TESTING TWO AUTOMATIC METHODS FOR ESTIMATION OF VOICE SOURCE PARAMETERS

Helmer Strik

In: H. Strik, N. Oostdijk, C. Cucchiarini, & P.A. Coppen (eds.)
Proceedings of the Department of Language and Speech
Vol. 19, pp. 105-127, Nijmegen, The Netherlands, 1996

TESTING TWO AUTOMATIC METHODS FOR ESTIMATION OF VOICE SOURCE PARAMETERS

Helmer Strik

1. Introduction

Techniques for glottal inverse filtering have been available for a long time now. By means of inverse filtering it is possible to obtain an estimate of the derivative of the glottal flow signal (dU_g), either from the acoustic speech signal or from the airflow signal recorded at the lips. However, for many applications it is not sufficient to get an estimate of dU_g; a parameterization of dU_g is also needed. By combining these two techniques (i.e. inverse filtering and parameterization of dU_g) an analysis method for the extraction of voice source parameters can be constructed.

Manual versions of such an analysis method are already available, and have been used very often in previous research (a long list of references is given in Strik, 1996). However, a completely automatic method that works satisfactorily does not seem to exist, while there clearly is an increasing need for automatic methods (see e.g. Fritzel, 1992; Fant, 1993; Ni Chasaide and Gobl, 1993).

An analysis method for automatic extraction of voice source parameters from the audio signal is proposed in Strik and Boves (1992), Strik *et al.* (1992), and Strik (1994). In this method of analysis dU_g is first calculated, and then parameterized by fitting the LF-model (Fant *et al.*, 1985) to dU_g. For natural speech this method gave plausible results (Strik and Boves, 1992; Strik *et al.*, 1992), while for synthetic speech voice source parameters could be estimated with reasonable accuracy (Strik *et al.*, 1992; Strik *et al.*, 1993).

Over the last years this method for estimation of voice source parameters has been improved substantially. The results of tests with this improved estimation method are presented in the current article. Furthermore, the results obtained with this (improved) estimation method are compared to these obtained with another type of estimation method that is often used in this kind of research.

A description of these two estimation methods is given in section 2. The evaluation method and the material used for the tests are described in section 3. Section 4 deals with the various tests performed, and the results obtained. Finally, some general conclusions are drawn in section 5.

Due to space limitations, only part of the tests performed are described here, and some less important details are also omitted. A more complete report can be found in Strik (1996).

2. Estimation methods

Two estimation methods, used to parameterize dU_g, are tested and compared in this article. Before going on to describe these two methods, I will give some definitions here, in order to avoid confusion later on.

2.1 Some definitions

First I will define some terms related to A/D-conversion, which are used often below. For A/D-conversion, a choice has to be made for some values like the sampling frequency (F_s), the input range ($\Delta = [X_{min}, X_{max}]$), and the number of bits used to code each sample (B_c). As the number of bits used for coding is B_c, the number of amplitude levels $L = 2^{Bc}$, and the step size $\delta = \Delta/L$. The *step size* is the smallest possible difference between two amplitude values. The distance between two neighbouring sample points is called the *sample interval* or the *sampling time* $T_s = 1/F_s$. 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.

Now I shall focus on the voice source parameters. To parameterize dU_g different sets of parameters can be used. Furthermore, different names for the same voice source parameter can be found in the literature. Therefore, it is important to give a clear definition of the voice source parameters that I shall discuss in this article. The definitions of the voice source parameters used throughout this article are based on the LF-model (see Figure 1).

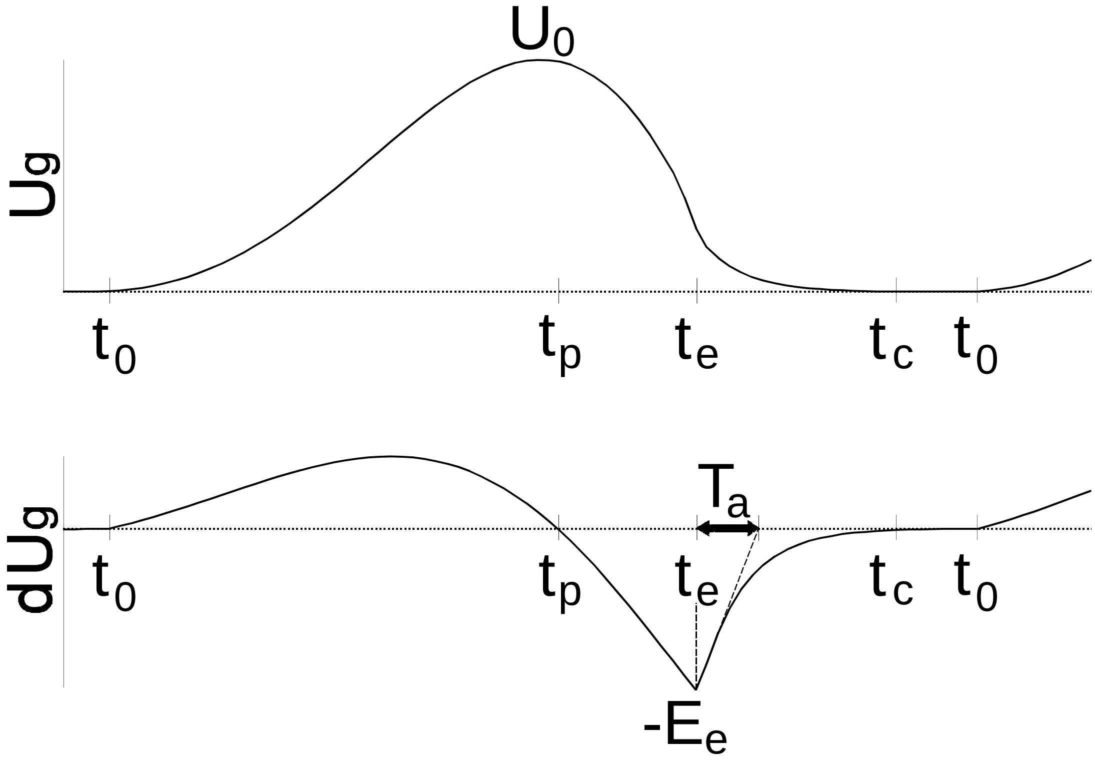


Figure 1. The LF-model and the LF-parameters.

These parameters can be divided into three groups:

1. amplitudes
   - ❒ E_e: excitation strength, $E_e = \min(dU_g)$
   - ❒ U_0: peak glottal flow, $U_0 = \max(U_g)$
2. moments
   - ❒ t_o: moment of opening
   - ❒ t_p: moment of peak in U_g, $t_p = \text{argmax}(U_g)$
   - ❒ t_e: moment of excitation, $t_e = \text{argmin}(dU_g)$
   - ❒ t_c: moment of closing
3. durations of time-intervals
   - ❒ T_0: duration of a pitch period, $T_0 = 1/F_0$,
   - ❒ T_a: duration of the interval between t_e and the projection of the tangent of dU_g in t_e.

2.2. Direct estimation method

In some 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. Some examples of estimations used quite often are: $U_0 = \max(U_g)$, $t_p = \text{argmax}(U_g)$, $E_e = -\min(dU_g)$, and $t_e = \text{argmin}(dU_g)$ (see e.g. Sundberg and Gauffin, 1979; Ananthapadmanabha, 1984; Gauffin and Sundberg, 1980; Gauffin and Sundberg, 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 zerocrossing is also used to estimate parameters. For instance, in that way t_o and t_c can be estimated (see Figure 1). Because in these methods the voice source parameters are estimated directly from the voice source signals, these methods will be called direct estimation methods (DE-methods).

With DE-methods, estimates of most voice source parameters can be obtained in a relatively simple way. However, DE-methods also have some disadvantages. An important disadvantage is that estimates are limited to the place or amplitude of samples in the discrete signals. Consequently, the estimated voice source parameters always have integer values. In practice, events generally will not coincide precisely with a sample point, and amplitudes will not always be exactly an integer multiple of the step size δ; i.e. the parameters will not have an integer value. The error in the estimated voice source parameters due to this property of the DE-methods will contribute to the total error.

Another drawback of the DE-methods is that a disturbance that is present in the estimated flow pulses can lead to large errors in the estimated parameters. For instance, noise or formant ripple can influence the position and the amplitude of certain events to a large extent. For a more thorough description of some drawbacks of DE-methods the reader is referred to Strik (1996) and Strik (to appear).

One of the aims of the research reported in this article is to test the performance of a DE-method, and to compare it with the performance of the estimation method proposed below. To that end I chose the DE-method described in Alku and Vilkman (1995), because they provide a fairly detailed description of their method (especially on page 765 of their article). After implementing this DE-method, numerous experiments were first carried out to improve the implementation. The goal was to make the estimation more robust, and thus to make the resulting average errors in the estimations smaller. In the following stage, the DE-method was used for the tests described below.

In their method Alku and Vilkman (1995) make estimations of several voice source parameters, of which E_e, t_o, t_p, and t_e are used here. They do not make estimations of T_a. Most probably, because it is very difficult to estimate T_a with a DE-method. Since an LF-model is not complete without T_a, another method had to be used to estimate T_a. Estimates of T_a were obtained by fitting the second phase of the LF-model to the return phase of the glottal pulse. Therefore, strictly speaking, only E_e, t_o, t_p, and t_e can be said to be the result of the DE-method. A complete description of the routine used for the DE-method can be found in Strik (1996).

2.3 Fit estimation methods

Voice source parameters can also be obtained by fitting a voice source model to the data. In the current research the LF-model (Fant *et al.*, 1985) is used as voice source model. Because in estimation methods of this kind a model fitting procedure is used, they will be referred to as 'fit estimation' methods (FE-methods).

It should be noted that the LF-model is a mathematically complex model, which is a disadvantage for a model used in a fit procedure. Nevertheless, I have chosen to use the LF-model, because this disadvantage is not crucial (its main effect is that it increases the CPU-time), and because the LF-model also has a number of advantages:

- ❒ in previous research the LF-model has already been used to estimate voice source parameters, with manual or (semi-)automatic methods, and this research has shown that it is a suitable model for description of the voice source signal (see e.g Karlsson, 1992);
- ❒ previous research has also proven that the LF-model is suitable for speech synthesis (see e.g. Carlson *et al.*, 1989);
- ❒ the model and its behaviour are well known; and finally,
- ❒ it seems that the LF-model is capable of giving a satisfactory description of most glottal flow pulses (at least for the types of speech that have been studied so far with this model).

In the present research the FE-method is used to estimate 5 parameters for each pitch period: E_e, t_o, t_p, t_e, and T_a. The FE-method consists of three stages:

1. initial estimate
2. simplex search algorithm
3. Levenberg-Marquardt algorithm

To fit the LF-model to dU_g, non-linear optimization techniques are used. These techniques require an initial estimate, which is made in the first stage. A description of the method used for initial estimation is given in Strik *et al.* (1993) and Strik (1996). In the second stage of the FE-method the simplex search algorithm of Nelder and Mead (1964) is used. Of the several optimization algorithms that were tested, the simplex search algorithm usually came closer to the global minimum than the gradient algorithms. Probably, discontinuities in the error function cause the gradient algorithms to get stuck in local minima more often than the simplex search algorithm does. However, in the neighbourhood of a minimum, the simplex algorithm may do worse (see e.g. Nelder and Mead, 1964). As a final optimization, the Levenberg-Marquardt algorithm (a gradient algorithm) is therefore used in the third stage.

3. Evaluation method and material

The performance of the DE-method and the FE-method, proposed above, was investigated in a systematic way, by independently testing the effect of several factors. These factors can roughly be divided into two groups. The factors in the first group are all properties of the pulses themselves, like e.g. exact position of the pulses, E_e, F_s, and B_c. The second group of factors consists of disturbances that are often present in recorded natural flow signals, like e.g. noise, formant ripple, and disturbances caused by low-pass filtering and phase distortion. In the current article only the results for three of these factors are presented, viz. position (shift), E_e, and low-pass filtering. The results of other tests can be found in Strik *et al.* (1993), Strik and Boves (1994), Strik (1994; 1996).

Many of the factors mentioned above will simultaneously influence the estimated voice source parameters. However, I think that the influence of each factor should be studied in isolation. First of all, because otherwise one will never know what the effect of each separate factor is. Second, because the relative size of each factor differs from one situation to the other. For instance, the magnitude of the disturbances mentioned above is certainly not constant for different experiments.

The question then was: What is the optimal way to study the effect of these factors? This is discussed in section 3.1 and 3.2 below.

3.1 Evaluation method

A first problem was the evaluation of the results. For natural speech the input (i.e. the voice source parameters) is unknown, and thus the estimated voice source parameters cannot be compared with the input values. Furthermore, for natural speech it is impossible to study the effect of most factors (mentioned above) in isolation. Therefore, I decided not to use natural speech, but to use synthetic glottal pulses in the following way.

First 11 base pulses were defined (see section 3.2). These 11 base pulses served as a starting point, and 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 x 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 off towards the nearest integer (as is done in standard A/D-conversion). Next, the voice source parameters were estimated with the DE-method and the FE-method, the resulting values were compared with the input values, and the errors were calculated:

$$\mathrm{ERR}(X) = 100\%*\mathrm{abs}(X_{est} - X_{inp})/X_{inp}, \text{ for } X = E_e$$

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

The experiments were carried out for a number (say N) of test pulses. After calculating the errors in the estimations of the 5 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. The absolute values were taken because otherwise positive and negative errors could cancel each other out. In that way the average error could be small, while the individual errors are (much) larger. The median was taken because (compared to the arithmetic mean) it is less affected by outliers that are sometimes present in the estimations. This method of averaging is the default method in the current article. Sometimes other ways of averaging were required. 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. Figure 3). In the upper left corner are the errors for E_e (in %), in the middle row are the errors for t_o and t_p, and in the bottom row are the errors for t_e and T_a. The errors in the time parameters t_o, t_p, t_e, and T_a are expressed in µsec. or in msec., depending on the magnitude of the errors.

3.2 Material

The effects of perturbations cannot always be studied by a single, isolated LF-pulse. For instance, a formant ripple will be present in dU_g when formant and bandwidth values are not estimated correctly in the inverse filtering procedure. To calculate the first samples of dU_g for the current pitch period, the speech signal resulting from the previous excitation is used. This speech signal is dependent on the amplitude and the shape of the previous flow pulse. Thus, the formant ripple at the beginning of the current pitch period will depend on the amplitude and the shape of the flow pulse in the previous pitch period. Therefore, I used sequences of three LF-pulses. Each time the voice source model was fitted to the (perturbed) pulse in the middle.

Furthermore, LF-pulses with different shapes were used. The reason is that the effect of a studied factor can depend on the shape of a pulse. Therefore, to get a general picture of the effect of that factor, the effect has to be studied for a number of pulses with different shapes. 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_o, and t_c were kept constant at 1024, 10 msec., 10 msec., and 20 msec., respectively. The values given for t_o and t_c are the values for the second of the three pulses. For the first pulse one should subtract 10 msec., and for the last pulse add 10 msec. T_0 and t_c were kept constant because varying these parameters had very little effect on the estimations. The influence of varying E_e and position (shift, which is more or less the same as t_o) were studied separately (see section 4.2).

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; Strik *et al.*, 1993; Strik and Boves, 1994; Strik, 1994) the following 11 base pulses were defined:

Table 1. Values of t_p, t_e, and T_a 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

For all tests F_s = 10 kHz and B_c = 12. If B_c = 12, the minimum value a sample can have is -2048, and thus the maximum value E_e can have is 2048. But even if the amplification during A/D conversion is optimal, the average value of E_e will be smaller than the maximum value of 2048. Therefore, the 11 base pulses were calculated with a value of 1024 for E_e.

4. Tests

Various tests were performed to test the DE-method and the FE-method. The results of some of these tests are presented in this article. First, the LF-routine used to generate the LF-pulses is tested in section 4.1. Subsequently, the influence of the amplitude (E_e) and the position (shift) of the pulse on the estimates is tested in section 4.2. Finally, in section 4.3 it is studied in which way low-pass filtering affects the estimations.

4.1 The LF-routine

4.1.1 Introduction

An important part of the FE-method is the error-function, and an essential part of the error-function is the routine used to calculate the voice source signal. Since I have used the LF-model to obtain a model fit of dU_g, this part of the FE-method is called the LF-routine. Because of the importance of this LF-routine for the FE-method, I shall first test the LF-routine in the current section (i.e. section 4.1).

In section 2.2 I argued that one of the drawbacks of the DE-methods is that only integer values for the parameters can be estimated. My intention was to develop an FE-method that would make it possible to estimate non-integer values too. In order to make this possible an LF-routine is needed which has a certain property: the LF-routine should be able to calculate correct LF-pulses for integer and non-integer values of the LF-parameters. Here I shall test whether my LF-routine has the required property, which will be called the ‘non-integer’ property below.

4.1.2 Method

A 10 kHz LF-pulse was calculated for the following values of the LF-parameters (which are not all integer): t_o = 10.05, t_p = 15.25, t_e = 17.25, t_c = 20.05, T_a = 1.0 msec., and E_e = 1.0. For this 10 kHz pulse all important events (i.e. t_o = opening, t_p = peak of U_g, t_e = excitation, and t_c = closing) are positioned exactly halfway between two sample positions. Next, a 20 kHz LF-pulse was calculated with the same values of the LF-parameters. In this case, all events coincide with sample positions.

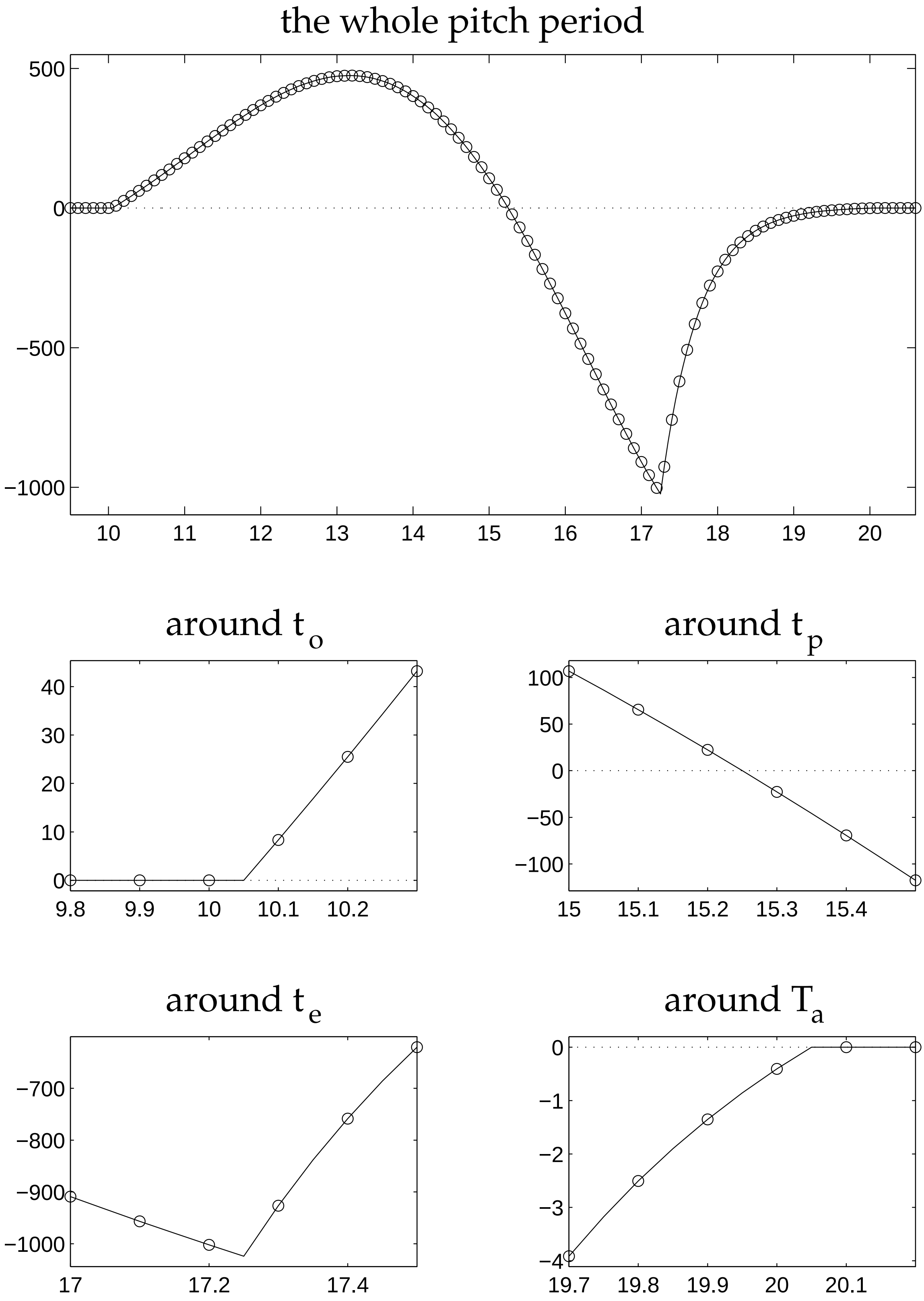


Figure 2. A 10 kHz (dotted) and a 20 kHz (solid) LF-pulse. Shown are the whole pitch period, and some details around important events.

4.1.3 Results and conclusions

As is apparent from Figure 2, the two pulses do not differ. A similar test was also performed for non-integer values of E_e, and different values of B_c (number of bits used for coding). In that case, too, the pulses did not differ. Therefore, the conclusion is that the proposed LF-routine succeeds in generating correct LF-pulses, also for non-integer values of the time and amplitude parameters. Results of ensuing tests can be found in the next subsection.

At this point it may seem more or less trivial to some readers that the LF-routine has the ‘non-integer’ property. However, this is not the case. For instance, the LF-routine I used first, i.e. the LF-routine described in Lin (1990), did not have the ‘non-integer’ property. The reason is that in Lin's routine all the input parameters are rounded off towards the nearest integer. Because Lin (1990) used his routine for speech synthesis, rounding off the input parameters was not a serious drawback for his application. For many implementations of a voice source model, rounding off the input seems a logical and practical operation.

In the current and the following subsection it is tested whether the LF-routine has the ‘non-integer’ property. These tests are presented here because I found that for the FE-method it is very important to use an LF-routine that has the ‘non-integer’ property. In fact, when the LF-routine used in my FE-method was changed from Lin's version to the current version, an enormous improvement was observed. Consequently, the errors in the estimates with the current version of the LF-routine are much smaller than those obtained with the previous (i.e. Lin's) version.

4.2 Shift and E_e

4.2.1 Introduction

In the previous subsection it was tested whether it is possible to calculate correct LF-pulses, with the proposed LF-routine, also for non-integer values of the LF-parameters. This was tested by studying some well-chosen examples of the LF-pulses. As the test gave positive results, I can now go one step further. In this section a more thorough test is presented. For both the DE-method and the FE-method it will be tested how well they succeed in estimating (non-integer values of) the voice source parameters.

4.2.2 Method

The definition of the 11 base pulses is such that all time parameters have an integer value (see section 3.2). 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 msec., from 0.0 up to 0.1 msec. (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 non-integer values. An example of a base pulse shifted over 0.05 msec. is the 10 kHz pulse in Figure 2 (dotted line).

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 x 11 shift values x 11 E_e values).

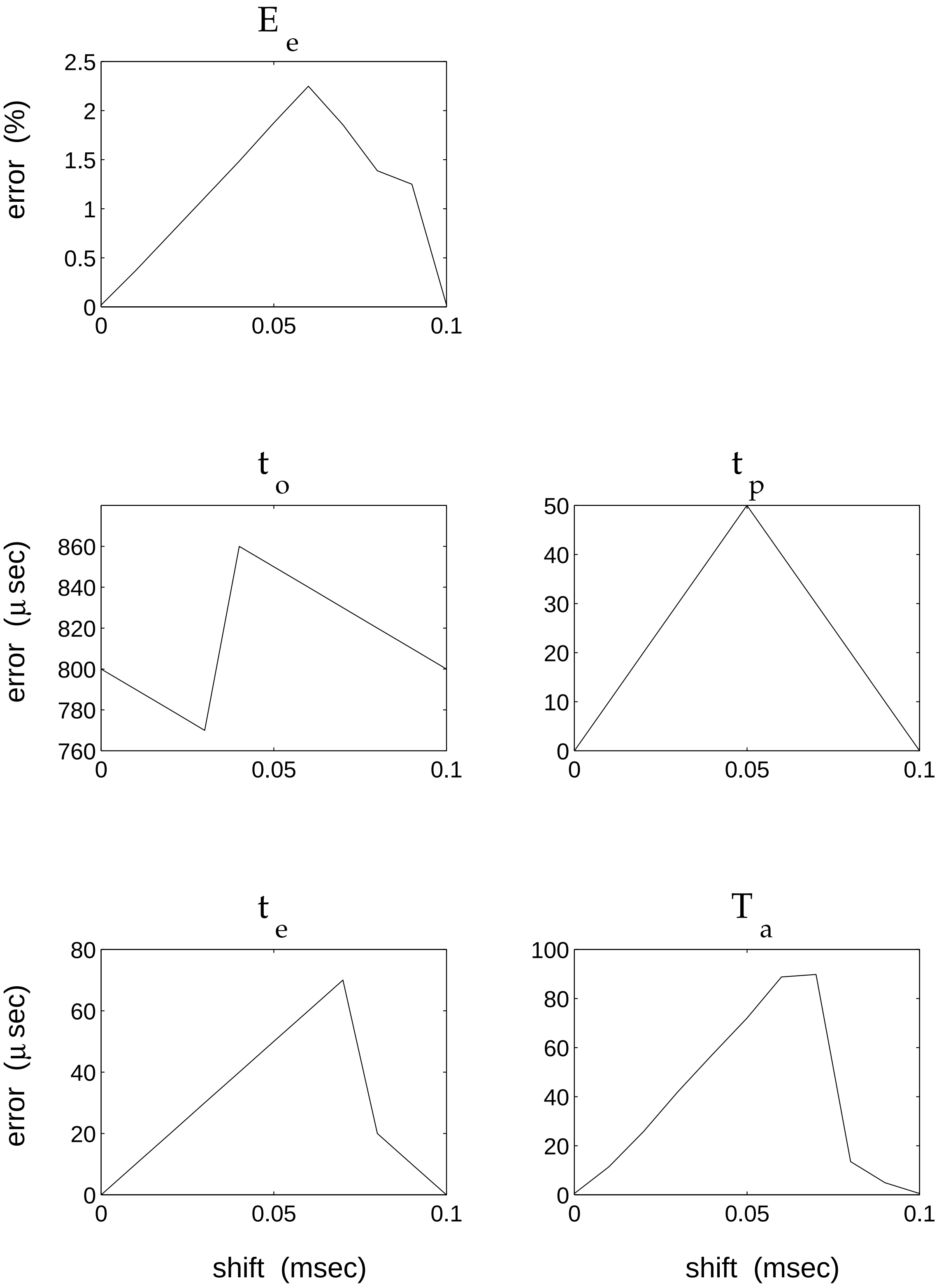


Figure 3. Median error for the estimated parameters for different values of shift.

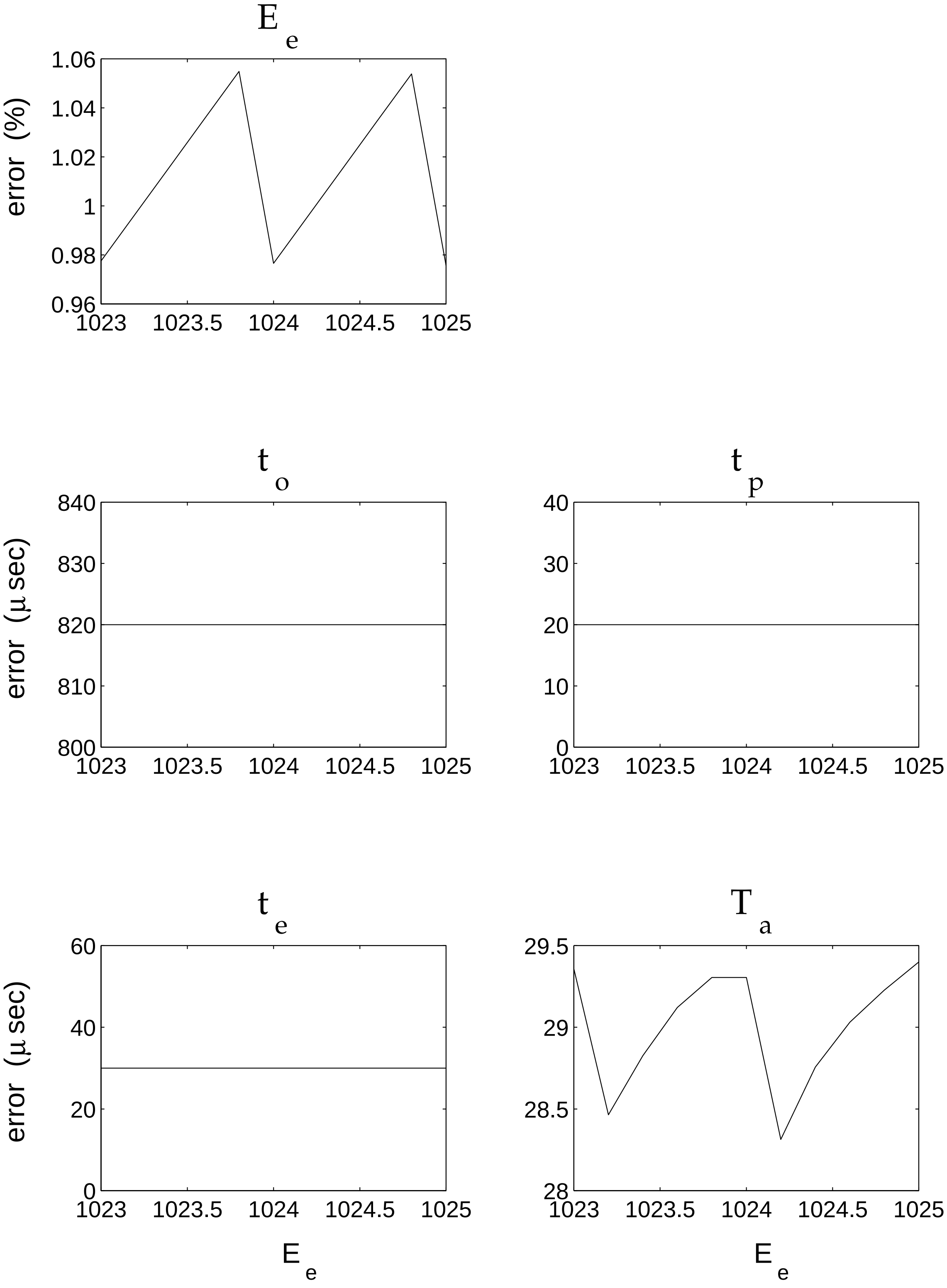


Figure 4. Median error for the estimated parameters for different values of E_e.

4.2.3 Results of the DE-method

First the results of the DE-method are presented in Figures 3 and 4. Each error in Figure 3 is the median of 121 errors (11 base pulses x 11 E_e values), while each error in Figure 4 is the median of another set of 121 errors (11 base pulses x 11 shift values).

Let us first look at the errors in Figure 3. To estimate t_o a threshold function is used in the DE-method. The consequence is that the estimate of t_o is always much too large (on the average about 820 μsec., see Figure 4). For a shift of 0.03 msec. the average error in t_o is minimal, while for a shift of 0.04 msec. it suddenly becomes maximal. The reason is that this extra shift of 0.01 msec. causes the threshold to be exceeded one sample later in many test pulses, and thus the average error in t_o suddenly increases.

Except for t_o, the figures of the average errors of the other parameters all have roughly the expected triangular shape. For a shift of 0.0 and 0.1 msec. the errors are zero, and for other shift values the errors are greater than zero. The fact that (except for t_p) the figures are not exactly triangular is caused by certain details of the implementation of the DE-method which are not relevant here.

The errors in the estimates for different values of E_e are shown in Figure 4. The errors in the time parameters t_o, t_p, and t_e obviously do not depend on the value of E_e. Therefore, the errors for these time parameters are constant. The figure of the errors in the estimates of E_e also has the expected triangular shape: 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. Again, the exact shapes of the figures with the errors of E_e and T_a are a corollary of details in the implementation of the DE-method which are not relevant.

4.2.4 Results of the FE-method

The resulting average errors for the FE-method are shown in Figures 5 and 6. 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 x 11 E_e values) were calculated, and the results are shown in Figure 5. Likewise, for each value of E_e the mean and standard deviation of 121 errors (11 base pulses x 11 shift values) were calculated, and the results are shown in Figure 6.

In Figures 5 and 6 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 errors for the DE-method. Except of course for the cases where all the LF-parameters have an integer value. In the latter case the errors for the DE-method are zero, which is smaller still than the tiny errors found for the FE-method. However, it is clear that in practice the voice source parameters will seldom have exactly an integer value.

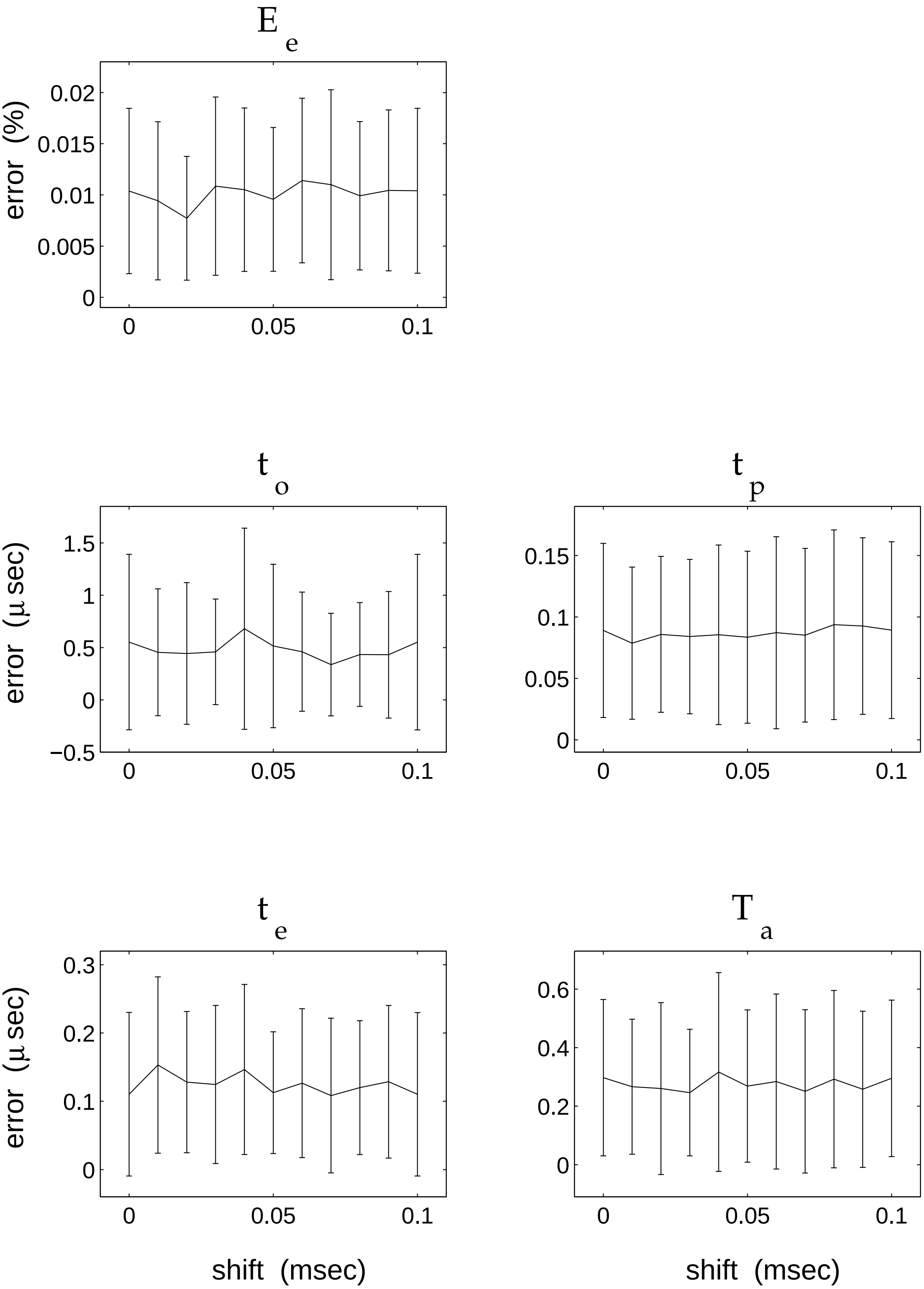


Figure 5. Mean and standard deviation of the errors in the estimated parameters for different values of shift.

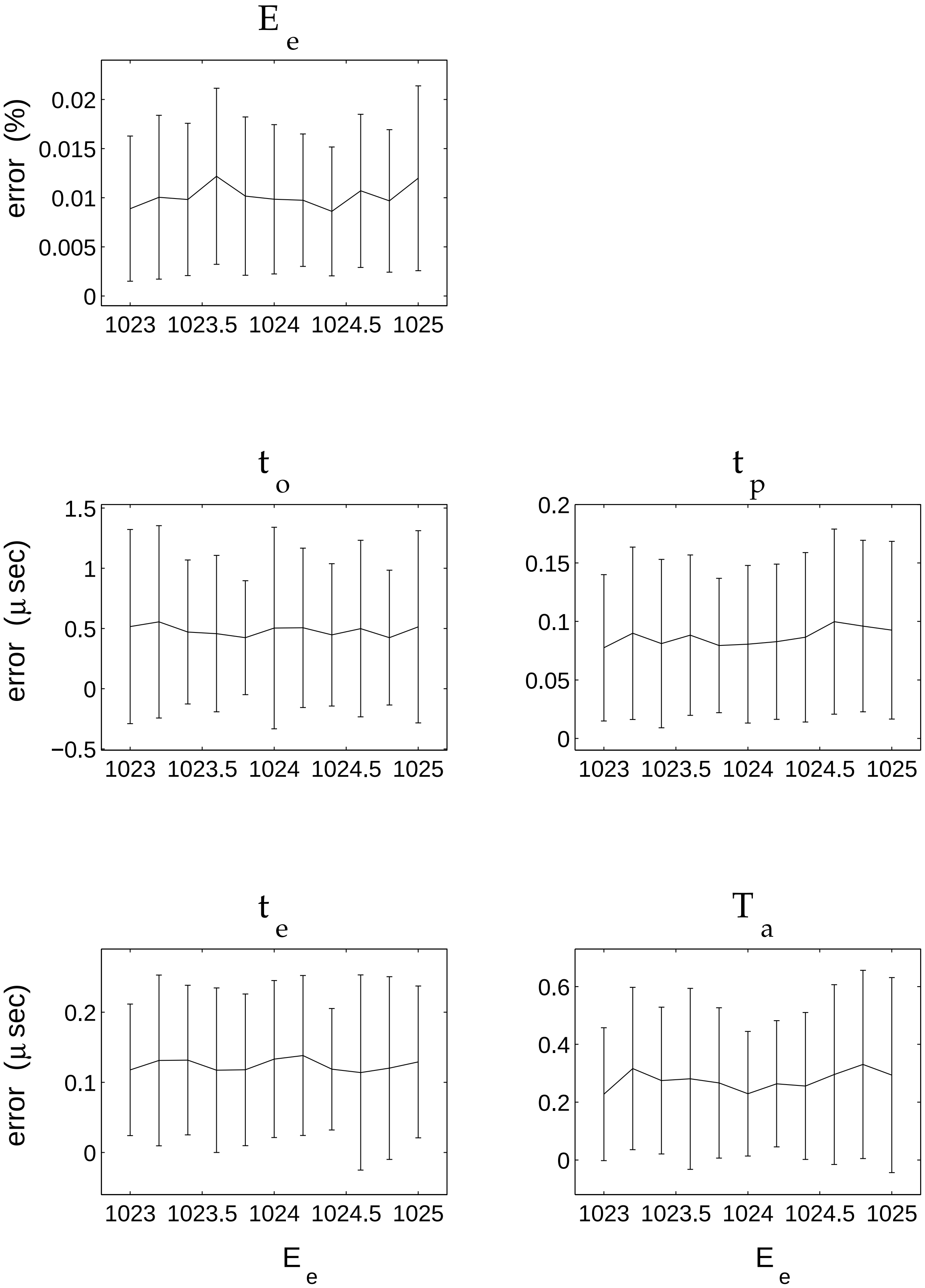


Figure 6. Mean and standard deviation of the errors in the estimated parameters for different values of E_e.

4.2.5 Conclusions

The conclusions that can be drawn from these tests are the following. For the DE-method the errors in t_o are always large in comparison to the average errors for the FE-method, because a threshold function is used to estimate t_o. For the other parameters the estimation errors are zero if the parameters have an integer value. As parameters will rarely have an integer value in practice, estimations of parameters with a DE-method will almost always contain an error due to this fact alone (even if the glottal pulses are perfect clean glottal pulses, as was the case in these tests).

The errors obtained with the FE-method are very small, much smaller than for the DE-method. It can be concluded that with the FE-method non-integer values can be estimated as accurately as integer values. Therefore, the quality of the 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_o and E_e could be kept constant in the definition of the base pulses (see section 3.2).

4.3 Low-pass filtering

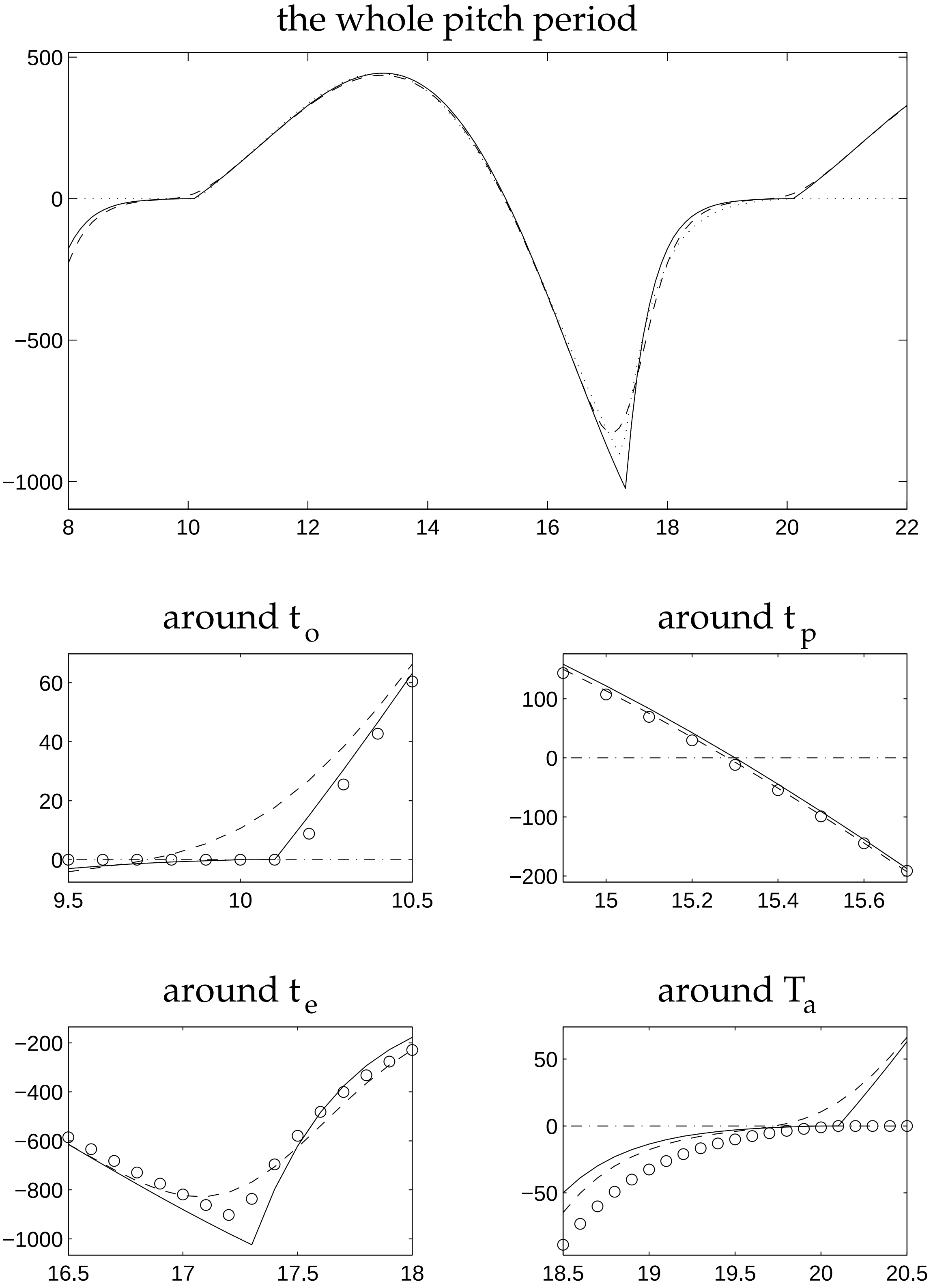
4.3.1 Introduction

Before the glottal flow signals are parameterized, 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 attenuate the noise component. The latter operation seems very sensible for DE-methods, because for these methods high-frequency disturbances can influence the estimated parameters to a large extent. However, low-pass filtering changes the shape of the glottal pulses, and, consequently, influences the estimated voice source parameters (see also Alku and Vilkman, 1995; and Strik, to appear).

An example of the distortion of a flow pulse caused by low-pass filtering is given in Figure 7. 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 fit on the low-pass filtered pulse (dotted). Besides a picture of the three signals for the whole pitch period, also some details around important events are provided. For this example the errors in the estimations obtained by means of the FE-method are: $\mathrm{Err}(E_e)$ = -11.2%, $\mathrm{Err}(t_o)$ = 46 µsec., $\mathrm{Err}(t_p)$ = -28 µsec., $\mathrm{Err}(t_e)$ = -52 µsec., and $\mathrm{Err}(T_a)$ = 144 µsec.

One can see in Figure 7 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 DE-method. This is most clear for the estimate of E_e, which will generally be too small. But also the estimates of the other parameters will be affected.

Low-pass filtering will also affect the estimates of an FE-method. After low-pass filtering the shape of the pulse is changed. The fit procedure will try to find an LF-pulse that resembles the filtered pulse as closely as possible. This is done by minimizing the RMS-value of the difference of the test pulse and the fitted LF-pulse. The result is a fitted LF-pulse that deviates from the original base pulse (see Figure 7). In Figures 7a and 7d it can be seen that the estimated values of E_e and t_e are too small, while the estimate of T_a is too large. Furthermore, one can see in Figure 7b that for this example pulse the estimate of t_o is too large, and in Figure 7c that the estimate of t_p is a bit too small.



the whole pitch period

around t_o

around t_p

around t_e

around T_a

Figure 7. An example of a flow pulse before (solid) and after (dashed) low-pass filtering, and a fit on the low-pass filtered pulse (dotted). Shown are the whole pitch period, and some details around important events.

The reader should be convinced by now, that low-pass filtering does affect the shape of the flow pulses, and consequently also the estimated parameters. In the present section the effect of low-pass filtering on the parameter estimates will be studied. The distortion of the glottal pulses depends on a numbers of factors, like e.g. the type and the bandwidth of the low-pass filter, the frequency contents of the glottal pulses, and the parameterization method used. I will study the effect of low-pass filtering for two parameterization methods (i.e. the DE-method and the FE-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. 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 some preliminary tests showed that the error in the estimations 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 the estimation errors will be larger than the errors presented below.

4.3.2 Method

The 11 base pulses were low-pass filtered by means of a convolution with a Blackman window of varying length. 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 x 9 window lengths) the parameters were estimated with the DE-method and the FE-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 Figures 8 and 9.

In the example provided in Figure 7 the test signal is low-pass filtered. An LF-model is then fitted to the low-pass filtered test pulse. This seems the most obvious way to apply low-pass filtering, and will be called the first version of the FE-method. However, there is an alternative (which will be called the second version of the FE-method): apart from the test pulse one could also low-pass filter the fitted LF-pulse. In that case, test pulse and fitted LF-pulse are altered in a similar fashion. In this way I 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 trick cannot be used in a DE-method, because in a DE-method the parameters are calculated directly from the (low-pass filtered) signal.

4.3.3. Results of the DE-method

In Figure 7a 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.

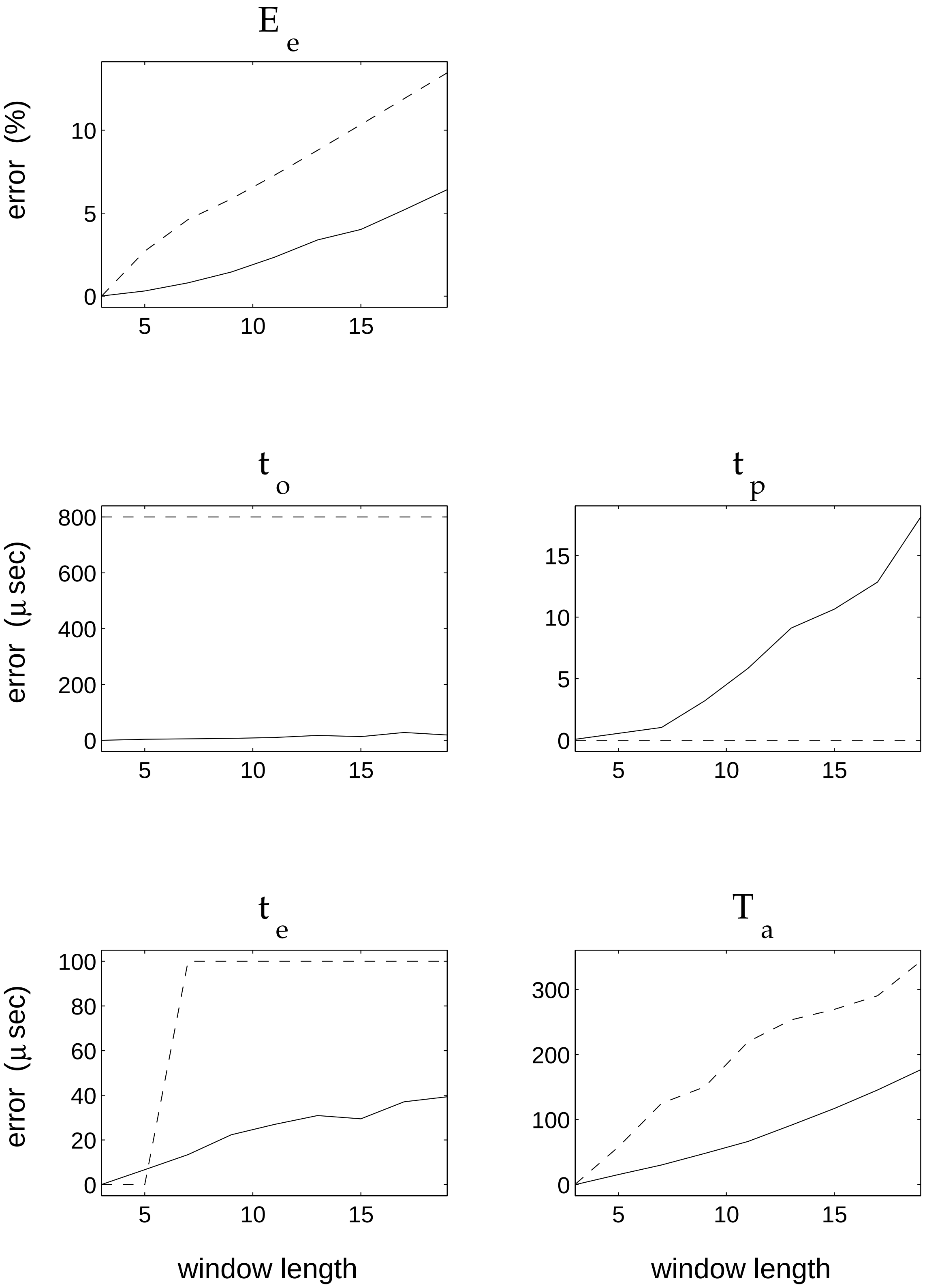


Figure 8. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the DE-method (dashed) and for the first version of the FE-method (solid).

Low-pass filtering does not have much influence on t_p (= the position of the zero-crossing in dU_g, see Figure 7c). Therefore, in the majority of the cases the error in the estimations 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 Figure 7d). 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, and as a result also the median error becomes larger.

Finally, the error in t_o remains constant, at the value of 820 μsec. (see also Figure 4). This can be explained with the help of Figure 7a and 7b. In these figures one can see that low-pass filtering has a large effect on the signal in the direct neighbourhood of t_o, and that this effect diminishes away from t_o. If the threshold is chosen high enough (which is the case in the DE-method), low-pass filtering will not have much influence on this estimate of t_o.

Here, I would like to repeat a remark made in the introduction of this subsection. The low-pass filters used in these tests have ripple-free impulse responses, and are chosen because their effect on the estimates is smaller than for most other low-pass filter. Therefore, it is most likely that for other low-pass filters the errors will be larger. Especially if a low-pass filter with a ripple in its impulse response is used, the errors for a DE-method will be much larger (see e.g. Strik, to appear).

4.3.4. Results of the FE-method

In Figure 8 not only the errors of the DE-method are presented, but also those of the first version of the FE-method (i.e. the version in which only the test pulses were low-pass filtered). If the median errors of the FE-method are compared with those of the DE-method, the following observations can be made:

- ❐ 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.
- ❐ In all other cases the errors of the first version of the FE-method are smaller than those of the DE-method.

The fact that in certain cases the error of the DE-method is smaller than the error of the FE-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 DE-method is zero, while that of the FE-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 that case does rounding off towards the nearest sample position mean rounding off towards the correct value. In practice, events almost never fall exactly on a sample position, and in section 4.2 we saw that this leads to substantial errors for the DE-method, and much smaller errors for the FE-method. Because I decided to study each phenomenon separately, the events of the test pulses used in this subsection coincide exactly with the corresponding sample point. Consequently, the errors of the DE-method are sometimes smaller than those of the FE-method. If the important events had been positioned randomly, the errors of the FE-method would have been slightly larger while those of the DE-method would have been substantially larger (see section 4.2). Therefore, for a realistic comparison of the errors obtained with the two estima-

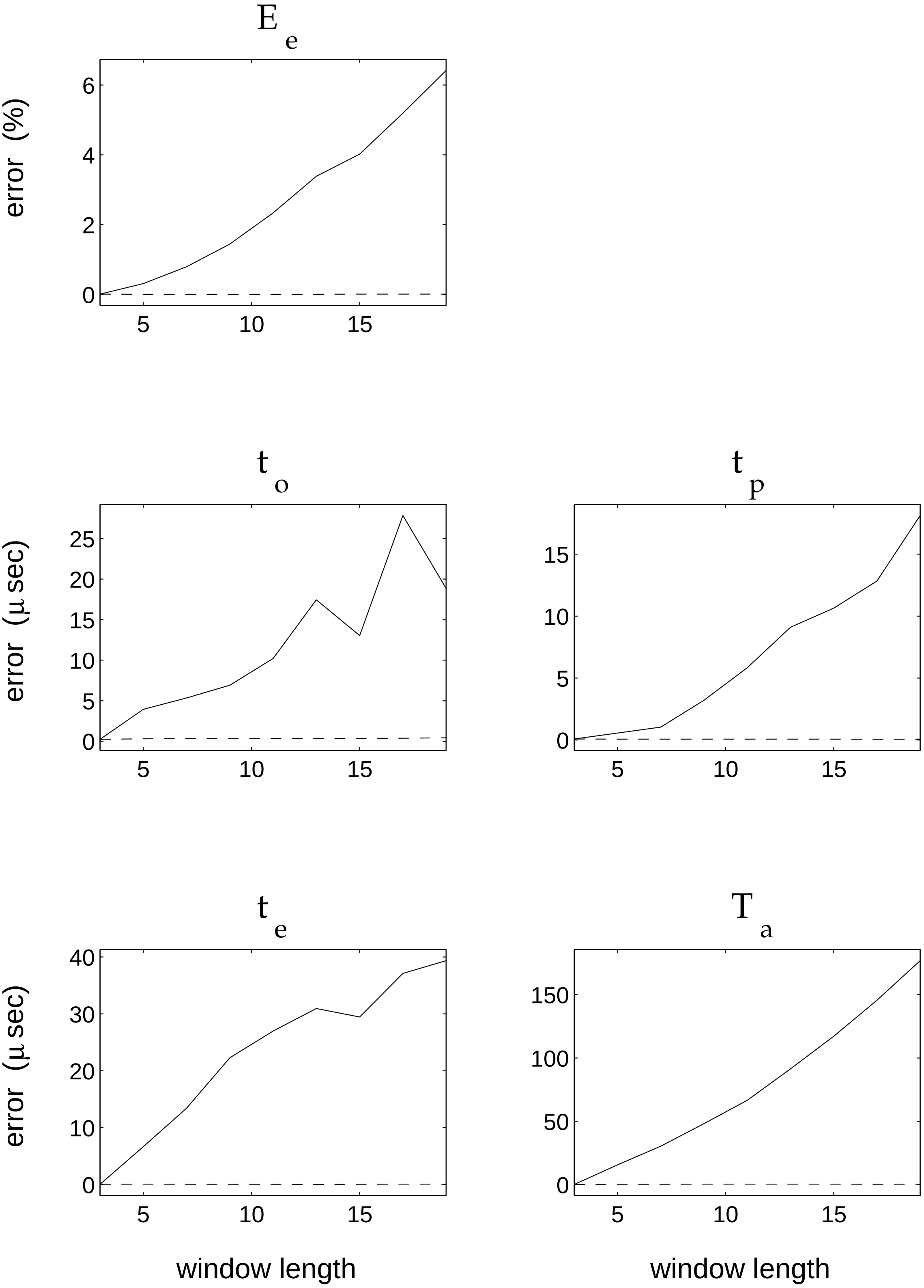


Figure 9. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the first (solid) and the second (dashed) version of the FE-method.

tion methods (for a certain studied factor, here low-pass filtering), these errors should be increased with the errors found in section 4.2 for both methods. If this is done the average errors of the DE-method are always larger than those of the FE-method.

In Figure 9 the results of the two versions of the FE-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 Figure 9, but they are extremely small. The largest error observed in the time parameters is 1 μsec., and the errors in E_e are always smaller than 0.03%.

4.3.5 Conclusions

In previous sections I explained why with the used test pulses the errors in the DE-method are sometimes smaller than those of the first version of the FE-method. However, for a realistic comparison the errors found in section 4.2 should be added. In that case the errors for the DE-method are always larger than those of the first version of the FE-method. In turn, these errors are larger than the errors of the second version of the FE-method. Therefore, the conclusion is that the second version of the FE-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 DE-method (as was already noted above).

5. General conclusions

In reality the value of voice source parameters will not exactly be integer, i.e. they can have all kind of non-integer values. Because in DE-methods the estimates of the parameters are limited to integer values, these estimates contain an intrinsic error (even if all other conditions are perfect). The FE-method proposed in this article does not have this drawback, and is also capable of estimating non-integer values. In fact, errors of a similar magnitude were found for estimates of integer and non-integer parameter values. Therefore, the average errors for various values of shift and E_e obtained with the FE-method are smaller than those of the DE-method.

Subsequently, the effect of the factor low-pass filtering was studied in isolation, i.e. independently of the other factors (like shift and E_e). For a realistic comparison of the different estimation methods, their intrinsic errors (as given in section 4.2) should be added to the errors found for low-pass filtering alone, as was explained above. If this is done, the methods can be arranged in order of decreasing average error: DE-method, first version of FE-method, and second version of FE-method. The conclusion that can be drawn on the basis of the tests presented in this article is that the second version of the FE-method is superior.

References

Alku, P. (1992), ‘An automatic method to estimate the time-based parameters of the glottal pulseform’, in: *Proceedings ICASSP’92*, 29-32.

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

Ananthapadmanabha, T.V. (1984), ‘Acoustic analysis of voice source dynamics’, *Speech Transmission Laboratory, Q. Prog. Status Rep.*, Royal Institute of Technology, Stockholm, 2-3/1984, 1-24.

Carlson, R., G. Fant, C. Gobl, B. Granstrom, I. Karlsson and Q. Lin (1989), ‘Voice source rules for text-to-speech synthesis’, in: *Proceedings ICASSP’89*, 223-226.

Fant, G. (1993), ‘Some problems in voice source analysis’, *Speech Communication*, 13, 7-22.

Fant, G., J. Liljencrants and Q. Lin (1985), ‘A four parameter model of glottal flow’, *Speech Transmission Laboratory, Q. Prog. Status Rep.*, Royal Institute of Technology, Stockholm, 4/1985, 1-13.

Fritzel, B. (1992), ‘Inverse filtering’, *Journal of Voice*, 6, 111-114.

Gauffin, J. and J. Sundberg (1980), ‘Data on the glottal voice source behavior in vowel production’, *Speech Transmission Laboratory, Q. Prog. Status Rep.*, Royal Institute of Technology, Stockholm, 2-3/1980, 61-70.

Gauffin, J. and J. Sundberg (1989), ‘Spectral correlates of glottal voice source waveform characteristics’, *Journal of Speech and Hearing Research*, 32, 556-565.

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

Koreman, J. (1996), *Decoding linguistic information in the glottal airflow*, PhD dissertation, Nijmegen University.

Lin, Q. (1990), *Speech production theory and articulatory speech synthesis*, Unpublished PhD dissertation, KTH, Stockholm.

Nelder, J.A. and R. Mead (1964), ‘A simplex method for function minimization’, *The Computer Journal*, 7, 308-313.

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

Strik, H. (1994), *Physiological control and behaviour of the voice source in the production of prosody*, PhD dissertation, University of Nijmegen.

Strik, H. (1996), *Factors affecting the error in estimated voice source parameters*, Internal report. Department of Language and Speech, University of Nijmegen.

Strik, H. (to appear), ‘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.* (accepted).

Strik, H. and L. Boves (1992), ‘On the relation between voice source parameters and prosodic features in connected speech’, *Speech Communication*, 11, 167-174.

Strik, H. and L. Boves (1994), ‘Automatic estimation of voice source parameters’, in: *Proceedings International Conference on Spoken Language Processing (ICSLP) ’94*, Yokohama, 155-158.

Strik, H., B. Cranen and L. Boves (1993), ‘Fitting a LF-model to inverse filter signals’, in: ESCA *3rd European Conference on Speech Communication and Technology: EUROSPEECH ‘93*, Berlin, 103-106.

Strik, H., J. Jansen and L. Boves (1992), ‘Comparing methods for automatic extraction of voice source parameters from continuous speech’, in: *Proceedings International Conference on Spoken Language Processing (ICSLP) ’92*, Banff, 121-124.

Sundberg, J. and J. Gauffin (1979), ‘Waveforms and spectrum of the glottal voice source’, in: Lindblom, B. and S. Öhman (eds.), *Frontiers of speech communication research*, Festschrift for Gunnar Fant. London: Academic Press, 301-320.