Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses

Helmer Strik[a)]
University of Nijmegen, Department of Language and Speech, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

(Received 14 July 1997; accepted for publication 14 January 1998)

The automatic parametrization of the first derivative of glottal flow is studied. Representatives of the two types of methods used most often for parametrization were tested and compared. The chosen representatives are all based on the Liljencrants–Fant model. As numerous tests were needed for a detailed comparison of the methods, a novel evaluation procedure is used which consists of the following stages: (1) use the Liljencrants–Fant model to generate synthetic flow pulses; (2) estimate voice source parameters for these synthetic flow pulses; and (3) calculate the errors by comparing the estimated values with the input values of the parameters. This evaluation procedure revealed that in order to reduce the average error in the estimated voice source parameters, the estimation methods should be able to estimate noninteger values of these parameters. The proposed evaluation method was also used to study the influence of low-pass filtering on the estimated voice source parameters. It turned out that low-pass filtering causes an error in all estimated voice source parameters. On average, the smallest errors were found for a parametrization method in which a voice source model is fitted to the flow derivative, and in which the voice source model is low-pass filtered with the same filter as the flow derivative. [S0001-4966(98)03204-4]

PACS numbers: 43.70.Aj, 43.72.Ar [AL]

INTRODUCTION

The technique of inverse filtering has been available for a long time now. This technique, which was first described in Miller (1959), can be used to decompose the speech signal into two components: the voice source and the filter (the vocal tract). In this way an estimate of the glottal volume velocity waveform (U_g) or its first derivative (dU_g) is obtained. For many applications, estimating a voice source signal (either U_g or dU_g) is not enough and the glottal flow signals have to be parametrized. Parametrization of the voice source signals and evaluation of the parametrization methods have received far less attention in the past. That is why we focus on these aspects in this study.

Parametrization of U_g or dU_g can be done in several ways. Often landmarks (like minima, maxima, zero crossings) are detected in the signals (e.g., Sundberg and Gauffin, 1979; Gauffin and Sundberg, 1980, 1989; Alku, 1992; Alku and Vilkman, 1995; Koreman, 1996). Because these landmarks are estimated directly from the voice source signals, these methods will be called direct estimation methods.

Voice source parameters are also calculated by fitting a voice source model to the data (e.g., Ananthapadmanabha, 1984; Schoentgen, 1990; Karlsson, 1992; Strik and Boves, 1992; Fant, 1993; Milenkovic, 1993; Alku *et al.*, 1997). Many different voice source models have been proposed in the literature (see, e.g., Rosenberg, 1971; Fant, 1979; Ananthapadmanabha, 1984; Fant *et al.*, 1985; Fujisaki and Ljungqvist, 1986; Lobo and Ainsworth, 1992; Cummings and Clements, 1995). Because in estimation methods of this kind a model fitting procedure is used, they will be referred to as ''fit estimation'' methods.

Estimation of voice source parameters can be useful for many applications. Although speech synthesis is the application most mentioned, the estimated voice source parameters are also used for fundamental research on speech production (e.g., Ní Chasaide and Gobl, 1993; Strik, 1994; Koreman, 1996). Other applications for which methods to measure voice source behavior could be useful are clinical use, speech analysis, speech coding, automatic speech recognition, and automatic speaker verification and identification. Since most of these applications require that the methods be fully automatic, there is an increasing need for automatic parametrization methods (see, e.g., Fritzell, 1992; Fant, 1993; Ní Chasaide and Gobl, 1993).

The development of an automatic parametrization method constitutes the long term goal of our research. Both direct and fit estimation methods can be made completely automatic. For this reason, and because they are the methods used most often, a representative of the direct estimation method will be compared with a representative of the fit estimation method. The representatives chosen are described in Secs. I E and I F.

The goals of the research reported on in this article are to find out what the pros and cons of each method are, to get a better understanding of the problems involved in estimating voice source parameters, and finally to determine which method performs best. In order to make it easier to compare the two methods, the same voice source model is used in both methods. To this end we use the Liljencrants–Fant (LF) model (Fant *et al.*, 1985). The LF model and the reasons for choosing it are described in Sec. I B. The evaluation method

[a)]Electronic mail: strik@let.kun.nl

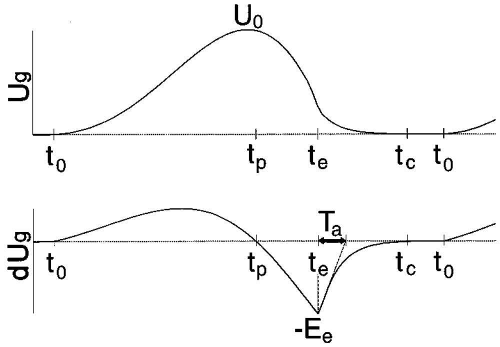

FIG. 1. Glottal flow (U_g) and glottal flow derivative (dU_g) with the parameters of the LF model: time of glottal opening (t_0); time (t_p) and value (U_0) of the maximum of U_g; time (t_e) and absolute value (Ee) of the minimum of dU_g; T_a describes the return phase, it is the length of the time interval between t_e and the projection of the tangent of dU_g in t_e; and the time of glottal closure (t_c).

and material are described in Secs. I C and I D, respectively. Because we want to focus on the parametrization method, we shall not evaluate inverse filtering in the current research. The performance of the parametrization methods is assessed in Secs. II and III. First, in Sec. II, it is studied how well the estimation methods succeed in estimating noninteger values of the parameters, which turned out to be a crucial property. Second, we focus on low-pass filtering in Sec. III. In Sec. IV the findings are discussed and some general conclusions are drawn.

I. GENERAL PROCEDURES

In this article two estimation methods used to parametrize dU_g are tested and compared. Before going on to describe these two methods (in Secs. I E and I F), we shall first give some definitions in Sec. I A, discuss the LF model in Sec. I B, and describe the method and material used for evaluation in Secs. I C and I D, respectively.

A. Definitions

In the current article it will be assumed that dU_g is a digital signal. In order to avoid confusion later on, we shall first define some terms related to sampling and quantization.

For all tests the sampling frequency $F_s = 10$ kHz, the number of bits used for quantization $B_c = 12$ and the amplitude range is $[-2048,2047]$. Consequently, the sampling time $T_s = 1/F_s = 1$ ms and the step size $\delta = 4096/2^{B_c} = 1$. Throughout this article a *time parameter* is said to have an integer value if its value is precisely an integer multiple of T_s. Likewise, an *amplitude parameter* is said to have an integer value if its value is exactly an integer multiple of δ.

B. Liljencrants–Fant model

In the current research the voice source model used is the LF model (see Fig. 1) because the LF model has the following advantages:

(1) In previous research the LF model has often been used to estimate voice source parameters, with manual or (semi-)automatic methods. This research has shown that it is a suitable model for description of the flow derivative (see, e.g., Fujisaki and Ljungqvist, 1986; Karlsson, 1992; Strik and Boves, 1992; Strik *et al.*, 1992, 1993; Childers and Ahn, 1995).

(2) Fujisaki and Ljungqvist (1986) compared several voice source models. Their results showed that the LF model and their own FL-4 model performed best (i.e., had the smallest prediction error).

(3) Previous research has also proven that the LF model is suitable for speech synthesis (see e.g., Carlson *et al.*, 1989).

(4) Due to all research already performed, the model and its behavior are well known.

The parameters shown in Fig. 1, in turn, can be used to derive many other parameters. For instance, the speed quotient is often calculated: $SQ = (t_p - t_0)/(t_c - t_p)$ (e.g., Alku and Vilkman, 1995). However, in our opinion these derived parameters are less suitable for evaluation of the parametrization methods, because whenever there is a change in a derived parameter, it is difficult to determine how this change came about (Strik, 1996). An increase in SQ could be the result of a larger t_p, a smaller t_0, a smaller t_c, or a combination of any of these three changes. On the other hand, whenever a derived parameter remains constant, this does not necessarily imply that the underlying parameters (i.e., the parameters which were used to calculate the derived parameters) remain constant. It is always possible that changes in these underlying parameters cancel each other out. Therefore, we prefer to use the LF parameters specified in Fig. 1 for the evaluation of estimation methods. Since the parameters E_e, t_0, t_p, t_e, and T_a give a complete description of an LF pulse, this set of parameters will be used in this article.

C. Evaluation method

Estimates of voice source parameters can be influenced by a large number of factors. So far, 11 of these factors have been studied: sampling frequency, number of bits used for quantization, position (shift) and amplitude (E_e) of the glottal pulses, t_c, T_0 (length of the fundamental period), signal-to-noise ratio (i.e., the effect of additive noise), phase distortion (which can be caused, e.g., by high-pass filtering), errors in the estimates of formant and bandwidth values during inverse filtering (which will bring about formant ripple in the estimated voice source signals), and low-pass filtering (Strik and Boves, 1994). We have performed over 1000 model fits for each of these 11 factors, making a total of much more than 11 000 model fits. The fact that so many tests had to be performed is the main reason for using the evaluation method described below (other reasons can be found in Strik, 1997).

In our experiments we first synthesize flow pulses (see Sec. I D). As we use the LF model for the fitting procedure, it is obvious that we also used the LF model to synthesize the flow pulses. Subsequently, the parametrization methods are used to estimate the voice source parameters. Finally, the estimated voice source parameters are compared with the correct values (used to synthesize the flow pulses), and the errors are calculated:

TABLE I. Values of t_p, t_e, and T_a (all in ms) for the 11 base pulses.

	Base pulse 1	2	3	4	5	6	7	8	9	10	11
t_p	14.0	14.0	16.0	16.0	16.0	16.0	14.0	14.0	15.2	15.2	15.2
t_e	15.2	15.2	17.2	17.2	18.8	18.8	16.0	16.0	17.2	17.2	17.2
T_a	0.4	1.6	0.4	1.6	0.4	0.8	0.4	1.6	0.4	1.0	1.6

$$\mathrm{ERR}(X)=|X_{\mathrm{est}}-X_{\mathrm{inp}}|/X_{\mathrm{inp}}, \quad \text{for } X=E_e$$

$$\mathrm{ERR}(Y)=|Y_{\mathrm{est}}-Y_{\mathrm{inp}}|, \quad \text{for } Y=t_0,\ t_p,\ t_e, \text{ and } T_a.$$

The experiments were carried out for a number (say N) of test pulses. After calculating the errors in the estimates of the five LF parameters for each test pulse, the errors had to be averaged. This can be done in a number of ways. Generally, averaging was done by taking the median of the absolute values of the errors. Absolute values were taken because otherwise positive and negative errors could cancel each other. The median was taken because (compared to the arithmetic mean) it is less affected by outliers which are occasionally present in the estimates. This method of averaging is the default method in the current article. Whenever another way of averaging was used, this is explicitly mentioned in the text.

In all figures below, the errors are arranged in a similar fashion (see, e.g., Fig. 2). In the upper left corner are the errors for E_e (in%), in the middle row are the errors for t_0 and t_p and in the bottom row are the errors for t_e and T_a. The errors in the time parameters t_0, t_p, t_e, and T_a are expressed in μs.

D. Material

The estimation methods used in this study are pitch synchronous. Among the parameters that have to be estimated are t_0 and t_c. Because these two parameters are not known beforehand, the pitch period cannot be segmented exactly. In practice, we first locate the main excitations (i.e., t_e) and then use a window with a width larger than the length of the longest (expected) pitch period. Generally, the pitch period will be situated between two other pitch periods (except for UV/V and V/UV transitions). Therefore, for each experiment sequences of three equal LF pulses were used. Each time voice source parameters were estimated for the (perturbed) pulse in the middle. Another reason for not using a single glottal pulse for evaluation is that the effects of perturbations cannot always be studied by a single, isolated LF pulse.

Since the effect of a studied factor can depend on the shape of a flow pulse, LF pulses with different shapes were used. These pulses will be called the base pulses. The base pulses were obtained by using the LF model for different values of the LF parameters. The parameters of E_e, T_0, t_0, and t_c were kept constant at 1024, 10 ms, 10 ms, and 20 ms, respectively. The values given for t_0 and t_c are the values for the second of the three pulses. For the first pulse one should subtract 10 ms from the values of t_0 and t_c, and for the last pulse add 10 ms. T_0 and t_c were kept constant because the results of our experiments showed that varying these parameters had very little effect on the estimations. The effects of varying E_e and shift (which is strongly related to t_0) were studied separately (see Sec. II).

For defining the base pulses the values of t_p, t_e, and T_a were varied. Based on the data given in Carlson *et al.* (1989), and the data from previous experiments (Strik and Boves, 1992; Strik *et al.*, 1992, 1993; Strik, 1994) the 11 base pulses shown in Table I were defined.

Subsequently, these 11 base pulses were used to generate the test pulses. For instance, to study the influence of the factor low-pass filtering, the 11 base pulses were filtered with M low-pass filters in order to generate $M\times 11$ test pulses. Calculation of the base pulses and the test pulses was first done in floating point arithmetic. After the test pulses had been created, the sample values were rounded towards the nearest integer (as is done in straightforward A/D conversion).

E. Direct estimation method

In direct estimation methods, voice source parameters are calculated directly from dU_g or U_g by means of simple arithmetic operators like min, max, argmin, and argmax. These arithmetic operators are used to detect landmarks in the signals. Some examples of estimations used quite often are: $U_0=\max(U_g)$, $t_p=\mathrm{argmax}(U_g)$, $E_e=-\min(dU_g)$, and $t_e=\mathrm{argmin}(dU_g)$ (see, e.g., Sundberg and Gauffin, 1979; Ananthapadmanabha, 1984; Gauffin and Sundberg, 1980, 1989; Alku, 1992; Alku and Vilkman, 1995; Koreman, 1996). Except for the value and the place of a maximum or minimum, the place of a zero crossing is also used to estimate parameters. For instance, in this way t_0 and t_c can be estimated (see Fig. 1).

One of the aims of the research reported in this article is to compare the performance of a typical direct estimation method with that of a fit estimation method. To this end we chose the direct estimation method described in Alku and Vilkman (1995), primarily because these authors provide a fairly detailed description of their method (see especially page 765 of their article), and because with this method it was possible to estimate the LF parameters E_e, t_0, t_p, and t_e (for which they use the terms $A_{\min}$, t_0, t_m, and t_{dm}, respectively).

In their method Alku and Vilkman (1995) do not estimate T_a. They use the parameter t_{ret} to describe the return phase. Since T_a cannot be derived from t_{ret} and an LF model is not complete without T_a, another method had to be used to estimate T_a. For the current research all estimates were made in the time domain. Because it is very difficult to estimate T_a in the time domain with a direct estimation method, estimates of T_a were obtained by fitting the LF model to the glottal pulse. More precisely, for given values of E_e, t_0, t_p, and t_e (made with the direct estimation method) the optimal value of T_a was estimated by fitting the LF model to the data. Therefore, strictly speaking, only E_e, t_0, t_p, and t_e can be said to be the result of the direct estimation method, while T_a is subsequently estimated with a fitting procedure. However, it is important to notice that the estimate of T_a does depend to a large extent on the estimates of E_e, t_0, t_p, and t_e made before with the direct estimation method. Furthermore, estimating one parameter (here T_a) with a fitting procedure, is a relatively simple operation. Consequently, the results showed that the error in the estimates of T_a is mainly the result of the errors in the estimates of E_e, t_0, t_p, and t_e made with the direct estimation method. For instance, if estimates of E_e and/or t_e are too large, the resulting estimates of T_a will generally be too small.

F. Fit estimation methods

In our fit estimation method five LF parameters (E_e, t_0, t_p, t_e, and T_a) are estimated for each pitch period. The method consists of three stages:

(1) initial estimate;
(2) simplex search algorithm;
(3) Levenberg–Marquardt algorithm (Marquardt, 1963).

The goal of the fit estimation method is to determine a model fit which resembles the glottal pulse as much as possible. This resemblance is quantified by means of an error function, which is calculated in the following way. The optimization procedure provides a set of LF parameters. These LF parameters and the analytical expression of the LF model are used to calculate a continuous LF pulse. The LF pulse is then sampled and zeros are added before t_0 and after t_c (until the length of the fitted signal is equal to that of the glottal pulse). These samples of the fitted signal together with the samples of the glottal pulse constitute the input to the error function that provides a measure of the difference between these samples. The fitting procedure tries to minimize this error.

We have experimented with several error functions which were defined either in the time domain, the frequency domain, or in both domains simultaneously. Defining a suitable error function in the frequency domain, for this automatic fitting procedure, turned out to be problematic. Probably the main reason is that the spectrum contains some details (e.g., the harmonics structure, the high-frequency noise) which need not be fitted exactly. With simple error measures, like, e.g., the root-mean-square (rms) error, we did not succeed in obtaining a reasonable model fit. More sophisticated error functions are needed for this task. A suitable error function should abstract away from the details which are not important, and emphasize the important aspects (e.g., the slope of the spectrum).

In the time domain it is much easier to obtain a fairly good model fit of dU_g. Here a simple rms error does yield plausible results. Still, also in the time domain some aspects of dU_g could be more important than others. It is likely that more sophisticated error functions could be defined which emphasize the relevant (e.g., perceptual) aspects. However, what is relevant depends on the application. In the current research we did not have a specific application in mind. The goal of this research was to develop a method for which the error in the estimated voice source parameters is small. Therefore, an important property of the error function is that it should decrease when the errors in the voice source parameters become smaller (this may sound trivial, but it is not). The rms error (defined in the time domain) did have this property and thus was suitable for this task, as our experiments revealed.

For the fitting procedure different nonlinear optimization techniques were tested: several gradient algorithms and some versions of a nongradient algorithm, i.e., the simplex search algorithm of Nelder and Mead (1964). Of the algorithms tested the simplex search algorithms usually came closer to the global minimum than the gradient algorithms. Owing to discontinuities in the error function, gradient algorithms are more likely to get stuck in local minima than simplex search algorithms are. Therefore the best version of the simplex search algorithm is used in the second stage of the fit estimation method. However, in the neighborhood of a minimum, the simplex algorithm may do worse (see Nelder and Mead, 1964). As a final optimization, the Levenberg–Marquardt algorithm (a gradient algorithm) is therefore used in the third stage (Marquardt, 1963).

In order to start the simplex search algorithm of stage 2 an initial estimate is required, which is made in the first stage. In principle, the best available direct estimation method should be used to provide the initial estimate. In this case the rms error for the fit estimation method can never be larger, and will almost always be smaller than the rms error for the direct estimation method used (because in stage 2 and 3 of our fit estimation method the rms error can never increase, and usually decreases gradually). Consequently, the errors in the voice source parameters estimated with the fit estimation method would almost always be smaller than those estimated with the direct estimation method used for initial estimation. Therefore, if we had used the direct estimation method described in the section above for initial estimation, the performance of this direct estimation method would probably have been worse than that of the fit estimation method. Because we considered this to be an unfair starting point, we decided to apply for initial estimation the routine used in our previous research (Strik *et al.*, 1993).

In Sec. III we will introduce a second version of this fit estimation method. This second version differs only slightly from the version described here. Together with the direct estimation method described in Sec. I E, the number of methods studied amounts to three.

Above we already mentioned that so far 11 different factors have been studied. In this article we shall confine ourselves to the most important results, namely those concerning the factors position (shift) and amplitude (E_e) (Sec. II) and those of low-pass filtering (Sec. III).

II. EXPERIMENT 1:SHIFT AND AMPLITUDE

A. Introduction

Direct estimation methods try to locate (important) events in the voice source signals. Thus the resulting estimates are generally limited to the place or amplitude of samples in the discrete signals, i.e., they are integers. Our intention was to develop a fit estimation method that would make it possible to estimate noninteger values too. Here we shall test how well the fit estimation method succeeds in estimating noninteger values of the voice source parameters, and what the resulting errors are for the two estimation methods for different values of shift and amplitude.

B. Material

The definition of the 11 base pulses is such that all time parameters have an integer value (see Sec. I D). In order to create test pulses in which the time parameters did not have integer values, the 11 base pulses were shifted in steps of 0.01 ms, from 0.0 up to 0.1 ms (11 values). This variable will be called *shift*. For only two of the chosen 11 values of shift (i.e., shift=0.0 and 0.1), the time parameters will have an integer value, while for the other 9 values of shift all time parameters will have noninteger values.

In order to create test pulses in which the amplitude (E_e) does not have integer values the amplitude E_e was varied from 1023 to 1025 in steps of 0.2 (11 values). This makes a total of 1331 test pulses (11 base pulses$\times$11 shift values$\times$11 E_e values). Next, the direct estimation method and the fit estimation method were used to estimate the voice source parameters for these 1331 test pulses. The errors in these estimations were then calculated.

C. Results of the direct estimation method

First, the results of the direct estimation method are presented in Figs. 2 and 3. Each error in Fig. 2 is the median of 121 errors (11 base pulses$\times$11 E_e values), while each error in Fig. 3 is the median of another set of 121 errors (11 base pulses$\times$11 shift values).

Let us first look at the errors in Fig. 2. To estimate t_0 a threshold function is used in the direct estimation method. The consequence is that the estimate of t_0 is always much too large (on average about 820 μs; see Fig. 3). For a shift of 0.03 ms the average error in t_0 is minimal, while for a shift of 0.04 ms it suddenly becomes maximal. The reason is that this extra shift of 0.01 ms causes the threshold to be exceeded one sample later in many test pulses, and thus the average error in t_0 suddenly increases. The average errors of the other parameters all behave as expected: the average errors are zero for a shift of 0.0 and 0.1 ms and larger in between.

The errors in the estimates for different values of E_e are shown in Fig. 3. The errors in the time parameters t_0, t_p, and t_e obviously do not depend on the value of E_e. Therefore, the errors for these time parameters are constant. If a large number of moments is randomly distributed, the average error (both the arithmetic mean and the median) due to rounding toward the nearest sample would be $T_s/4=25$ μs. The average errors of t_p, t_e, and T_a do not deviate much from this theoretical average. The reason why the error in t_0 is much larger was already explained above.

The average errors in the estimates of E_e behave as was expected: the average errors are minimal for integer values of E_e, and are larger in between. The median error in E_e is never zero, because it is obtained by averaging over different values of shift, and for most values of shift the error in E_e is larger than zero. The estimate of T_a depends on the estimates of E_e and t_e, and thus is not constant as a function of E_e.

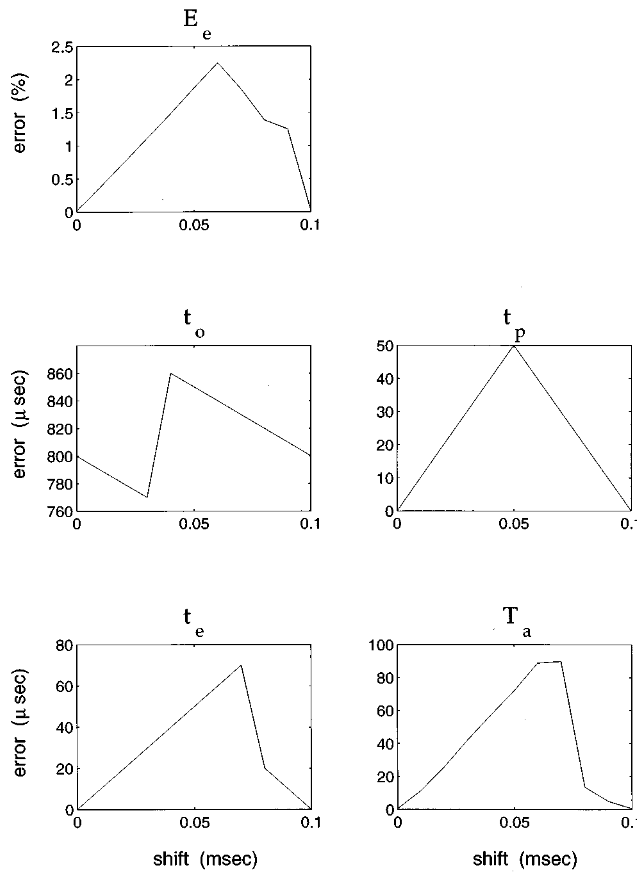

FIG. 2. Results of the direct estimation method: median error for the estimated parameters for different values of shift.

D. Results of the fit estimation method

The resulting average errors for the fit estimation method are shown in Figs. 4 and 5. In this case the errors were averaged by taking the mean value. This was done for two reasons: (1) since there are no outliers, median and mean values do not differ much; (2) by taking the mean it is also possible to calculate standard deviations. In turn, this makes it possible to test whether there is a significant difference between two mean values.

In this case for each value of shift the mean and standard deviation of 121 errors (11 base pulses$\times$11 E_e values) were calculated. The results are shown in Fig. 4. Likewise, for each value of E_e the mean and standard deviation of 121 errors (11 base pulses$\times$11 shift values) were calculated. The results are shown in Fig. 5.

In Figs. 4 and 5 one can observe that the mean errors do not differ significantly from each other. Furthermore, no trend can be observed in the errors. Put otherwise, the magnitude of the error in all estimated parameters does not depend on the value of the factors shift and E_e. Furthermore, all errors are very small, in general much smaller than the

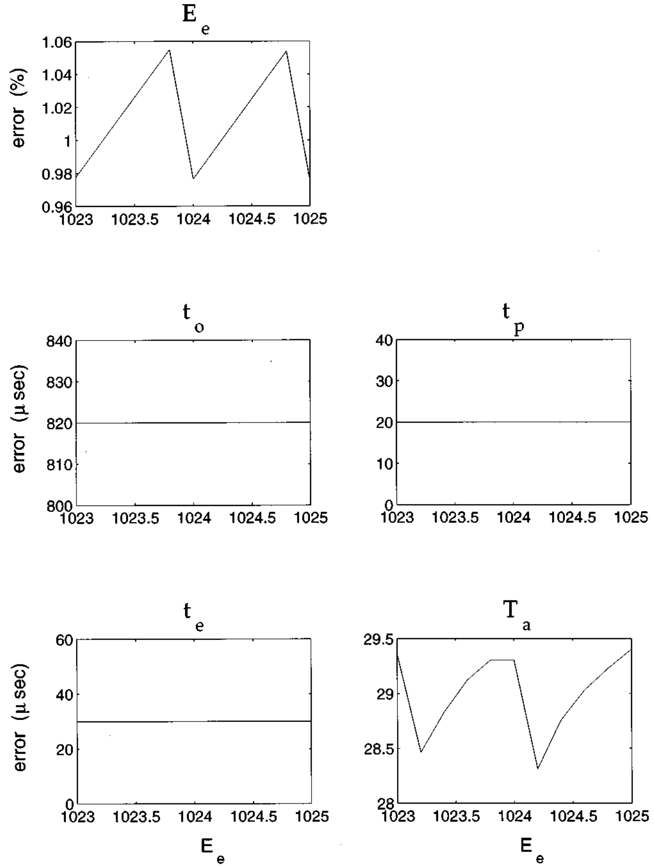

FIG. 3. Results of the direct estimation method: median error for the estimated parameters for different values of E_e.

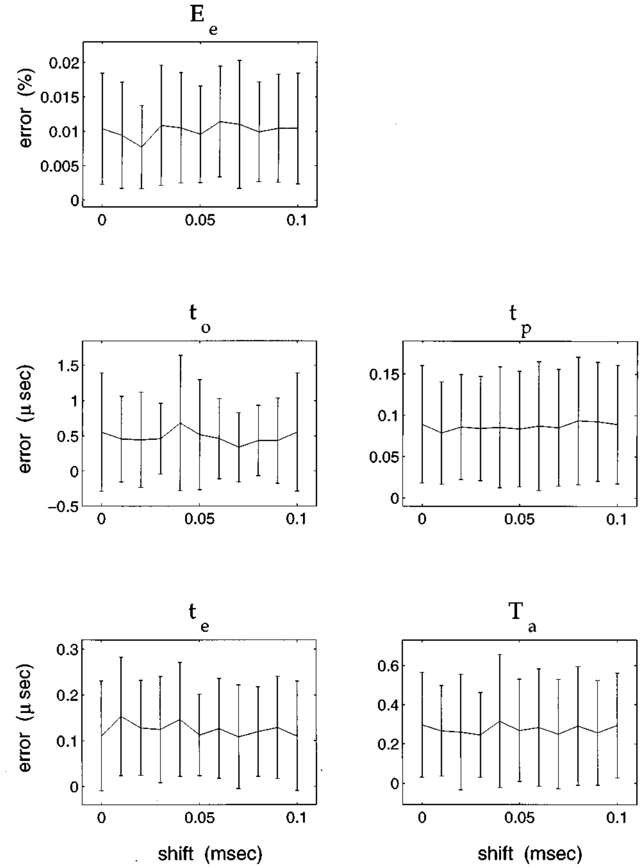

FIG. 4. Results of the fit estimation method: mean and standard deviation of the errors in the estimated parameters for different values of shift.

errors for the direct estimation method. Except of course for the cases in which all the LF parameters have an integer value. In the latter case the errors for the direct estimation method are zero, which is smaller still than the tiny errors found for the fit estimation method. However, it is clear that in practice the voice source parameters will seldom have exactly an integer value.

E. Conclusions

The conclusions that can be drawn from these tests are the following. The errors obtained with the fit estimation method are very small, in general much smaller than those for the direct estimation method. With the fit estimation method noninteger values can be estimated as accurately as integer values. Therefore, the quality of the model fit does not depend on the exact value of E_e and the position of the pulse (which is determined here by the variable shift). This explains why t_0 and E_e could be kept constant in the definition of the base pulses (see Sec. I D).

For the direct estimation method the average errors in t_0 are always larger than for the fit estimation method, because in the former a threshold function is used to estimate t_0. In fact, the error in t_0 can be substantially reduced, simply by subtracting a constant from its estimate. For the other parameters the estimation errors for the direct estimation method are zero if the parameters have exactly an integer value. Since in practice parameters rarely have an integer value, the estimates of the parameters will almost always contain an error due to this fact alone. These errors will be called the intrinsic errors, because they are intrinsic to the estimation methods. They will always be present, even if the glottal pulses are perfectly clean glottal pulses, as was the case in these tests. The results presented in this section make it possible to estimate what the average intrinsic errors are. For the direct estimation method the average error in the time parameters (except t_0) is about $T_s/4 = 25\ \mu$s, which is the theoretical average for randomly distributed values, while for E_e it is about 1% (see Fig. 3). For the fit estimation method the average error in the time parameters is less than 0.5 μs, while the average error for E_e is about 0.01% (see Figs. 4 and 5).

III. EXPERIMENT 2: LOW-PASS FILTERING

A. Introduction

Before the glottal flow signals are parametrized, they are low-pass filtered at least once in all methods, viz., before A/D conversion. Often, they are low-pass filtered again after A/D conversion, usually to cancel the effects of formants that were not inverse filtered or to attenuate the noise component. The latter operation seems very sensible for direct estimation methods, because in these methods high-frequency disturbances can influence the estimated parameters to a large extent. Although parametrization of inverse filtered signals has been done in many studies for almost 40 years now (i.e.,

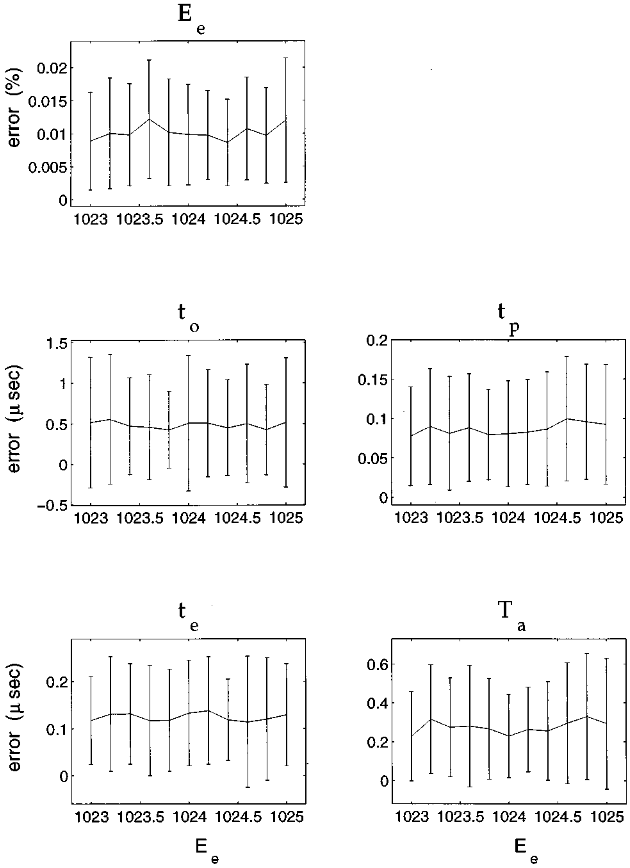

FIG. 5. Results of the fit estimation method: mean and standard deviation of the errors in the estimated parameters for different values of E_e.

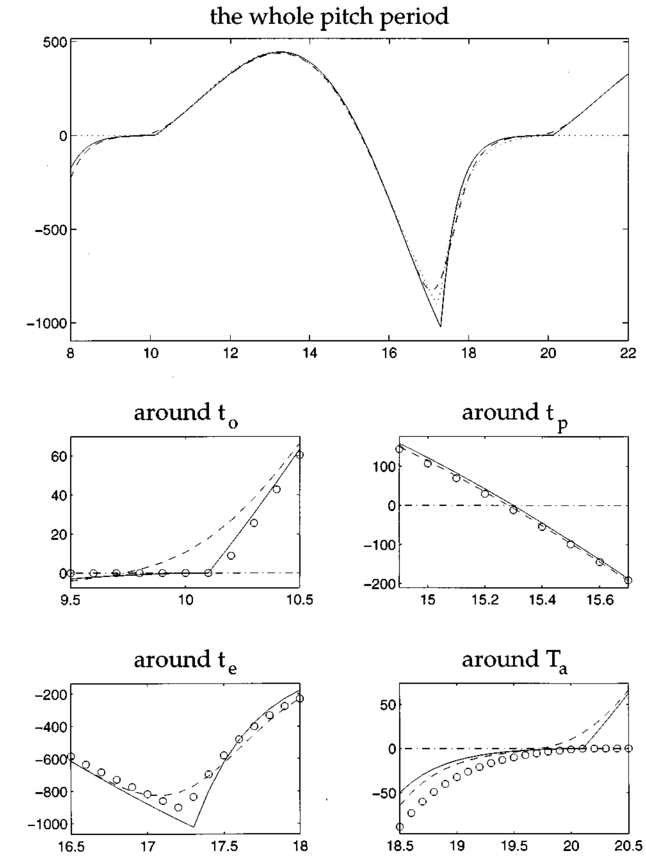

FIG. 6. An example of a differentiated flow pulse before (solid) and after (dashed) low-pass filtering, and a fit on the low-pass filtered pulse (dots in the top panel, open circles in the lower four panels). Shown are the whole pitch period, and some details around important events. For clarity, the zero line (dashed-dotted) has been omitted in the top panel.

since Miller, 1959), it has only recently been noted that low-pass filtering can influence the estimated voice source parameters (Strik *et al.*, 1992, 1993; Perkell *et al.*, 1994; Alku and Vilkman, 1995; Strik, 1996; Koreman, 1996). Thus it becomes very important to study what the effect of low-pass filtering exactly is. This will be done in the present section.

An example of the distortion of a differentiated flow pulse caused by low-pass filtering is given in Fig. 6. For low-pass filtering a convolution with a 19-point Blackman window was used. Shown are a base pulse before (solid) and after (dashed) low-pass filtering, and a model fit on the low-pass filtered pulse (dotted). Besides a picture of the three signals for the whole pitch period, some details around important events are also provided.

One can see in Fig. 6 that low-pass filtering does influence the shape of the pulse. From this figure one can deduce that the change in shape can have a large impact on the estimates obtained by means of a direct estimation method. This is most clear for the estimate of E_e, which will generally be too small, but the estimates of the other parameters will also be affected.

Low-pass filtering will also affect the estimates of a fit estimation method. After low-pass filtering the shape of the pulse is changed. The fitting procedure will try to find an LF pulse that resembles the filtered pulse as closely as possible. This is done by minimizing the rms error, which is a measure of the difference between the test pulse and the fitted LF pulse. The result is a fitted LF pulse that deviates from the original base pulse (see Fig. 6).

The distortion of the differentiated glottal flow signals depends on a number of factors, like, e.g., the type and the bandwidth of the low-pass filter, the frequency contents of the differentiated glottal flow signals, and the parametrization method used. We will study the effect of low-pass filtering for two parametrization methods (i.e., the direct estimation and the fit estimation method), for glottal pulses with different frequency contents (i.e., the 11 base pulses), and for different values of the bandwidth of the low-pass filter.

Low-pass filtering is done by means of a convolution with a Blackman window.[1] The bandwidth of this low-pass filter is varied by changing the length of the Blackman window (the longer the window, the smaller the bandwidth). This type of low-pass filtering was chosen because preliminary tests had shown that the error in the estimates induced by this filter was smaller than that of other tested filters. In part this can be explained by the fact that this low-pass filter does not have a ripple in its impulse response, while a ripple is present for many other low-pass filters. Therefore, for most other low-pass filters (including the generally used standard FIR filters) the estimation errors will be (much) larger than the errors presented below (Strik, 1996).

In the example provided in Fig. 6 the test signal is low-pass filtered. An LF model is then fitted to the low-pass

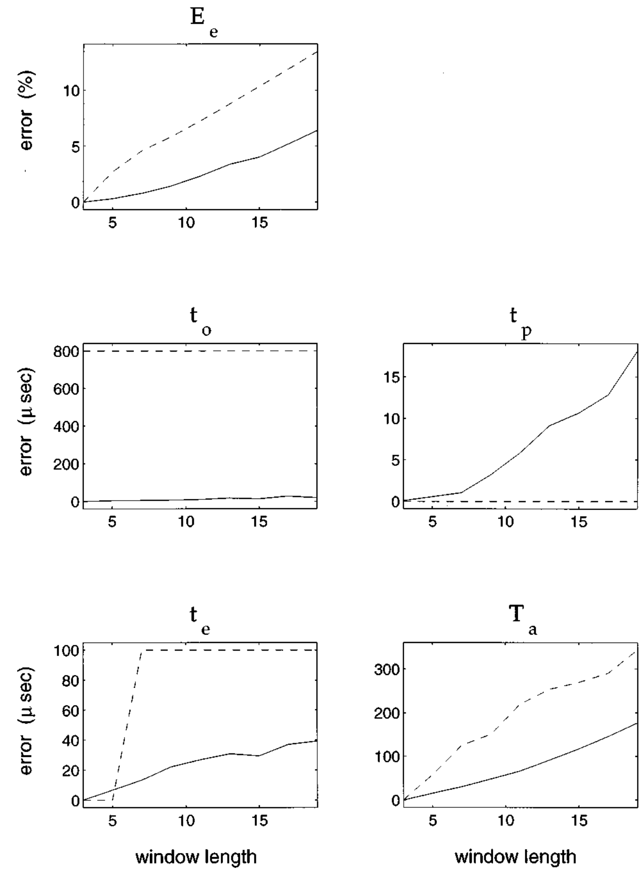

FIG. 7. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the direct estimation method (dashed) and for the first version of the fit estimation method (solid).

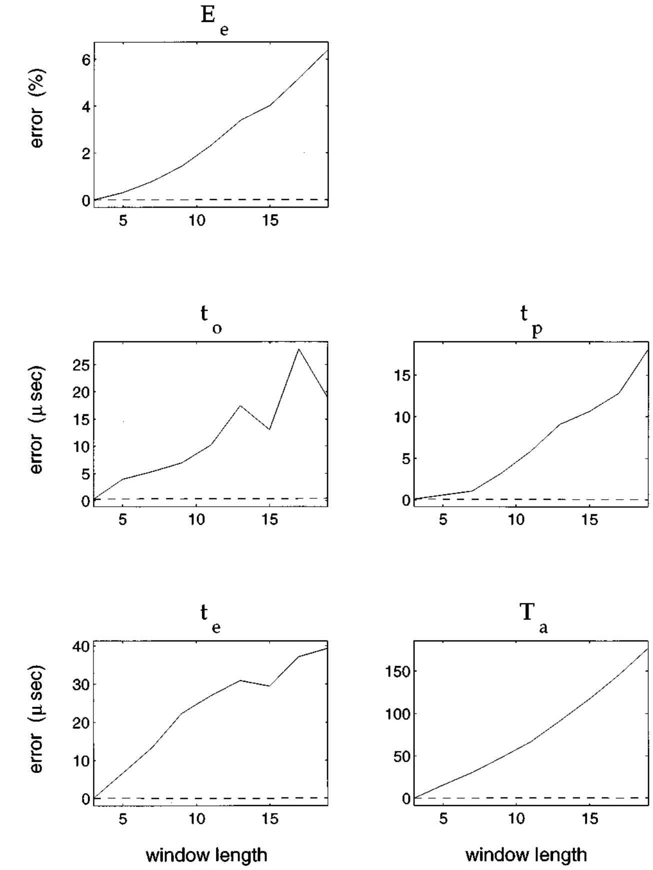

FIG. 8. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the first (solid) and the second (dashed) version of the fit estimation method. Note that the vertical scales are different from those in Fig. 7.

filtered test pulse. This seems the most obvious way to apply the fit estimation method, and will be called the first version of the fit estimation method. However, there is an alternative (which will be called the second version of the fit estimation method): apart from the test pulse one could also low-pass filter the fitted LF pulse. In this case, the test pulse and fitted LF pulse are altered in a similar fashion. In this way we hope to achieve that the error in the estimated parameters (which is due to low-pass filtering) will be smaller than when only the test pulses are low-pass filtered. It is obvious that the same procedure cannot be used in a direct estimation method, because in this case the parameters are calculated directly from the (low-pass filtered) signal.

B. Material

The 11 base pulses were low-pass filtered by means of a convolution with a Blackman window. The length of the window was varied from 3 to 19 samples in steps of 2 samples (9 lengths). For the resulting 99 test pulses (11 base pulses$\times$9 window lengths) the parameters were estimated with the direct estimation method and the fit estimation method. For each length of the Blackman window the results of the 11 base pulses were pooled and the median values of the absolute errors were calculated. These median values are shown in Figs. 7 and 8.

C. Results of the direct estimation method

In Fig. 6 one can see that low-pass filtering has most effect on the amplitude of the signal (E_e) and the shape of the return phase. Low-pass filtering causes the excitation peak to be smoother, and thus the estimate of E_e will be too small. Low-pass filtering also makes the return phase less steep, and therefore the estimate in T_a too large. These effects are enhanced if the length of the Blackman window increases (i.e., if the bandwidth of the low-pass filter is reduced). Therefore, the median errors of E_e and T_a increase with increasing window length.

Low-pass filtering does not have much influence on t_p (=the position of the zero crossing in dU_g; see Fig. 6). Therefore, in the majority of the cases the error in the estimates remains within half a sample, and the median of the errors is zero.

Usually, low-pass filtering causes the estimates of t_e to be too small (see Fig. 6). If the window length is 3 or 5, most of the errors in t_e remain within half a sample, and thus the median error is zero. However, for larger window lengths the errors in t_e become larger. As a result the median error increases too.

Finally, the error in t_0 remains constant, at the value of 820 μs (see also Fig. 3). This can be explained with the help of Fig. 6. In this figure one can see that low-pass filtering has a large effect on the signal in the direct neighborhood of t_0, and that this effect diminishes away from t_0. If the threshold chosen is high enough (which is the case for the direct estimation method used in the current research), low-pass filtering will not have much influence on this estimate of t_0.

D. Results of the fit estimation method

In Fig. 7 not only the errors of the direct estimation method are presented, but also those of the first version of the fit estimation method (i.e., the version in which only the test pulses were low-pass filtered). If the median errors of the fit estimation method are compared with those of the direct estimation method, the following observations can be made:

(i) The median errors are larger for t_p for all window lengths, and for t_e for windows with a length of 3 or 5.

(ii) In all other cases the errors of the first version of the fit estimation method are smaller than those of the direct estimation method.

The fact that in certain cases the error of the direct estimation method is smaller than the error of the fit estimation method can be explained quite easily. If the effect of a studied phenomenon (here low-pass filtering) on an event (here t_p or t_e) is such that the event is shifted by less than half a sample, the error with the direct estimation method is zero, while that of the fit estimation method is larger than zero. However, one should keep in mind that this is only the case for pulses in which all events coincide exactly with a sample position, as is the case with the test pulses. Only in this case does rounding towards the nearest sample position mean rounding towards the correct value.

In Fig. 8 the results of the two versions of the fit estimation method are compared, i.e., the first version, in which only the test pulses are low-pass filtered (solid lines), and the second version, in which both test pulses and fitted LF pulses are low-pass filtered (dashed lines). Clearly, the errors for the second version are much smaller. The errors are not zero, as may seem to be the case from Fig. 8, but they are extremely small. The largest error observed in the time parameters is 1 μs, and the errors in E_e are always smaller than 0.03%.

E. Conclusions

From our research we can conclude that low-pass filtering changes the shape of the flow pulses, and thus affects the estimates of all voice source parameters. The error due to low-pass filtering does depend on a lot of factors, e.g., the shape of the flow derivative, the low-pass filter and the estimation method used. So even for a given low-pass filter and estimation method (i.e., within one experiment) the error is not constant, because the shape of the glottal pulses is generally not constant. Furthermore, for a low-pass filter with a ripple in its impulse response (like the often used standard FIR filters) the average errors will be larger than for the low-pass filter used in this study, i.e., a convolution with a Blackman window (Strik, 1996).

Generally, the errors for the direct estimation method are larger than those of the first version of the fit estimation method. In turn, these errors are larger than the errors of the second version of the fit estimation method. Therefore, the conclusion is that the second version of the fit estimation method is superior. Low-pass filtering both the test pulse and the fitted voice source model seems to be a very good way to reduce the error caused by low-pass filtering. Of course, it cannot be used in a direct estimation method (as was already noted above).

IV. DISCUSSION AND GENERAL CONCLUSIONS

Before we draw our conclusions regarding the comparison of the three estimation methods, we first discuss some aspects of the fit estimation methods used in this study. The first aspect is the voice source model used in the fit estimation method, in our case the LF model. In the literature several voice source models have been described (see, e.g., Rosenberg, 1971; Fant, 1979; Ananthapadmanabha, 1984; Fant *et al.*, 1985; Fujisaki and Ljungqvist, 1986; Lobo and Ainsworth, 1992; Cummings and Clements, 1995). All voice source models for which an analytical expression exists can be used with the proposed fit estimation method to parametrize either U_g or dU_g. In our program there is a subroutine which calculates the fitted signal. The model fit is now calculated with the LF model, but this part can easily be replaced by the analytical expression of any voice source model. Furthermore, any number of voice source parameters can be used for parametrization. However, increasing the number of parameters makes the optimization problem (i.e., the error space) more complex, thus increasing the probability that the fitting procedure gets stuck in a local minimum.

Using a voice source model for parametrization has some advantages, one of them being the possibility that the estimated voice source parameters can subsequently be used for speech synthesis. Of course, for fit estimation methods a voice source model is mandatory. However, probably the most important disadvantage of a voice source model used for this purpose is that it cannot describe all the observed glottal pulses. Although the LF model is capable of describing many different glottal pulse shapes, it cannot describe all details. Whether a voice source model is suitable for a certain type of research depends on the goals of this research. Above we explained that with our fit estimation method it is possible to use many voice source models. The reasons for choosing the LF model in this study are given in Sec. I B.

The second aspect of the fit estimation method we want to discuss concerns the properties of the LF routine, which is the routine used to calculate the LF pulses. The way in which the LF routine is implemented turned out to be extremely important. The first version of our LF routine was taken from Lin (1990). Since in this version all input parameters are rounded toward the nearest integer, the shapes of the resulting LF pulses do not change gradually but abruptly. The consequence is that also the calculated rms error jumps from one value to the next. Thus the error function has the shape of a staircase, which is problematic for many optimization algorithms: they often get stuck in a local minimum. This is especially the case for gradient algorithms, because the gradient is zero for each stair.

In the second version of the LF routine, oversampling was used within the LF routine. For instance, we tried oversampling by a factor 10. Thus not only integer values can be estimated, but also nine values between these integers. However, the error function still has the shape of a staircase. Since the stairs are ten times smaller (compared to the first version of the LF routine), the resulting estimates were better. Still, the optimization often did not come close to the global minimum.

Our conclusion is that oversampling can reduce the width of the stairs in the error function, and thus improve the estimates, but it can never take away the fundamental problem for optimization, i.e., that the error function is a staircase. That is why we tried to find an implementation of the LF routine for which the error function changes smoothly. This property will be called the ''smooth property.'' The third version of the LF routine, which is described in Sec. I F, did have this property. In this version the analytical expression of the LF model is used to calculate a continuous LF pulse, which is then sampled. An enormous improvement in the fit estimation method was observed when the third version of the LF routine was used (compared to the first and second version). The reason is that a smooth error function is an enormous advantage for both simplex search and gradient algorithms. All results presented in this article are obtained with the third version of the LF routine.

The third aspect of the fit estimation method which will be discussed is that no anti-aliasing low-pass filter is used. In the LF routine a continuous LF pulse is first calculated and is then sampled with the same sampling frequency (F_s) as the flow derivative which has to be parametrized (here, 10 kHz). We did not use an anti-alias low-pass filter here, because we wanted to be able to study each factor in isolation. If we had used an anti-alias low-pass filter, this factor (and its effect on the estimated voice source parameters) would always have been present, thus making it impossible to study it independently of other factors.

If no anti-aliasing low-pass filter is used, aliasing effects can be present in the digital signals. Careful inspection showed that this was not the case for the LF pulses used in this study. The dU_g signals on average have a slope of -6 dB/oct. The first fundamental is at 100 Hz, so at 5 kHz the attenuation is usually more than 30 dB. Using a F_s of 10 kHz made it possible to study the effect of the factor low-pass filtering independently of other factors (like, e.g., shift and E_e).

If aliasing is a problem (e.g., because F_s is smaller than 10 kHz), an anti-alias low-pass filter has to be used. The most straightforward way to do this is to sample the continuous LF signal first with a sampling frequency F_s, and next use a digital low-pass filter with a bandwidth smaller than $F_s/2$. However, in that case the smooth property is lost, and the error function (which quantifies the difference between the LF signal and the flow derivative) becomes a staircase. The result is that the average error in the estimated voice source parameters becomes larger, as mentioned above. A somewhat better solution is to oversample the LF signal before digital low-pass filtering. By oversampling noninteger values can also be estimated. Furthermore, the stairs of the staircase become smaller. Consequently, the average error in the estimated voice source parameters also becomes smaller. Probably the best solution would be to use the analytic anti-alias low-pass filter proposed by Milenkovic (1993), which can be applied in continuous time. In this way the smooth property is preserved, and the error function remains a function that changes smoothly (instead of being a staircase).

In the current study two factors were studied in detail. As parameters rarely have an integer value, we first estimated what the resulting intrinsic errors are for the two methods. For the direct estimation method they turned out to be much larger than for the fit estimation method.

Next, the effect of the factor low-pass filtering was studied independently, i.e., with all input parameters having an integer value. For low-pass filtering we found that the errors of the direct estimation method are sometimes smaller than those of the fit estimation method. However, if the important events had been positioned randomly, the errors of the fit estimation method would have been slightly larger while those of the direct estimation method would have been substantially larger. For a realistic comparison of the two methods the intrinsic errors should be added to the errors found for low-pass filtering alone. If this is done the average errors of the direct estimation method are always larger than those of the first version of the fit estimation method, and these in turn are larger than the average errors of the second version of the fit estimation method.

The conclusion which can be drawn on the basis of the tests presented in this article is that the second version of the fit estimation method is superior. However, the effect of more single factors and factors in combination should be studied to get a more thorough understanding of the intricacies of the various parametrization methods.

In order to test and compare the parametrization methods we have used a novel evaluation method in which synthetic test material is generated by a production model. Subsequently, the same production model is used to re-estimate the synthesis parameters. This evaluation method turned out to be useful for our research, e.g., it helped us find the importance of the properties of the implementation of the LF routine and the effects of the factor low-pass filtering. We are convinced that with other evaluation methods this would have been much more difficult or even impossible (see also Strik, 1996).

Since in the present research we want to focus on the estimation of voice source parameters from the flow derivative, without being distracted by the problems of inverse filtering, we use a voice source model (the LF model) as the production model. For other purposes a vocal tract model or a complete synthesizer could be used.

A similar method was used by McGowan (1994) to evaluate the estimation of vocal tract parameters. In our research, just as in McGowan's work (1994), all details of the generating procedure are explicitly known. We therefore agree with him that these kinds of studies should be regarded as best case studies which can be used to study the limitations of estimation procedures and to optimize these estimation procedures. There are two other reasons why the present study is a best case study. First of all, because the test signals are clean LF pulses, and besides the influence of low-pass filtering contain none of the other disturbances that are generally present in natural speech. And second, because for a standard FIR filter, which is used most often as a low-pass filter, the resulting average errors are larger than for the low-pass filter used in this study. Consequently, when estimation methods are used to parametrize inverse filtered natural speech signals, the errors in the resulting parameters will generally be (much) larger.

The final topic we want to discuss is how the proposed estimation methods can be used to estimate voice source parameters for natural speech. The answer is straightforward: first use inverse filtering to obtain estimates of the glottal flow signals, and then apply the estimation methods. In Strik and Boves (1992) and Strik *et al.* (1992) we showed that this is possible for previous versions of the fit estimation method. We only have to exchange the previous version of the fit estimation method with the new improved version. The best solution would be to take the second version of the fit estimation method, and in the error routine use the same low-pass filter as used during the inverse filter procedure.

ACKNOWLEDGMENTS

The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. I would like to thank Loe Boves, Bert Cranen, and Jacques Koreman for fruitful discussions. Furthermore, I am grateful to Paavo Alku, Anders Löfqvist, and an anonymous reviewer for their useful comments on a previous version of this paper.

[1]This idea was suggested to me by Bert Cranen.

Alku, P. (**1992**). ''An automatic method to estimate the time-based parameters of the glottal pulseform,'' Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, Vol. 2, 29–32.

Alku, P., and Vilkman, E. (**1995**). ''Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering,'' J. Acoust. Soc. Am. **98**, 763–767.

Alku, P., Strik, H., and Vilkman, E. (**1997**). ''Parabolic Spectral Parameter—A new method for quantification of the glottal flow,'' Speech Commun. **22**, 67–79.

Ananthapadmanabha, T. V. (**1984**). ''Acoustic analysis of voice source dynamics,'' Speech Transmiss. Lab. Q. Prog. Stat. Rep. **2-3**, 1–24.

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., and Lin, Q. (**1989**). ''Voice source rules for text-to-speech synthesis,'' Proceedings of the IEEE International Conference on Acoustic Speech Signal Process, Glasgow, Scotland, Vol. 1, 223–226.

Childers, D. G., and Ahn, C. (**1995**). ''Modeling the glottal volume-velocity waveform for three voice types,'' J. Acoust. Soc. Am. **97**, 505–519.

Cummings, K. E., and Clements, M. A. (**1995**). ''Analysis of the glottal excitation of emotionally styled and stressed speech,'' J. Acoust. Soc. Am. **98**, 88–98.

Fant, G. (**1979**). ''Glottal source and excitation analysis,'' Speech Transmiss. Lab. Q. Prog. Stat. Rep. **1**, 70–85.

Fant, G. (**1993**). ''Some problems in voice source analysis,'' Speech Commun. **13**, 7–22.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). ''A four-parameter model of glottal flow,'' Speech Transmiss. Lab. Q. Prog. Stat. Rep. **4**, 1–13.

Fritzell, B. (**1992**). ''Inverse filtering,'' J. Voice **6**, 111–114.

Fujisaki, H., and Ljungqvist, M. (**1986**). ''Proposal and evaluation of models for the glottal source waveform,'' Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, Japan, Vol. 4, 1605–1608.

Gauffin, J., and Sundberg, J. (**1980**). ''Data on the glottal voice source behavior in vowel production,'' Speech Transmiss. Lab. Q. Prog. Stat. Rep. **2-3**, 61–70.

Gauffin, J., and Sundberg, J. (**1989**). ''Spectral correlates of glottal voice source waveform characteristics,'' J. Speech Hear. Res. **32**, 556–565.

Karlsson, I. (**1992**). ''Analysis and synthesis of different voices with emphasis on female speech,'' Ph.D. dissertation, KTH, Stockholm.

Koreman, J. (**1996**). ''Decoding linguistic information in the glottal airflow,'' Ph.D. dissertation, University of Nijmegen.

Lin, Q. (**1990**). ''Speech production theory and articulatory speech synthesis,'' Ph.D. dissertation, KTH, Stockholm.

Lobo, A. P., and Ainsworth, W. A. (**1992**). ''Evaluation of a glottal ARMA model of speech production,'' Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, Vol. 2, 13–16.

Marquardt, D. (**1963**). ''An algorithm for least-squares estimation of nonlinear parameters,'' SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **11**, 431–441.

McGowan, R. (**1994**). ''Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests,'' Speech Commun. **14**, 19–48.

Milenkovic, P. H. (**1993**). ''Voice source model for continuous control of pitch period,'' J. Acoust. Soc. Am. **93**, 1087–1096.

Miller, R. L. (**1959**). ''Nature of the Vocal Cord Wave,'' J. Acoust. Soc. Am. **31**, 667–677.

Nelder, J. A., and Mead, R. (**1964**). ''A simplex method for function minimization,'' Comput. J. (Switzerland) **7**, 308–313.

Ní Chasaide, A., and Gobl, C. (**1993**). ''Contextual variation of the vowel voice source as a function of adjacent consonants,'' Language and Speech **36**, 303–330.

Perkell, J. S., Hillman, R. E., and Holmberg, E. B. (**1994**). ''Group differences in measures of voice production and revised values of maximum airflow declination rate,'' J. Acoust. Soc. Am. **96**, 695–698.

Rosenberg, A. E. (**1971**). ''Effect of glottal pulse shape on the quality of natural vowels,'' J. Acoust. Soc. Am. **49**, 583–590.

Schoentgen, J. (**1990**). ''Non-linear signal representation and its application to the modelling of the glottal waveform,'' Speech Commun. **9**, 189–201.

Strik, H. (**1994**). ''Physiological control and behavior of the voice source in the production of prosody,'' Ph.D. dissertation, University of Nijmegen.

Strik, H. (**1996**). ''Comments on ''Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering'' [J. Acoust. Soc. Am. **98**, 763–767 (1995)],'' J. Acoust. Soc. Am. **100**, 1246–1249.

Strik, H. (**1997**). ''Automatic parametrization of voice source signals: A novel evaluation procedure is used to compare methods and test the effects of low-pass filtering,'' Internal Report, University of Nijmegen (available at http://lands.let.kun.nl/TSpublic/strik/).

Strik, H., and Boves, L. (**1992**). ''On the relation between voice source parameters and prosodic features in connected speech,'' Speech Commun. **11**, 167–174.

Strik, H., and Boves, L. (**1994**). ''Automatic estimation of voice source parameters,'' Proc. Int. Conf. Spoken Language Process., Yokohama, Japan, Vol. 1, 155–158.

Strik, H., Cranen, B., and Boves, L. (**1993**). ''Fitting an LF-model to inverse filter signals,'' Proc. of the 3rd European Conf. on Speech Technology, Berlin, Germany, Vol. 1, 103–106.

Strik, H., Jansen, J., and Boves, L. (**1992**). ''Comparing methods for automatic extraction of voice source parameters from continuous speech,'' Proc. Int. Conf. Spoken Language Process., Banff, Canada, Vol. 1, 121–124.

Sundberg, J., and Gauffin, J. (**1979**). ''Waveforms and spectrum of the glottal voice source,'' in *Frontiers of Speech Communication Research*, Festschrift for Gunnar Fant, edited by B. Lindblom and S. Öhman (Academic, London), pp. 301–320.