis are usually larger than the bandwidths obtained by CC-LP. These bandwidths can not be compared in a straightforward manner, and consequently the weight of the local cost of the bandwidths is set to zero.

RESULTS

Natural speech

For the current article data are used that are obtained by applying both methods to a natural utterance with a length of about 2.5 seconds. In Fig. 4 the audio signal of part of the utterance is shown, together with T_0 and the four LF-parameters as calculated by both methods.

The part of the utterance shown in Fig. 4 clearly demonstrates the dynamics of the LF-parameters. Generally, it is observed that U_0 covaries with the amplitude of the speech signal. The same holds for the other amplitude related parameters of the LF-model, like E_e and E_i. During transitions from vowels to consonants it is often observed that U_0 decreases, while the time parameters T_0, T_e, T_p and T_a increase. The same effects were found by Bickley and Stevens for artificial vocal tract constrictions [8].

The four LF-parameters and T_0 can be used to calculate all other glottal waveform parameters like open quotient, speed quotient, skewing etc. As an example the 3 dimensionless wave shape parameters have been calculated: $R_g = T_0/2T_p$, $R_k = T_e/T_p - 1$, and $R_a = T_a/T_0$. The average values of R_g, R_k, and R_a for all 194 voiced periods of the utterance are 111, 42, and 6.7 for method 1, and 110, 41, and 7.1 for method 2. These values are in accordance with the values given in [9].

Although the results of both methods are slightly different, both methods seem to give plausible results. Given these results, two questions emerged: what is the reliability of the results, and which method is the best? As there is no objective test procedure for natural speech, we also tested both methods for synthetic speech.

Synthetic speech

The DEC-Talk waveform and the LF waveform are fundamentally different. LF-parameters that are derived by both methods from synthetic speech can not be compared directly to source signal that is used during synthesis. For evaluation of the test results a reference was required. This reference was obtained by performing the same operations on the source signal as on the inverse filter results, i.e. low-pass filtering, integration and fitting of a LF-model.

The rationale behind this is that the differentiated source signal should really be compared to the inverse filtered signals. All operations that are necessary to obtain LF-parameters from the inverse filtered signals should therefore also be applied to the differentiated source signal. In Fig. 5 the low-pass filtered and the fitted signal are shown. Apart from the fitted signal, the fit also yields the four LF-parameters for each pitch period. These reference parameters are used below for evaluation. The utterance used for evaluation was a random concatenation of all vowels, liquids and glides that are used in the synthesis system.

In Fig. 5 it can be seen that low-pass filtering and fitting mainly influences the excitation strength and the return phase. The effects are clear for a synthetic glottal pulse (as in Fig. 5), but the same ef-

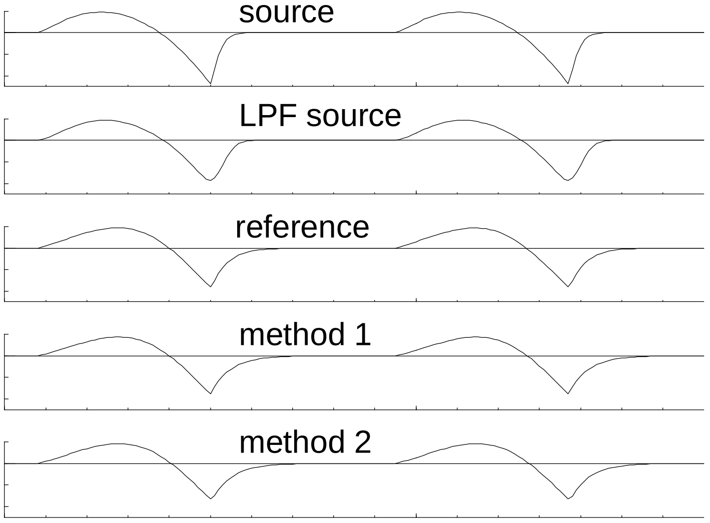

Figure 5 Shown are from top to bottom: 2 periods of the DEC-Talk voice source for a consonant /l/, the low-pass filtered source signal, the reference signal, and the source signals that have been calculated from the synthetic speech signal by method 1 and 2.

fects occur for the inverse filtered signals that are derived from the speech signals. In order to calculate the true values of the voice source parameters, a correction is mandatory after both methods. The amount of both corrections can be calculated from synthetic speech, and should be verified for natural speech. In the following part we will only test the similarity of the reference parameters (set 0) and the parameters obtained by both methods (set 1 and 2).

In Fig. 5 one can see that the reference signal and the source signals obtained by both methods from the synthetic speech signal are very much alike. A regression analysis on the voice source parameters has been done to test the degree of resemblance. The results are given in Table I.

Table I. Results of regression analysis for different combinations of the four LF-parameters. For each combination of two variables y and x a straight line is fitted through the data: *Y = intercept + slope∗X.* The correlation coefficient R denotes the goodness of the fit (N = 232). In the top box are given the comparisons of the results of both methods (subscript 1 and 2) with the reference (subscript 0), and in the bottom box the comparisons between both methods.

Y	X	intercept	slope	R
$U_{0,1}$	$U_{0,0}$	-9.7	1.03	0.96
$U_{0,2}$	$U_{0,0}$	-11.4	1.05	0.97
$T_{e,1}$	$T_{e,0}$	-0.33	0.95	0.75
$T_{e,2}$	$T_{e,0}$	0.07	1.01	0.76
$T_{p,1}$	$T_{p,0}$	0.25	0.96	0.60
$T_{p,2}$	$T_{p,0}$	-0.09	1.05	0.62
$T_{a,1}$	$T_{a,0}$	0.05	0.95	0.69
$T_{a,2}$	$T_{a,0}$	0.00	1.00	0.64
$U_{0,2}$	$U_{0,1}$	3.8	0.99	0.98
$T_{e,2}$	$T_{e,1}$	-0.05	1.01	0.97
$T_{p,2}$	$T_{p,1}$	-0.09	1.02	0.96
$T_{a,2}$	$T_{a,1}$	0.02	0.90	0.80

The results of the comparisons between both methods and the reference are given in the upper part of Table I. All slopes are about 1, and all intercepts are small which means that the parameters calculated by both methods have, on the average, the same value as the reference parameters. (The intercept of U_0 has a larger absolute value, but the value of the intercept relative to the average value of U_0 is even smaller than those of T_e, T_p, and T_a.) The correlation coefficients in Table I for U_0 are almost 1, and those of T_e, T_p, and T_a are somewhat smaller but still highly significant. This means that the values calculated by both methods for T_e, T_p, and T_a closely resemble the reference, while the calculated values for U_0 and the reference values are almost identical. Generally, the results of method 2 are slightly better than those of method 1.

In the lower part of Table I both methods are compared. For U_0, T_e, and T_p the intercept is small, and the correlation coefficient and the slope are almost 1. This means that for U_0, T_e, and T_p the results of both methods are very much alike. The resemblance of the values of T_a for both methods is somewhat less.

CONCLUSIONS

In this article 2 methods are proposed for the automatic extraction of voice source parameters from continuous speech. For natural speech both methods produce comparable and reasonable results. There is a need for an objective procedure to test the reliability of the extracted parameters. As long as this test is not available, the best alternative is to test these methods on synthetic speech.

The various operations that are used during the extraction of the voice source parameters from speech, have influence on the magnitude of these parameters. In order to re-estimate the true magnitude of the parameters these effects have to be corrected.

From the tests on synthetic speech it appeared that both methods succeed in estimating the voice source parameters quite accurately. The results obtained for the amplitude parameter U_0 are better than those of the time parameters T_e, T_p, and T_a. For synthetic speech the second method is slightly better than the first method.

REFERENCES

[1] L. Boves, J. Kerkhoff, and H. Loman. "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, 1987.

[2] D.H. Klatt and L. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers, " J. Acoust. Soc. Am. 87(2), pp. 820-857, 1990.

[3] T.V. Anantapadmanabha and G. Fant. "Calculation of true glottal flow and its components, " Speech Communication, Vol. 1, pp. 167-184, 1982.

[4] J. de Veth, B. Cranen, H. Strik, and L. Boves. "Extraction of control parameters for the voice source in a text-to-speech system, " Proc. Int. Conf. Acoust. Speech Signal Processing, Vol. 1, pp. 301-304, 1990.

[5] H. Strik and L. Boves. "On the relation between voice source parameters and prosodic features in connected speech, " to appear in Speech Communication, Vol. 11, 1992.

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

[7] D.G. Childers and C.K. Lee. "Vocal quality factors: Analysis, synthesis, and perception, " J. Acoust. Soc. Am. 90(5), pp. 2394-2410, 1991.

[8] C.A. Bickley and K.N. Stevens. "Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies, " J. of Phon., Vol. 14, pp. 373-382, 1986.

[9] R. Carlson, G. Fant, C. Gobl, B. Granstrom, I. Karlsson and Q. Lin. "Voice source rules for text-to-speech synthesis, " Proc. Int. Conf. Acoust. Speech Signal Processing, Vol. 1, pp. 223-226, 1989.