Automatic parametrization of voice source signals: a novel evaluation procedure is used to compare methods and test the effect of low-pass filtering

Helmer Strik

Internal Report [1]
June 1997
Dept. of Language and Speech
University of Nijmegen
P.O. Box 9103
6500 HD Nijmegen
the Netherlands

strik@let.kun.nl
http://lands.let.kun.nl/TSpublic/strik

ABSTRACT

There is a need for automatic methods for parametrization of the voice source signals. Representatives of the two types of methods that have been used most often for parametrization were tested and compared. For this purpose a novel evaluation procedure is proposed which makes it possible to perform the numerous tests needed for a detailed comparison of the methods. 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 non-integer 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. The factor low-pass filtering was chosen because low-pass filtering is probably used in all methods in which voice source parameters are estimated. 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 voice source signals, and in which the voice source model is low-pass filtered with the same filter as the voice source signals.

1. INTRODUCTION

The technique of inverse filtering has been available for a long time now. It was first described in Miller (1959). Inverse filtering is based on the linear source-filter model of speech production (Fant, 1960; Flanagan, 1965).

The signal that is inverse filtered most often is the acoustic sound pressure wave recorded with a microphone placed a few centimeters in front of the mouth. In this way an estimate of the first derivative of glottal flow (dU_g) can be obtained (Miller, 1959; Rosenberg, 1971; Strube, 1974; Gauffin and Sundberg, 1980; Schoentgen, 1990, 1995; de Veth *et al.*, 1990; Jansen *et al.*, 1991; Alku, 1992; Karlsson, 1992; Strik and Boves, 1992a, 1992b; Fant, 1993). Subsequently, the lip radiation effect can be canceled by integrating dU_g to obtain an estimate of true glottal flow (U_g).

It is also possible to inverse filter the airflow signal recorded at the lips to calculate U_g (e.g. Rothenberg, 1973, 1977; Sundberg and Gauffin, 1979; Holmberg, 1993; Hertegard, 1994; Koreman, 1996). This type of research has shown a strong increase since the introduction of the so called Rothenberg mask (Rothenberg, 1973, 1977), which consists of a differential pressure transducer attached to a circumferentially vented face mask covered with a wire mesh. The frequency response of this system is flat up to about 1.5 kHz (Sundberg and Gauffin, 1979; Gauffin and Sundberg, 1989; Hertegard and Gauffin, 1992). The mask is usually held as tightly as possible against the subject's face, in order to ensure a tight seal between face and mask.

Inverse filtering has already been studied extensively, and many different methods have been proposed in the literature (see e.g. Funaki and Mitome, 1990; Alku and Vilkman, 1994; Hong *et al.*, 1994; Ding and Kasuya, 1996). However, estimating a voice source signal (either dU_g or U_g) is usually not enough. For many applications it is necessary to parametrize the glottal flow signals. Parametrization of the voice source signals, and evaluation of these parametrization methods, has received far less attention in the past. That is why we focus on these aspects in this study.

Parametrization of dU_g or U_g can be done in several ways. Usually landmarks (like minima, maxima, zero crossings) are detected in the signals (e.g. Sundberg and Gauffin, 1979; Gauffin and Sundberg, 1980; Gauffin and Sundberg, 1989; Alku, 1992; Strik and Boves, 1992a; Holmberg, 1993; 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 (DE methods).

Voice source parameters are also calculated by fitting a voice source model to the data (e.g. Ananthapadmanabha, 1984; Karlsson, 1990; Schoentgen, 1990, 1995; Jansen *et al.*, 1991; Karlsson, 1992; Strik and Boves, 1992b; Fant, 1993; Milenkovic, 1993; Riegelsberger and Krisnamurthy, 1993). Because in estimation methods of this kind a model fitting procedure is used, they will be referred to as 'fit estimation' methods (FE methods).

In an FE method the voice source model is essential. Many different models have been proposed in the literature (see e.g. Rosenberg, 1971; Fant, 1979; Ananthapadmanabha, 1984; Fant *et al.*, 1985, Fujisaki and Ljungqvist, 1986; Funaki and Mitome, 1990; Lobo and Ainsworth, 1992; Hong *et al.*, 1994; Cummings and Clements, 1995). The Liljencrants-Fant model (LF model) (Fant *et al.*, 1985) is used most often as the voice source model (e.g. Jansen *et al.*, 1991; Karlsson, 1992; Strik and Boves, 1992a, 1992b; Fant, 1993; Riegelsberger and Krisnamurthy, 1993). Since a voice source model is not required in a DE method, some studies do not use it (e.g. Sundberg and Gauffin, 1979; Gauffin and Sundberg, 1980; Gauffin and Sundberg, 1989; Alku, 1992; Holmberg, 1993; Alku and Vilkman, 1995). However, other studies based on DE methods do use a voice source model (Rosenberg, 1971; Fujisaki and Ljungqvist, 1986; Gobl, 1988; Lobo and Ainsworth, 1992; Strik and Boves, 1992a; Koreman, 1996). An important reason for using a voice source model is that the estimated voice source parameters can be subsequently used for speech synthesis.

The parametrization is usually done in the time domain (e.g. Ananthapadmanabha, 1984; Fujisaki and Ljungqvist, 1986; Schoentgen, 1990, 1995; Jansen *et al.*, 1991; Strik and Boves, 1992b; Milenkovic, 1993; Riegelsberger and Krisnamurthy, 1993), sometimes simultaneously in time and frequency domain (e.g. Gobl and Ní Chasaide, 1988; Karlsson, 1990, 1992; Fant, 1993; Ní Chasaide and Gobl, 1990, 1993), and occasionally in the frequency domain alone (Funaki and Mitome, 1990; Hong *et al.*, 1994; Alku and Vilkman, 1996; Ding and Kasuya, 1996; Alku, Strik and Vilkman, to appear). What the optimal domain is depends on the application and the method used.

Besides the method used to estimate the voice source parameters, it is important to have a look at the method and material used for evaluation. The analyzed material is often limited to a small number of pitch periods of vowels; most often natural vowels (e.g. Fujisaki and Ljungqvist, 1986; Funaki and Mitome, 1990; Jansen *et al.*, 1991; Holmberg, 1993; Milenkovic, 1993), sometimes synthetic vowels (e.g. Strik and Boves, 1994; Darsinos *et al.*, 1995), or both (e.g. Strube, 1974; Alku, 1992; Strik *et al.*, 1992, 1993; Riegelsberger and Krisnamurthy, 1993). Furthermore, the analyzed material usually consists of sustained vowels or carefully produced (e.g. read) utterances. Hardly ever were voice source parameters estimated for all pitch periods of a complete spontaneous sentence. As far as we know the only exception is Strik and Boves (1992a, 1992b). Three important reasons why the material is often limited to a small number of pitch periods of sustained or carefully produced vowels are:
[1] because almost none of the methods is automatic, it is too laborious to process large amounts of speech,

[2] inverse filtering and parametrization are generally easier for vowels than for consonants, and
[3] they are usually more difficult for spontaneous speech compared to sustained vowels or carefully produced utterances.

If voice source parameters are estimated only for a limited number of pitch periods of vowels, there is not much material that can be used for evaluation of the proposed methods. This is one of the reasons why a thorough evaluation generally is not provided. In fact, in the majority of the articles no evaluation is presented at all. In the few cases in which an evaluation was provided, it often consisted merely of a simple qualitative (usually visual) comparison of the glottal flow signals and the model fits for a small number of pitch periods of vowels (Strube, 1974; Fant *et al.*, 1985; Lobo and Ainsworth, 1992; Riegelsberger and Krisnamurthy, 1993; Hong *et al.*, 1994; Ding and Kasuya, 1996).

This qualitative evaluation was generally done for natural speech (Fant *et al.*, 1985; Lobo and Ainsworth, 1992; Hong *et al.*, 1994; Ding and Kasuya, 1996), although it is also possible to use synthetic speech for evaluation (e.g. Strik *et al.*, 1992, 1993; Strik and Boves, 1994; Darsinos *et al.*, 1995). Natural speech has the advantage that it is the kind of speech the method will be used for eventually. However, an important drawback of natural speech is that the correct voice source parameters are not known [2]. This makes it hard to perform a quantitative and detailed evaluation of the estimated voice source parameters. On the other hand, for synthetic speech the correct voice source parameters are known: they are simply the voice source parameters used during synthesis, or some transformation of these parameters. They can be used to calculate the error in the estimated parameters (e.g. Strik *et al.*, 1992, 1993; Strik and Boves, 1994; Darsinos *et al.*, 1995). A drawback of synthetic speech is that it (usually) does not contain all effects (especially the non-linear effects) that are present in natural speech.

We think that evaluation of the estimation methods is important, and therefore should get more attention than it has received so far. That is why we elaborate on this topic in the current article.

Estimation of voice source parameters can be useful for many applications. Without doubt, the application mentioned most often is speech synthesis. However, the estimated voice source parameters are also used for fundamental research on speech production (Ní Chasaide and Gobl, 1993; Holmberg *et al.*, 1994; Strik, 1994; Koreman, 1996). Other areas in 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. However, in order to be applicable in these areas the methods should be fully automatic. Also for research on speech synthesis and fundamental research on speech production the use of automatic methods would be advantageous. Thus, for various reasons there is an increasing need for automatic methods (see e.g. Fritzel, 1992; Fant, 1993; Ní Chasaide and Gobl, 1993; Ding and Kasuya, 1996). Although a lot of research has already been carried out on this topic, a completely automatic method that works satisfactorily does not seem to exist yet.

The long term goal of our research therefore is to develop such an automatic method. Both DE methods and FE methods can be made completely automatic. For this reason, and because they are the methods used most often, a representative of the DE method will be compared with a representative of the FE method. The representatives chosen are described in section 2.3 and 2.4.

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 LF model. The LF model and the reasons for choosing it are described in section 2.2.

To achieve these goals we tried to develop an evaluation procedure with which it is possible to make a thorough and systematic evaluation. The method and material chosen for evaluation are described in sections 3.1 and 3.2, respectively.

This evaluation procedure is then used to study voice source estimation. First, in section 4.1 and 4.2, it is studied how well the estimation methods succeed in estimating non-integer values of the parameters. This turned out to be a very crucial property of the estimation methods.

The evaluation procedure proposed in section 3 can be used to study the effect of different factors. As an example we have chosen to study the effect of low-pass filtering (see section 4.3). The reason is that low-pass filtering influences the estimated parameters (Strik *et al.*, 1992, 1993; Perkell *et al.*, 1994; Alku and Vilkman, 1995; Strik, 1996a; Koreman, 1996). Because low-pass filtering is used in (almost) all methods, it becomes very important to study what the effect of low-pass filtering exactly is. Previously proposed methods are not optimally suited for this task (see Strik, 1996a). We will show that the evaluation procedure proposed here is suitable for studying the effect of the factor low-pass filtering.

In section 5 the findings are discussed and some general conclusions are drawn.

2. ESTIMATION METHODS

In this article two estimation methods used to parametrize dU_g are tested and compared. Before going on to describe these two methods (in sections 2.3 and 2.4), we shall first give some definitions in section 2.1 and describe the LF model in section 2.2.

2.1. Some definitions

In the current article it will be assumed that dU_g is a discrete signal. Some terms related to these voice source signals, and the A/D conversion used to obtain them, are often used below. In order to avoid confusion later on, we shall first define some of these terms in this section.

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). Here, F_s = 10 kHz, Δ = [-2048, 2047], and B_c = 12. 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 neighboring 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.

2.2. LF model

In the current research the LF model is used as voice source model (see Figure 1). It should be noted that the LF model is a mathematically complex model, which is a disadvantage for a model used in a fitting procedure. Nevertheless, we 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 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 voice source signal (see e.g Fujisaki and Ljungqvist, 1986; Jansen *et al.*, 1991; Karlsson, 1992; Strik and Boves, 1992b; Strik *et al.*, 1992, 1993; Riegelsberger and Krisnamurthy, 1993; Childers and Ahn, 1995; Darsinos *et al.*, 1995).
- Fujisaki and Ljungqvist (1986) compared several voice source models. Their results showed that the LF model and their own FL-4 model performed best.
- Previous research has also proven that the LF model is suitable for speech synthesis (see e.g. Carlson *et al.*, 1989).
- Due to all research already performed, the model and its behavior are well known.

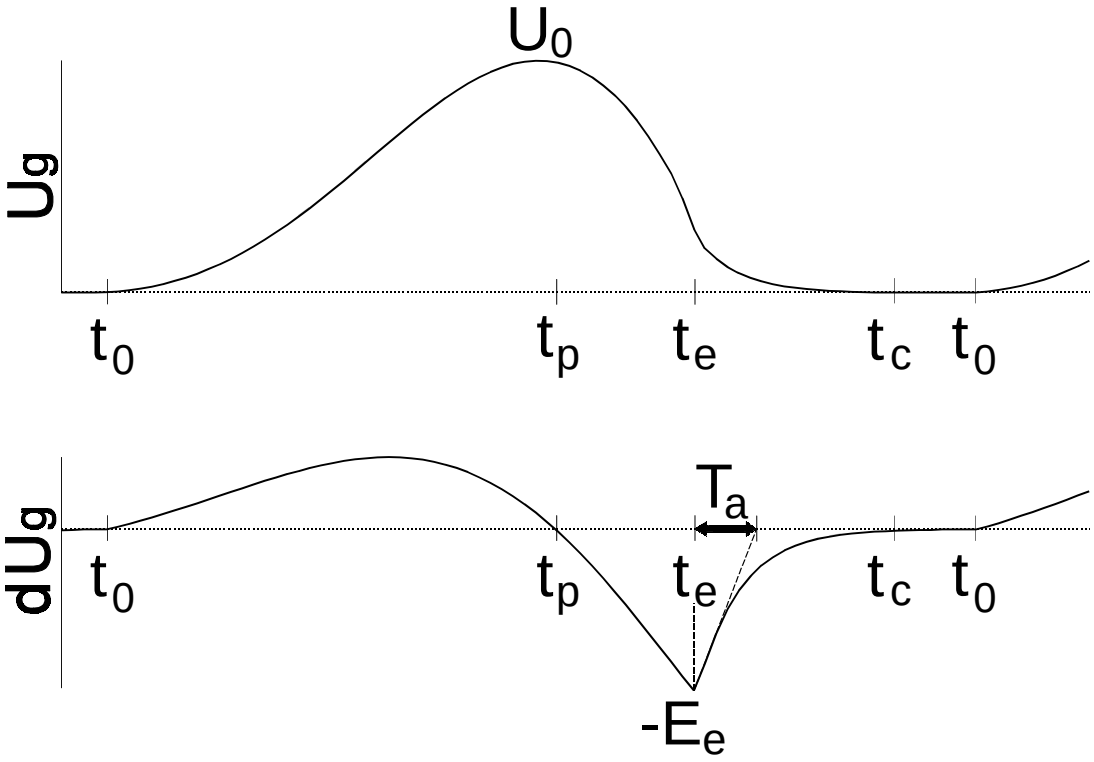

Figure 1. The LF-model and the LF-parameters.

The parameters of the LF model can be divided into three groups (see Table I).

Table I. The LF parameters.

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.

These parameters, in turn, can be used to derive many other parameters. For instance, speed quotient is often calculated: $SQ = (t_p-t_o)/(t_c-t_p)$ (e.g. Alku and Vilkman, 1995). However, in our opinion these derived parameters are less suitable for evaluation of the parametrization methods. The reason is that the derived parameters have an important drawback: whenever there is a change in a derived parameter, it is difficult to determine how this change came about (Strik, 1996a). An increase in SQ could be the result of a larger t_p, a smaller t_o, 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 (from which the parameter was derived) 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 Table I for the evaluation of estimation methods. Since the parameters E_e, t_o, 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.

2.3. Direct estimation method

In DE 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: $U_0 = \max(U_g)$, $t_p = \text{argmax}(U_g)$, $E_e = -\min(dU_g)$, 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 zero crossing is also used to estimate parameters. For instance, in this way t_o and t_c can be estimated (see Figure 1).

With DE methods, estimates of most voice source parameters can be obtained in a relatively simple way. However, DE methods also have some disadvantages. DE methods try to locate (important) events in the voice source signals. Thus the resulting estimates are limited to the place or amplitude of samples in the discrete signals. In other words, the estimated voice source parameters always have integer values. In practice, these (important) 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, as we will show in section 4. This is a major drawback of DE methods.

Another drawback of DE methods is that a disturbance 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. Some other drawbacks of DE methods can be found in Strik (1996a).

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 an FE method. To that end we chose the DE method described in Alku and Vilkman (1995), because their method seemed promising and because the authors provide a fairly detailed description of their method (see especially page 765 of their article). Furthermore, with this method it was possible to estimate the LF parameters E_e, t_o, t_p, and t_e (for which they use the terms A_{min}, t_o, 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 DE 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_o, t_p, and t_e (made with the DE method) the optimal value of T_a was estimated by fitting the LF model to the data. Therefore, strictly speaking, only E_e, t_o, t_p, and t_e can be said to be the result of the DE 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_o, t_p, and t_e made before with the DE 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_o, t_p, and t_e made with the DE 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.

After implementing this method for parameter estimation, numerous experiments were first carried out to improve the implementation. The goal was to make the estimations more robust, and thus to make the resulting average errors in the estimates smaller. In the following stage, the DE method was used for the tests described below.

2.4. Fit estimation methods

Voice source parameters can also be obtained by fitting a voice source model to the data (e.g. Ananthapadmanabha, 1984; Karlsson, 1990; Schoentgen, 1990, 1995; Jansen *et al.*, 1991; Karlsson, 1992; Strik and Boves, 1992b; Fant, 1993; Milenkovic, 1993; Riegelsberger and Krisnamurthy, 1993). In our FE method five LF parameters (E_e, t_o, t_p, t_e, and T_a) are estimated for each pitch period. The FE method consists of three stages:

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

The goal of the FE method is to determine a model fit which resembles the glottal pulse as good 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. A

routine (called the LF routine) uses the analytical expression of the LF model to calculate a continuous LF pulse for these LF parameters. Subsequently, this LF pulse is sampled and zeros are added before t_o and after t_c (until the length of the fitted signal is equal to that of the glottal pulse). The output of the LF routine are the samples of the fitted signal. In turn, the samples of the fitted signal together with the samples of the glottal pulse are the input to the error function, which 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. The desired 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 does depend on the application. In the current research we did not have a specific application in our 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 non-linear optimization techniques were tested: several gradient algorithms and some versions of a non-gradient 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 FE 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.

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 DE method should be used to provide the initial estimate. In that case the rms error for the FE method can never be larger, and will almost always be smaller than the rms error for the DE method used (because in stage 2 and 3 of our FE method the rms error can never increase, and usually decreases gradually). Consequently, the errors in the voice source parameters estimated with the FE method would almost always be smaller than those estimated with the DE method used for initial estimation. Therefore, if we had used the DE method described in the section above for initial estimation, the performance of this DE method would probably have been worse than that of the FE method. Because we considered this to be an unfair starting point, we decided to use another

routine for initial estimation. We simply selected the one we had used in previous research (Strik *et al.,* 1993).

In section 4.3 we will introduce a second version of this FE method. This second version differs only slightly from the version described here. Together with the DE method described in section 2.3 this makes a total of three estimation methods that were studied.

3. EVALUATION METHOD AND MATERIAL

3.1. 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: F_s, B_c, position (shift) and amplitude (E_e) of the glottal pulses, t_c, T_0, signal-to-noise ratio (i.e. the effect of additive noise), phase distortion (which can be caused e.g. by high-pass filtering), low-pass filtering, and errors in the estimates of formant and bandwidth values during inverse filtering (which will bring about formant ripple in the estimated voice source signals). We have performed at least 1000 model fits for each of these 11 factors, making a total of much more than 11.000 model fits.

Due to space limitations it is not possible to present the results of all the tests here. Therefore, we shall confine ourselves to the most important results, viz. those of the factors shift, E_e, and low-pass filtering. The results of other tests can be found in Strik *et al.* (1993), Strik and Boves (1994), and Strik (1994).

In natural speech many of these factors will simultaneously affect the estimated voice source parameters. Still, we think that it is better to conduct a systematic study of each factor in isolation. First of all, because otherwise it would be difficult to find out what the effect of each factor is. Second, because the contribution of these factors differs from one situation to the other, even within one experiment. If for different magnitudes of each factor it is known what the effects on the voice source parameters are, then for a given setting it could be estimated what the magnitude of each factor is and thus what the errors in the voice source parameters are. And third, because it is impossible to study all combinations (1000 cases for 11 factors make a total of 1000^{11} combinations). Certainly, some of the factors will interact. Therefore, after studying the effect of each factor in isolation, some relevant combinations should also be studied later.

Next, we had to decide whether to use natural or synthetic speech for evaluation. Natural speech has the advantage that it is realistic: it is the kind of speech (with all its properties) for which the method eventually should be used. A previous version of the FE method tested with natural speech produced plausible results (Strik and Boves, 1992b; Strik *et al.*, 1992). Plausible in the sense that visual inspection revealed that glottal flow signals and model fits were very much alike. This kind of qualitative (visual) evaluation is about the only evaluation which is usually done. Furthermore, we also checked whether the voice source parameters changed slowly in time (this is what one would expect if the voice source parameters are related to articulation), and they did. Of course the latter type of evaluation is not possible if only some pitch periods of vowels are processed (as is often done). In this case analysis of longer stretches of speech is required.

If a voice source model is used during analysis, another type of qualitative evaluation can be done. The estimated voice source parameters can be used to resynthesize the utterance, and by using perception one could try to minimize the difference between natural and resynthesized utterance (analysis-by-synthesis). This method has the disadvantage that (almost) similar percepts can be obtained with different articulatory settings. So with this method one is never sure whether the estimated voice source parameters approximate the correct voice source parameters, or whether it is another set of voice source parameters that just sounds (almost) similar. For some types of research, e.g. fundamental research on speech production, this is an important distinction.

Furthermore, all the qualitative evaluation methods mentioned above have some other disadvantages. First of all, for natural speech it is almost impossible to control all the factors which influence the estimated voice source parameters, and thus to examine the effect of each of these factors in isolation. And even if this were possible, these methods are much too laborious to be used in the numerous (more than 11.000) cases that were studied so far. After all, for every new model fit of each pitch period a qualitative evaluation has to be done by looking at or listening to the signals.

Finally, natural speech has to be inverse filtered before one can start with the parametrization of the glottal flow signals. Current inverse filter techniques work quite well, but they are certainly not perfect. Imperfections in inverse filtering lead to errors in the glottal flow signals. These errors contribute to the final errors in the estimated voice source parameters, and it becomes impossible to determine which part of the error is caused by inverse filtering and which one by parametrization. Inverse filtering has already been studied a lot in the past. Here we want to concentrate on the estimation methods.

Instead of natural speech synthetic speech can be used for evaluation. The most important drawback of synthetic speech is that it is only an approximation of natural speech, and does not contain all the properties of natural speech. However, synthetic speech also has many advantages. First of all, with synthetic speech inverse filtering and parametrization of the glottal flow signals can be studied in isolation (if a synthesizer is used that outputs both speech and the glottal flow signal). Furthermore, one can control and vary each factor, and thus each factor can be studied in isolation. Desired glottal waveforms with different kinds of shapes can easily be produced. For all these glottal flow signals the correct voice source parameters are known. They are simply the voice source parameters used to synthesize these pulses (or some transformation of them). This makes it easy to calculate the error between estimated and correct voice source parameters. Finally, the experimental cycle is fast, much faster than with the qualitative methods mentioned above. This is very important, because for a systematic and thorough evaluation many experiments have to be done (so far, already more than 11.000 model fits have been carried out).

Given the considerations presented above, we decided to use synthetic speech for our evaluations. Because we want to focus on the parametrization method, we shall not evaluate inverse filtering in the current research. In our experiments we first synthesize glottal flow signals. Subsequently, the three parametrization methods are used to estimate the voice source parameters. Finally, the estimated voice source parameters are compared with the correct ones (used to synthesize the glottal flow signals). In this way the experimental cycle is short, and can be used to perform the numerous tests which are needed. As we use the LF model for the fitting procedure, it is obvious that we also used the LF model to synthesize the glottal flow signals.

This evaluation method is equivalent to the method used by McGowan (1994) to estimate vocal tract parameters. He used the same articulatory synthesizer to produce formant tracks and to recover the articulatory trajectories from these formant tracks. His research showed that this is a useful evaluation method, which can be used to gain insight in the estimation procedure. For example, he found that the estimation could be improved by using additional acoustic information, such as rms amplitude.

In our research, just as in the research by McGowan (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.

For evaluating the estimation methods 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). Subsequently, for these test signals voice source parameters were estimated with the DE method and the FE method. The resulting values were compared with the correct values, and the errors were calculated:

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

$ERR(Y) = abs(Y_{est} - Y_{inp})$, for $Y = t_o$, t_p, t_e 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 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 this 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 which are occasionally present in the estimates. 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 three estimation methods used in this study are pitch-synchronous. This implies that a pitch period of dU_g first has to be located before it can be parametrized. Among the parameters that have to be estimated are t_o 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 (perturbated) 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.

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 the results of our experiments showed that varying these parameters had very little effect on the estimations. The influence of varying E_e and shift (which is strongly related to 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, 1992a, 1992b, 1994; Strik *et al.*, 1992, 1993; Strik, 1994) the following 11 base pulses were defined:

Table II. 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 in practice (when natural speech is used), 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 position (shift) and amplitude (E_e) of the

glottal pulses 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 estimates.

4.1. The LF routine

4.1.1. Introduction

In section 2.3 we argued that one of the drawbacks of the DE methods is that only integer values for the parameters can be estimated. Our 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 we shall test whether our LF routine has the required property, which will be called the 'non-integer' property.

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.

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 this 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.

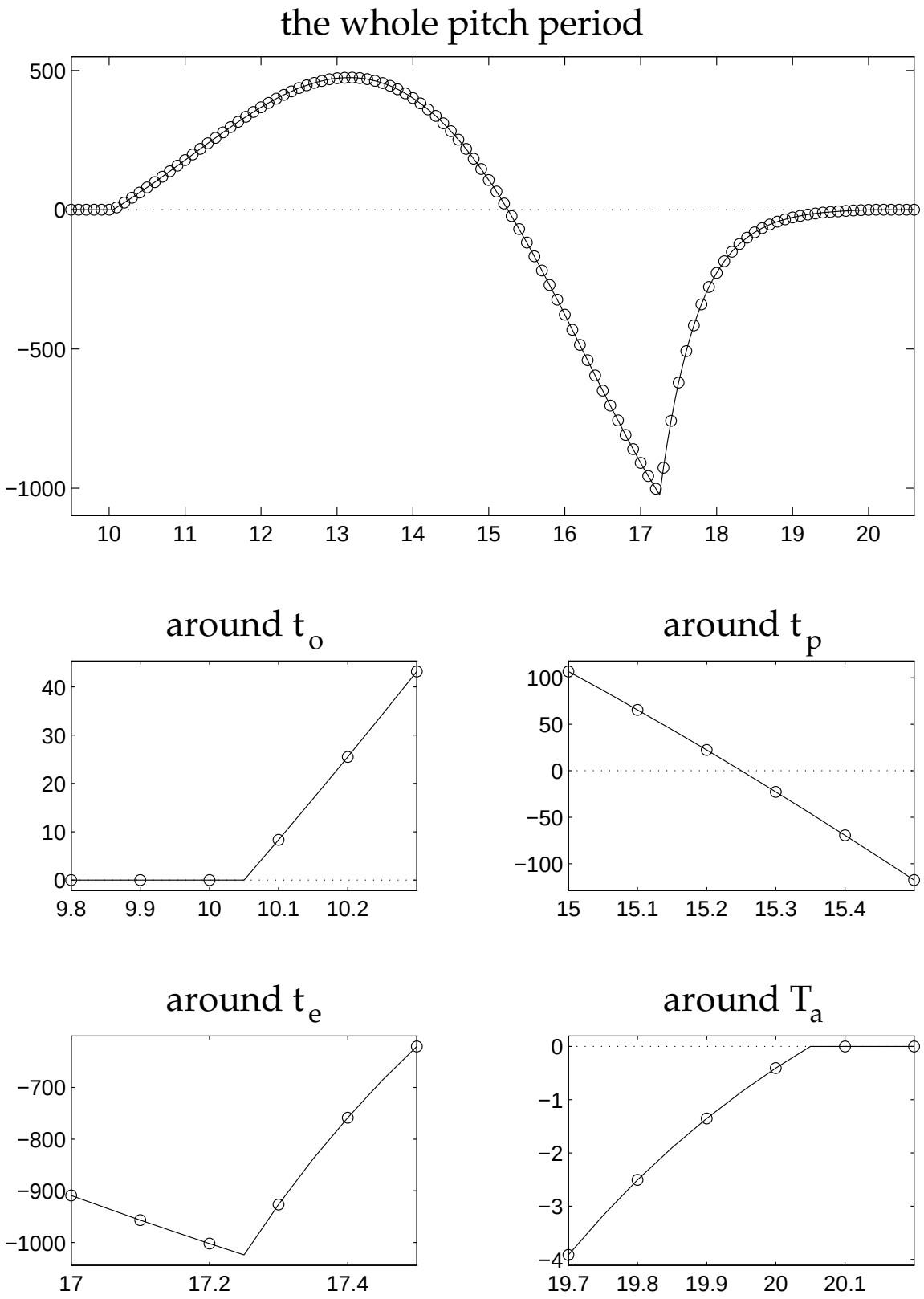

Figure 2. A 10 kHz (dotted) and a 20 kHz (solid) LF-pulse. Shown are the whole pitch period, and some details around important events.

4.2. 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, we 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).

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 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.

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).

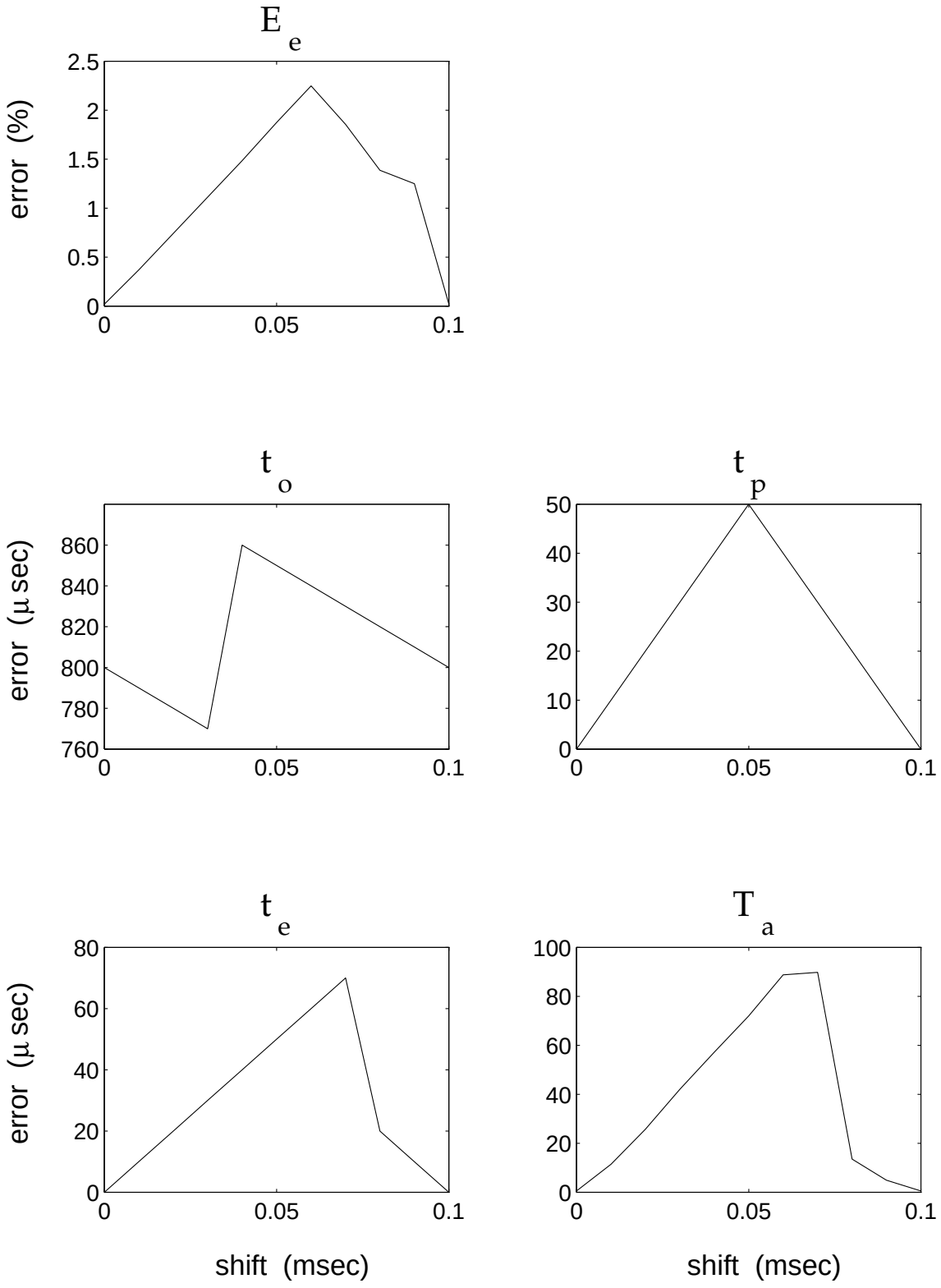

Figure 3. Median error for the estimated parameters for different values of shift.

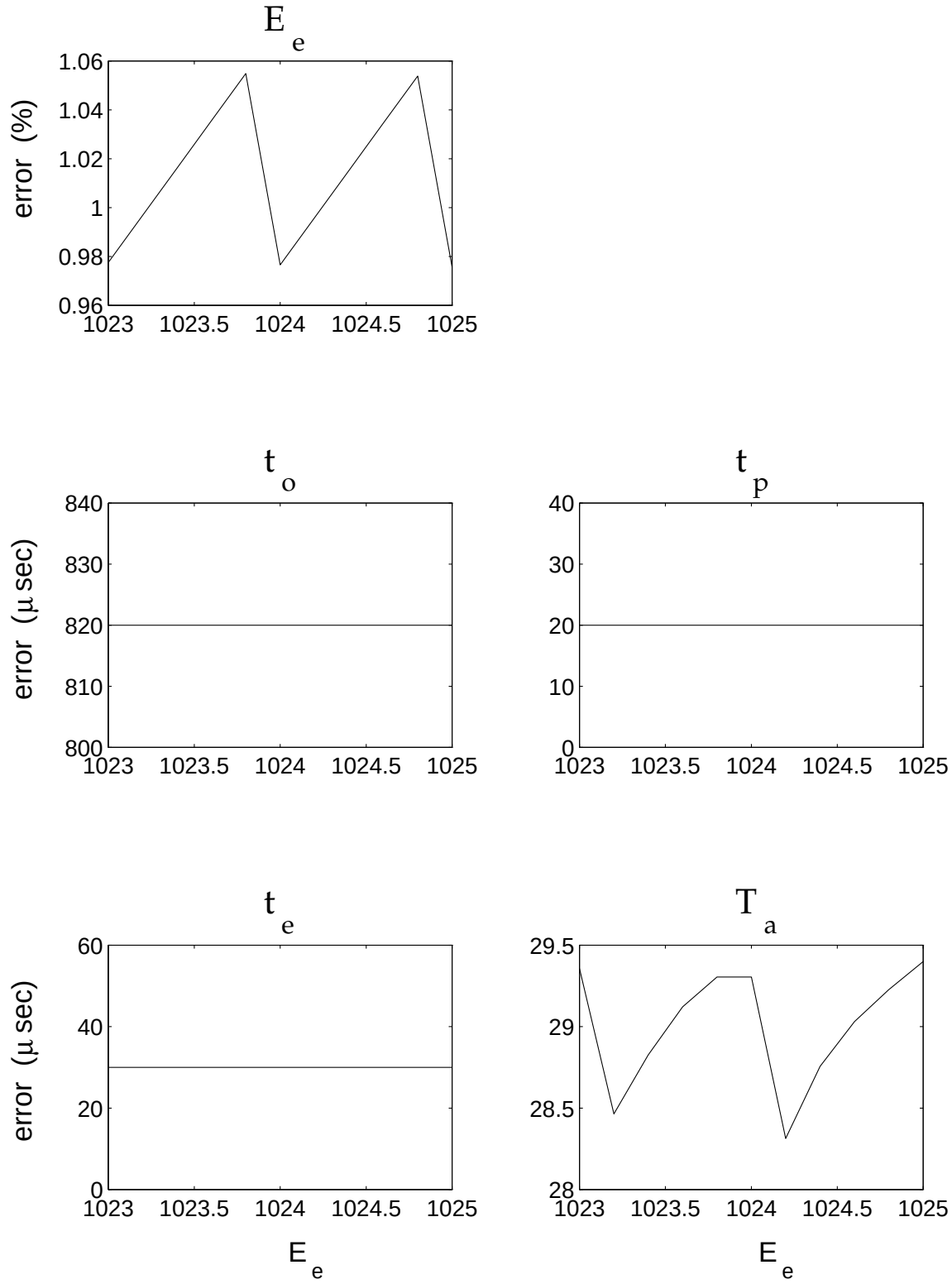

Figure 4. Median error for the estimated parameters for different values of E_e.

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. If a large number of moments is randomly distributed, the average error (both the arithmetic mean and the median) due to rounding off towards the nearest sample would be $T_s/4$ = 25 μsec. The average errors of t_p, t_e, and T_a do not deviate much from this theoretical average. The reason that the average errors are not exactly equal to 25 μsec. is that the related moments are not positioned randomly. The reason why the error in t_o is much larger was already explained above.

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. 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. 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.

4.2.5. Conclusions

The conclusions that can be drawn from these tests are the following. The errors obtained with the FE method are very small, in general much smaller than those 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 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_o and E_e could be kept constant in the definition of the base pulses (see section 3.2).

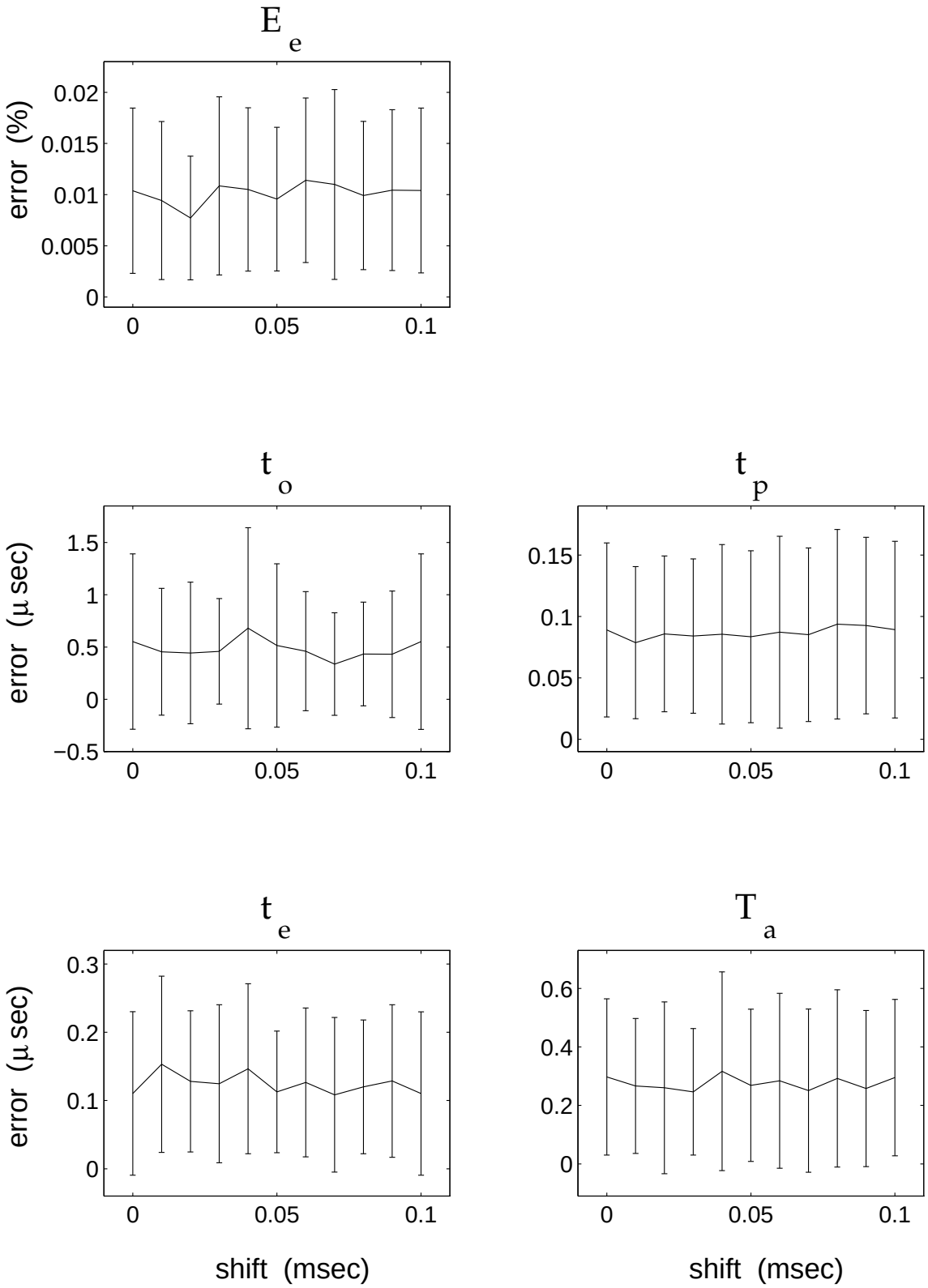

Figure 5. Mean and standard deviation of the errors in the estimated parameters for different values of shift.

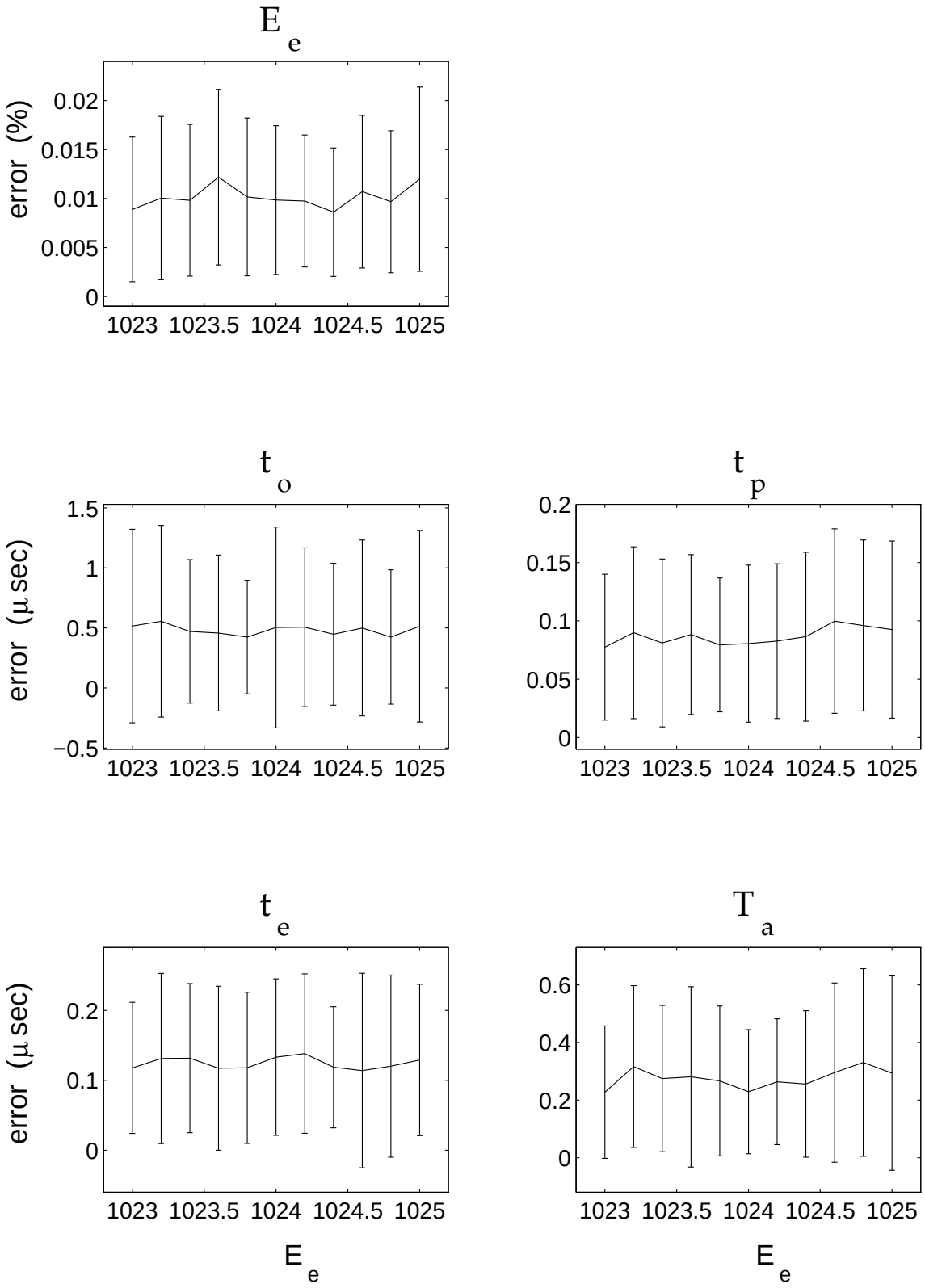

Figure 6. Mean and standard deviation of the errors in the estimated parameters for different values of E_e.

For the DE method the average errors in t_o are always larger than for the FE method, because in the former a threshold function is used to estimate t_o. In fact, the error in t_o can be substantially reduced, simply by subtracting a constant from its estimate. For the other parameters the estimation errors for the DE method are zero if the parameters have exactly an integer value. Since parameters will rarely have an integer value in practice, estimates of 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 perfect 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 DE method the average error in the time parameters (except t_o) is about $T_s/4$ = 25 μsec., which is the theoretical average for randomly distributed values, while for E_e it is about 1% (see Figure 4). For the FE method the average error in the time parameters is less than 0.5 μsec, while the average error for E_e is about 0.01% (see Figures 5 and 6).

At this point it may seem more or less trivial that the LF routine has the non-integer property. However, this is not the case. For instance, the first version of our LF routine, 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.

Since in this first version of the LF routine the input parameters are rounded off towards the nearest integer, the resulting parameters do not change gradually but instead jump from one integer value to the next. The consequence is that also the calculated rms error jumps from one value to the next, because the shape of the generated LF pulse changes abruptly. Thus the error function has the shape of a staircase. A staircase-like error function is problematic for many optimization algorithms, especially for gradient algorithms. They often get stuck in a local minimum, because the gradient is zero for each stair. Although the simplex search algorithms generally come closer to the global minimum than the tested gradient algorithms, the staircase-like error function also proved to be problematic for this algorithm. The explanation is the following. The simplex is formed by N+1 points in a N-dimensional space. During optimization the size of simplex often diminishes gradually. At a certain point the distance between two points of the simplex can become smaller than the width of a stair, and then it is usually stuck on that 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 9 values between these integers. Therefore, the second version of the LF routine has the required non-integer property. However, the error function still has the shape of a staircase. Since the stairs are 10 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 define an error function which changes gradually. The solution was simple: do not round off the input parameters to integer values; instead use the real values. Subsequently, the analytical expression of the LF model is used to calculate a continuous LF pulse. Finally, this continuous LF pulse is sampled. This is the third version of the LF routine. Lin (1990) did not use the

analytical expression, most probably because it requires too much CPU time for his application (real-time speech synthesis). Instead he used an approximation procedure for which it is necessary that the parameters are rounded off towards an integer. For our application CPU time is not essential, and therefore we can use the analytical expression.

The third version of the LF routine has the required non-integer property. More important, for this version the shape of the calculated LF pulse changes gradually when the input parameters change gradually. Consequently, the error function is no longer a staircase but a gradual function. We will call this the 'gradual' property. It is clear that LF routines which have the gradual property also have the non-integer property, i.e. LF routines with the non-integer property form a subset of the LF routines with the gradual property. This gradual property turned out to be essential. An enormous improvement in the FE 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 gradual 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.

4.3. Low-pass filtering

4.3.1. 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 DE methods, because in these methods high-frequency disturbances can influence the estimated parameters to a large extent. However, low-pass filtering changes the shape of the glottal flow signals, and, consequently, influences the estimated voice source parameters (Strik *et al.*, 1992, 1993; Perkell *et al.*, 1994; Alku and Vilkman, 1995; Strik, 1996a; Koreman, 1996).

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 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 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 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 which 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. For this example the errors in the estimates obtained by means of the FE method are: $Err(E_e)$ = -11.2%, $Err(t_o)$ = 46 μsec., $Err(t_p)$ = -28 μsec., $Err(t_e)$ = -52 μsec., and $Err(T_a)$ = 144 μsec.

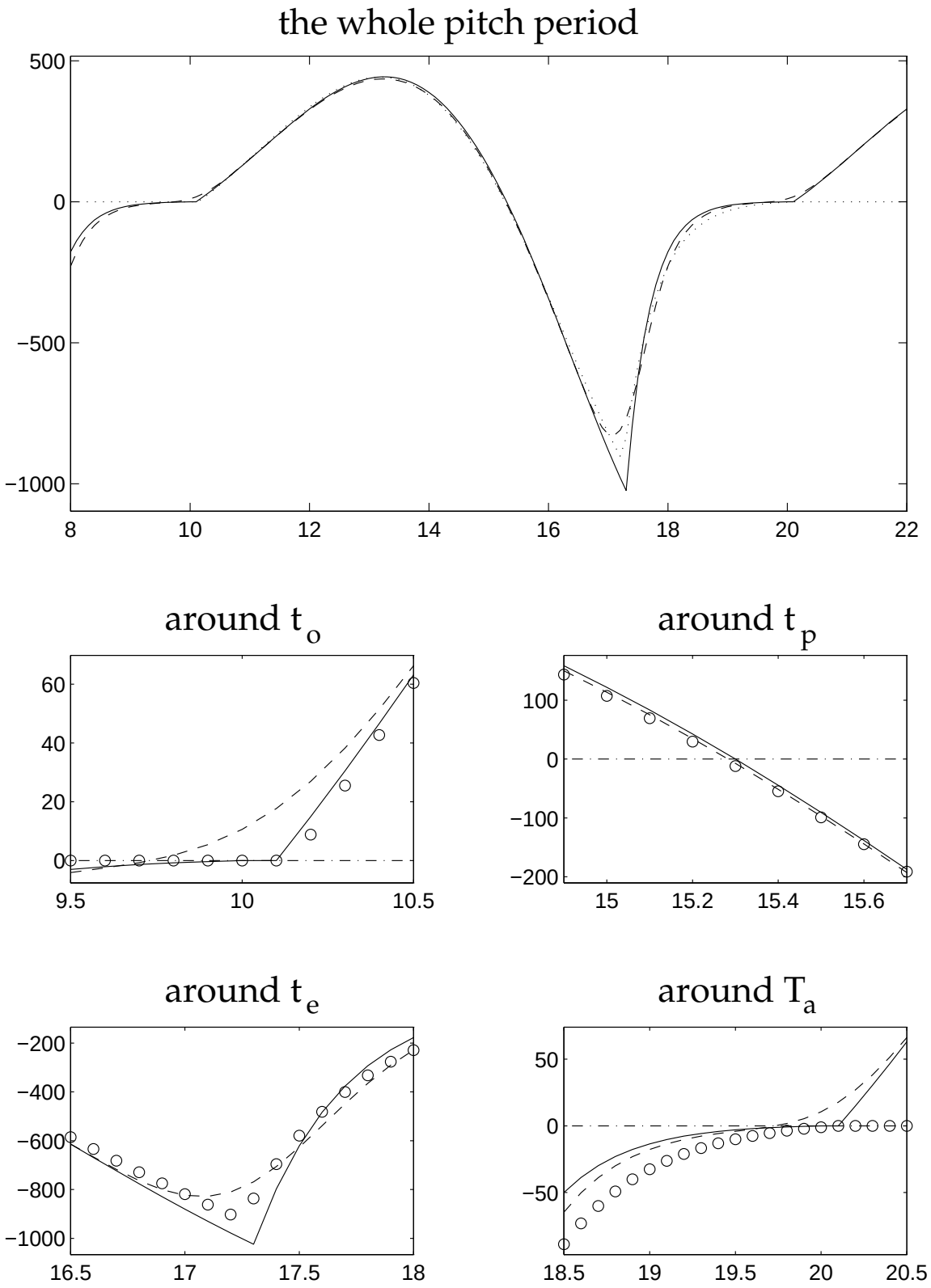

Figure 7. An example of a flow pulse before (solid) and after (dashed) low-pass filtering, and a fit on the low-pass filtered pulse (dotted). Shown are the whole pitch period, and some details around important events.

Since low-pass filtering does affect the shape of the flow pulses, and consequently also the estimated parameters, it becomes important to study the effect of low-pass filtering on the parameter estimates. This will be done in the present section. The distortion of the 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 glottal flow signals, and the parametrization method used. We will study the effect of low-pass filtering for two parametrization 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 [3]. 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 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.

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 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 trick cannot be used in a DE method, because in this case 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.

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 estimates remains within half a sample, and the median of the errors is zero.

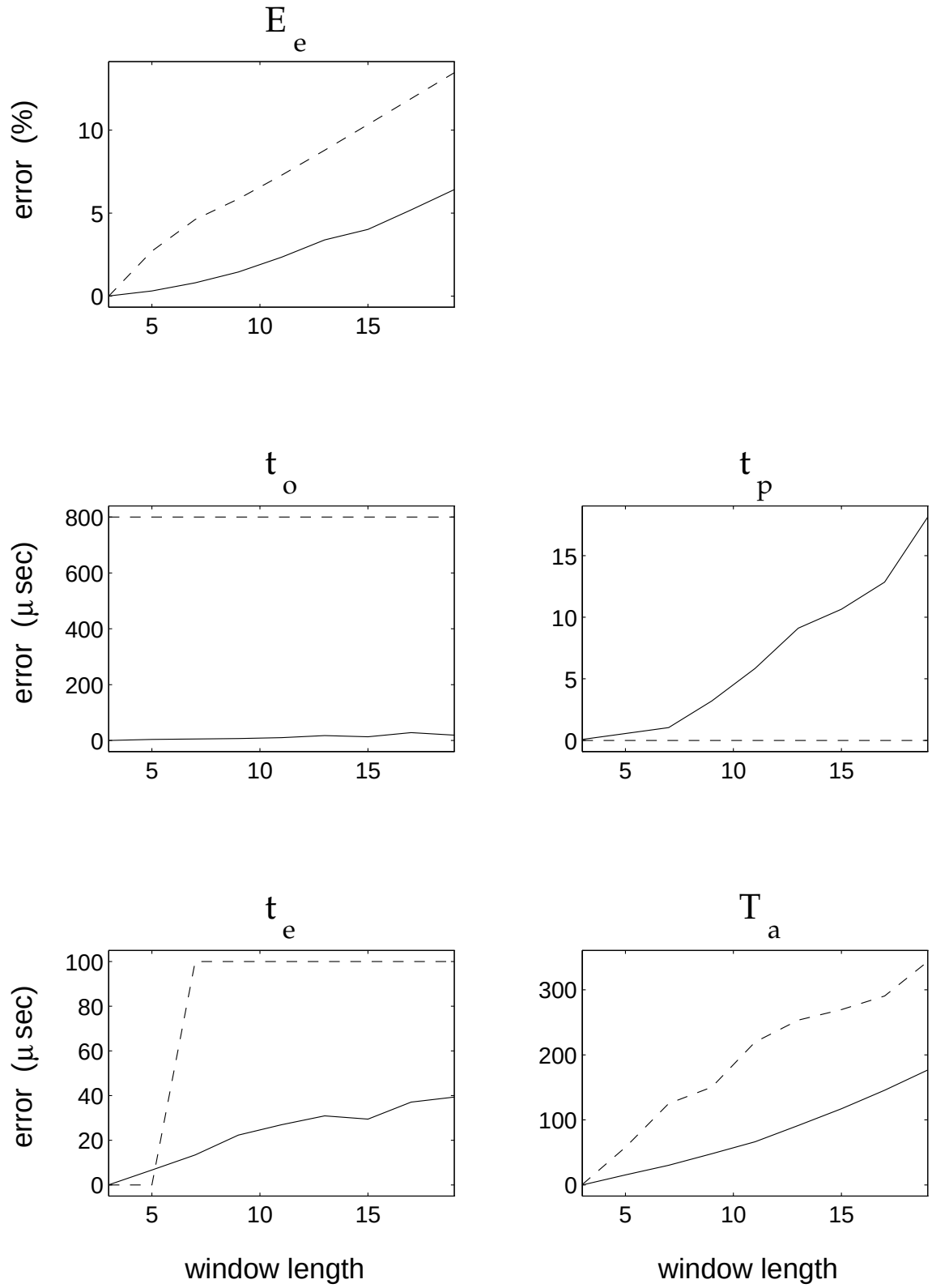

Figure 8. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the DE-method (dashed) and for the first version of the FE-method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

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. As a result the median error increases too.

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 neighborhood of t_o, and that this effect diminishes away from t_o. If the threshold is chosen high enough (which is the case for the DE method used in the current research), low-pass filtering will not have much influence on this estimate of t_o.

Here, we would like to repeat a remark made in the introduction to 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 that of most other low-pass filters. 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 (Strik, 1996a).

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. 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 we 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. In section 4.2 we estimate what the average intrinsic errors are. For the DE method this is about 1% and 25 μsec, and for the FE method 0.01% and 0.5 μsec. For a realistic comparison of the two methods these errors should be added to the errors found in this section. If this is done the average errors of the DE method are always larger than those of the FE method.

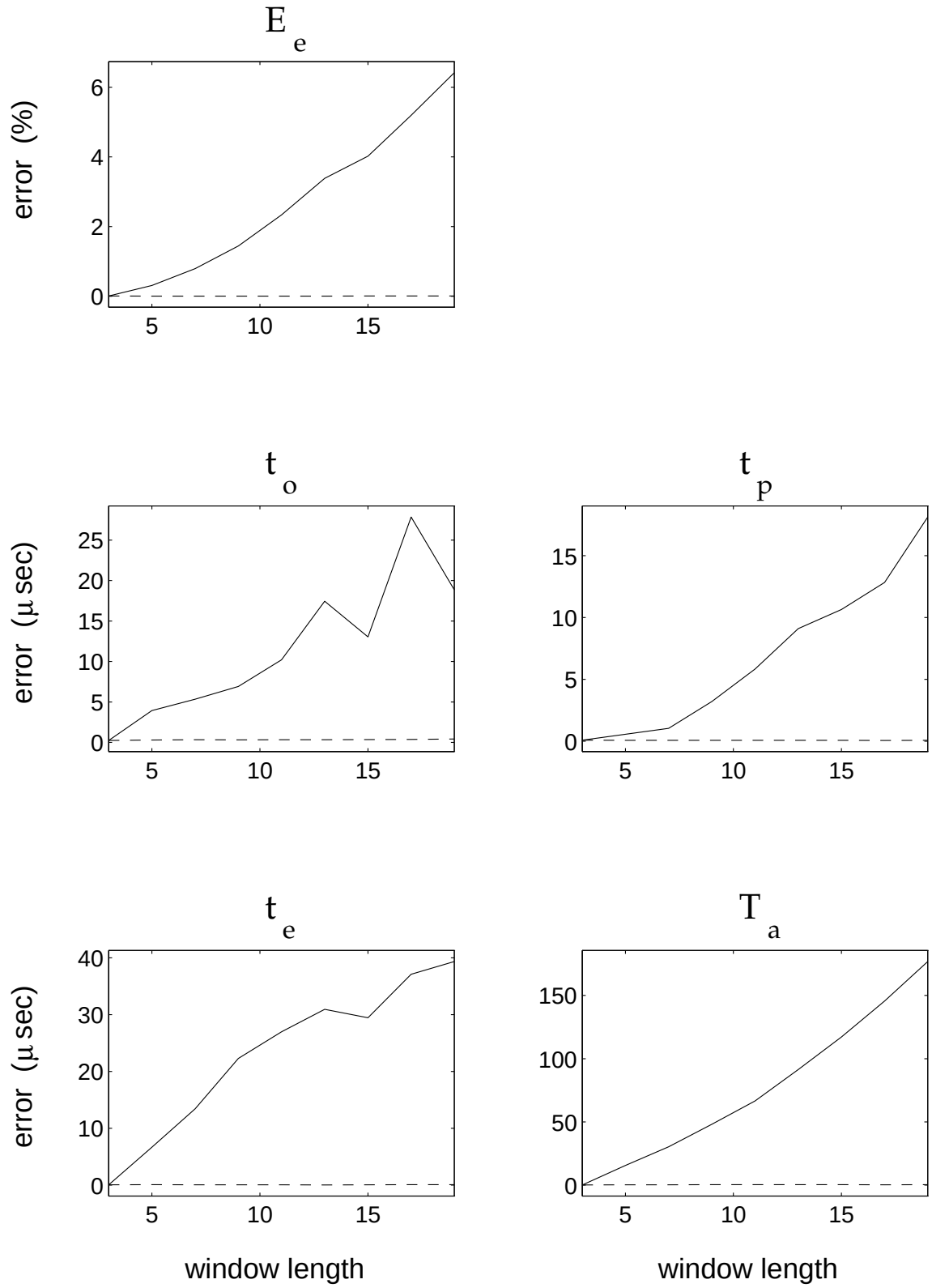

Figure 9. Median errors in the estimated voice source parameters due to low-pass filtering by means of a convolution with a Blackman window. The length of the Blackman window varies from 3 to 19 in steps of 2. Shown are the errors for the first (solid) and the second (dashed) version of the FE-method.

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 the previous sections we have explained why with the test pulses used 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 this 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. Discussion and general conclusions

In the current article the estimation of voice source parameters from flow signals is studied. Since for this purpose DE and FE methods are used most often, a representative of each method is chosen (see sections 2.3 and 2.4, respectively). In section 4.3 a second version of the FE method is proposed, making a total of three estimation methods. The goals of the research are to find out what the advantages of each estimation method are, to get a better understanding of the problems involved in these estimation methods, and finally to determine which method performs best.

In order to do this an evaluation method is needed. In section 3.1 several evaluation methods are discussed. The evaluation method used in this study is best suited for our goals. In this evaluation method synthetic test material is generated by a production model. Subsequently, the same production model is used to re-estimate the synthesis parameters. A similar method was used by McGowan (1994) to evaluate the estimation of vocal tract parameters. This evaluation method was useful for his research, and it also turned out to be useful for our own research. Since in the present research we want to focus on the estimation of voice source parameters from voice source signals, 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.

The evaluation procedure proposed here is used to test the three estimation methods described in this article. For a quantitative evaluation the LF parameters E_e, t_o, t_p, t_e, and T_a are used. Other parameters can be derived from these 5 LF parameters. These derived parameters are often used in other studies. However, in section 2.2 we argued that using derived parameters for evaluation has some disadvantages. Therefore, we prefer to use E_e, t_o, t_p, t_e, and T_a themselves for evaluation (see also Strik, 1996a).

With this evaluation method the effect of several factors can be studied in isolation. For instance, in this article results for the factors shift, E_e, and low-pass filtering are presented. However, studying each factor in isolation is not enough because some factors can interact. For example, both low-pass filtering and high-frequency disturbances present in the voice source signals (e.g. noise or formant ripple) cause errors in the estimated voice source parameters. But the errors due to the high-frequency disturbances can be reduced by using an appropriate low-pass filter. What the optimal low-pass filter for this purpose is, depends on a number of factors like e.g. the estimation method and the voice source model used, and the kind and magnitude of the disturbances. With this evaluation method the effect of factors in combination can also be studied. Thus, e.g., the optimal low-pass filter for a given situation can be determined experimentally.

With the proposed evaluation method the effect of the factor low-pass filter was studied. Low-pass filtering is probably used in all methods in which voice source parameters are estimated from inverse filtered signals. Although parametrization of inverse filtered signals has been done in many studies for almost 40 years now (i.e. 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, 1996a; Koreman, 1996).

In Strik *et al.* (1992, 1993) we mentioned that low-pass filtering changes the shape of the glottal flow signals, and consequently the estimated voice source parameters. We concluded that E_e and the return phase (i.e. T_a) are affected most by low-pass filtering (Strik *et al.*, 1992). This conclusion is supported by the results presented in section 4.3. Since the amount of change cannot easily be determined for natural speech, we suggested that a correction which is based on calculations for synthetic speech be used (Strik *et al.*, 1992, 1993).

Perkell *et al.* (1994) describe that in a first version of their data analysis procedure they used a low-pass filter “with a roll-off that began at 700 Hz and achieved 40 dB of attenuation by 1350 Hz” (ibid, p. 697). Subsequently, this procedure was used for some years to analyze large amounts of data (see references to other studies in Perkell *et al.* 1994). In a second version of the data analysis procedure somewhat less excessive low-pass filters were used. Voice source parameters estimated with the two versions of the software were compared, and differences were observed. So, more or less by accident, they observed that (the amount of) low-pass filtering influences the estimates. Indeed, for natural speech the effect of low-pass filtering cannot easily be observed, if only because for natural speech the correct voice source parameters are not known.

Perkell *et al.* (1994) concluded that the effect of the excessive low-pass filtering in the data obtained with the first version of the software appears to be confined to mfdr (which is equal to our parameter E_e). Indeed, the largest percentual differences were observed for mfdr. However, low-pass filtering will not only affect the estimates of E_e (their mfdr) but also those of all other voice source parameters. Probably, the evaluation method used by Perkell *et al.* (1994) was not sensitive enough to observe the (smaller) differences in the other parameters.

To study the effect of low-pass filtering Alku and Vilkman (1995) used a method which was similar to that of Perkell *et al.* (1994), in the sense that voice source parameters obtained with two different low-pass filters were compared. First voice source parameters were estimated for a low-pass filter with a bandwidth of 4 kHz. These voice source parameters were used as the reference values. Subsequently, the voice source parameters were estimated again for low-pass filters with a bandwidth of 2 and 1 kHz. The resulting values were compared to the reference values. Strik (1996a) showed that in only three cases was the measured difference larger than the standard deviation (in all cases for t_{ret}, the length of the return phase). The

differences found for A_{min} (our E_e) were always much smaller than the standard deviation. Our conclusion is that the evaluation method used by Perkell *et al.* (1994) and Alku and Vilkman (1995) is not optimal for studying the effect of low-pass filtering.

Koreman (1996) uses low-pass filters with small bandwidths (varying from 200 to 1500 Hz) in his data analysis method. He notes that low-pass filtering reduces the value of E_e, and concludes that low-pass filtering does not affect the relative amplitude of E_e (ibid, p. 60). This is certainly not the case. The amount of decrease in E_e due to low-pass filtering does depend on a lot of factors, an important factor being the shape of the glottal pulse. To illustrate this, let us take two pitch periods of dUg with the same E_e. The first one has a sharp negative peak, the other is more sinusoidal. The reduction in E_e due to low-pass filtering will be larger for the first pulse than for the second. Furthermore, if a low-pass filter with a ripple in its impulse response is used (like the standard FIR filters used by Koreman, 1996) the resulting low-pass filtered signals will also contain a ripple (see also Strik, 1996a). The estimates of many voice source parameters will be influenced by this ripple, and in general the error in the estimates is larger for the first pulse with a sharp peak. Since the shape of the glottal pulse changes continuously, the errors in the voice source parameters generally are not constant.

To sum up, low-pass filtering changes the shape of the glottal flow signals, and thus affects the estimates of the voice source parameters. The error due to low-pass filtering does depend on a lot of factors, e.g. the shape of the glottal flow signal, and 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 flow signal 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 (a convolution with a Blackman window).

Before we draw our conclusions regarding the comparison of the three estimation methods, we first discuss some aspects of the FE methods used in this study. The first aspect is the voice source model used in the FE 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; Funaki and Mitome, 1990; Lobo and Ainsworth, 1992; Hong *et al.*, 1994; Cummings and Clements, 1995). All voice source models for which an analytical expression exists can be used with the proposed FE method to parametrize either U_g or dU_g. In the 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 substituted 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, and thus the probability that the fitting procedure gets stuck in a local minimum is increased.

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 FE 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 flow signals. Although the LF model is capable of describing many different glottal pulse shapes, it cannot describe all details. For instance, it has been noted that there often is a second (smaller) excitation after the main excitation (Cranen, 1987; Hertegard, 1994; Koreman, 1996). The LF model cannot describe this second excitation, and therefore is not suitable to study this phenomenon. Whether a voice source model is suitable for

research depends on the goals of this research. Above we explained that with our FE method it is possible to use many voice source models. The reasons for choosing the LF model in this study are given in section 2.2.

The second aspect of the FE method we want to discuss is the non-integer property and the gradual property. In practice the value of voice source parameters will not exactly be integer, i.e. they can have all kind of non-integer values. This fact alone will bring about a substantial error in estimates obtained with a DE method, because a DE method can only estimate integer values. In section 4.2 we estimated these average errors to be about 1% for E_e and 25 μsec. for the time parameters.

Therefore, our goal was an FE method that could also estimate non-integer values. In this way we would e.g. be able to estimate moments between sample positions and thus reduce the error in the estimates. This was possible with the second and third version of our LF routine (described in section 4.2.5), which both have the non-integer property. However, another property of the LF routine turned out to be more important, i.e. the gradual property. The reason is that with an LF routine that has the gradual property it is not only possible to estimate instants between sample positions, but, more important, the optimization usually comes closer to the global minimum. This finding can probably be generalized to other FE methods and/or other voice source models: a reduction in the errors can be achieved if a routine is used (for calculation of the voice source signal) which has the gradual property.

Milenkovic (1993) also describes an estimation method which has the non-integer property. This method was not used in our research because it has some disadvantages compared to the FE methods used here. First of all, with the method of Milenkovic (1993) only t_o and t_e can be estimated, while with our FE method all parameters can be estimated. Furthermore, t_o and t_e are calculated with an iterative full search procedure. For two parameters this is feasible. However, for a larger number of parameters the number of combinations that should be tested grows exponentially, which makes this method less attractive.

The third aspect of the FE 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, which is then sampled with the same sampling frequency (F_s) as the flow signal 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 usually is more than 30 dB. Using a F_s of 10 kHz made it possible to study the effect of the factor low-pass filter 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 non-integer property is lost, and the error function (which quantifies the difference between the LF signal and the flow signal) 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 also non-integer values can 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 gradual property is preserved, and the error function remains a function that changes gradually (instead of being a staircase).

The comparison of DE- and FE methods revealed what the pros and cons of each method are. DE methods have the advantage that they are mathematically simple, and require little CPU time. However, DE methods also have many disadvantages. First of all only integer values can be estimated. Consequently, the intrinsic errors are large. The quality of the estimates depends on how well the corresponding landmarks can be determined. For instance, for t_o this is problematic because the flow signals generally change slowly during the beginning of opening. Therefore, it is difficult to determine the moment at which opening begins and the error in the estimates of t_o is generally large. Since the exact beginning of opening cannot easily be determined, a threshold function is generally used. However, our results showed that using a threshold function yields large errors in the estimates of t_o. Generally, the error in estimates of t_c is large too, as the flow signals also change slowly around t_c. Furthermore, for parameters for which a clear corresponding landmark is not present in the flow signals, estimates cannot (easily) be obtained with a DE method. An example of such a parameter is T_a, which describes the return phase. In DE methods T_a is generally not estimated. Finally, disturbances present in the signals (like noise and ripple) often will change the position of a minimum, maximum, or zero crossing, and thus result in (large) errors in the estimates obtained with a DE method.

An FE method has many advantages compared to a DE method. With an FE method it is possible to estimate non-integer values, making the intrinsic errors smaller. In fact, errors of a similar magnitude were found for estimates of integer and non-integer parameter values. Furthermore, estimates of all parameters of a voice source model can be obtained, i.e. not only for parameters related to clearly distinguishable events (as was the case for a DE method). The optimal model fit is determined for the whole pitch period, which makes the method more robust for disturbances present in the flow signals. Finally, in an FE method it is relatively easy to exchange voice source models, which is certainly not the case for a DE method. A disadvantage of an FE method is that each voice source model has its limitations, the most important one probably being that the voice source model cannot model all glottal flow pulses that occur in practice. However, as voice source models can be easily exchanged, this is not a major drawback.

In the current study two aspects were examined in detail. As parameters rarely have an integer value, we first estimated what the resulting intrinsic errors are for the two methods. For the DE method they turned out to be much larger than for the FE method. These intrinsic errors will always be present. Therefore, when the errors due to other factors are studied independently (i.e. with all input parameters having an integer value), the errors found for these factors should be increased with the intrinsic errors in order to make a realistic comparison possible. When this is done for the factor low-pass filtering, the arrangement in order of decreasing average error is: DE method, first and second version of the FE method. The factor low-pass filter was chosen because a low-pass filter is probably used in all methods in which voice source parameters are estimated from inverse filtered signals. Consequently, the resulting errors will be present in the estimated voice source parameters.

The conclusion which can be drawn on the basis of the tests presented in this article is that the second version of the FE 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.

Note, that in several ways this is a best case study. First of all, because all details of the generation of the test signals are explicitly known, as was already mentioned in section 3.1. Second, 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 third, 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.

In the introduction we already noted that DE methods and FE methods are the methods used most often. Therefore, and because they can be made completely automatic, we have compared representatives of both methods. Before we started to compare these estimation methods, we first tried to improve each estimation method as much as possible. The evaluation method proposed in section 3.1 is very suitable for this purpose. This evaluation method makes it possible to perform numerous different tests relatively easily and fast. However, during improvement of the DE method we never changed the basic algorithm. The reason is that we wanted to use the method as it is described in Alku and Vilkman (1995). We only tried to omit as many (obvious) errors as possible, i.e. we made the implementation of the DE method more robust. It is likely that the DE method can be improved, e.g. by using interpolation or by trying to reconstruct the analogous signal from the discrete signal (this can be done with the use of sync functions). However, this is generally not done. Since we wanted to use a representative of an often used method, we also did not do it. Furthermore, we are convinced that it is very unlikely that the improvement in the DE method will be such that its final performance is better than that of the FE methods (especially, that of the second version of the FE method).

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 next apply the estimation methods. In Strik and Boves (1992b) and Strik *et al.* (1992) we showed that this is possible for previous versions of the FE method. We only have to exchange the previous version of the FE method with the new improved version. The best solution would be to take the second version of the FE 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 their comments on a previous version of this paper.

Footnotes

[1] This paper is available at http://lands.let.kun.nl/TSpublic/strik/publications/. It is an elaborated and improved version of Strik (1996b).

[2] The term 'correct voice source parameters' will be used for the voice source parameters which would be obtained if the whole estimation method (i.e. inverse filtering and parametrization of the resulting flow signals) were perfect. Consequently, if a linear source-filter model is used for speech synthesis, the 'correct voice source parameters' are equal to the voice source parameters used as input for the voice source model during synthesis.

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

References

Alku, P. (**1992**). "An automatic method to estimate the time-based parameters of the glottal pulseform," Proc. Int. Conf. on Acoustic Speech Signal Process., San Francisco, USA, **2**, 29-32.

Alku, P., Strik, H., and Vilkman, E. (to appear). "Parabolic Spectral Parameter - A new method for quantification of the glottal flow," Accepted for publication in Speech Communication.

Alku, P., and Vilkman, E. (**1994**). "Estimation of the glottal pulseform based on discrete all-pole modeling," Proc. Int. Conf. Spoken Language Process., Yokohama, Jpn., **3**, 1619-1622.

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., and Vilkman, E. (**1996**). "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," Speech Communication **18**, 131-138.

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., Granstrom, B., Karlsson, I., and Lin, Q. (**1989**). "Voice source rules for text-to-speech synthesis," Proc. Int. Conf. on Acoustic Speech Signal Process, Glasgow, Scotland, **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.

Cranen, L.I.J. (**1987**). "The acoustic impedance of the glottis: Measurements and Modelling," Ph.D. thesis, Univ. of Nijmegen.

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.

Darsinos, V., Galanis, D., and Kokkinakis, G. (**1995**). “A method for fully automatic analysis and modelling of voice source characteristics,” Proc. ESCA 4th European Conf. On Speech Communication and Technology, Madrid, Spain, **1**, 413-416.

De Veth, J., Cranen, B., Strik, H., and Boves, L. (**1990**). “Extraction of control parameters for the voice source in a text-to-speech system,” Proc. Int. Conf. on Acoustic Speech Signal Process., **1**, 301-304

Ding, W., and Kasuya, H. (**1996**). “A novel approach to the estimation of voice source and vocal tract parameters from speech signals,” Proc. Int. Conf. Spoken Language Process., Philadelphia, USA, **2**, 1257-1260.

Fant (**1960**). Acoustic Theory of speech Production (Mouton, The Hague), 2nd ed., 1970.

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 Communication, **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.

Flanagan, J.L. (**1965**). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Berlin), 2nd ed., 1972.

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

Fujisaki, H., and Ljungqvist, M. (**1986**). “Proposal and evaluation of models for the glottal source waveform,” Proc. Int. Conf. on Acoustic Speech Signal Process., **4**, Tokyo, Jpn., 1605-1608.

Funaki, K., and Mitome, Y. (**1990**). “A speech analysis method based on a glottal source model,” Proc. Int. Conf. Spoken Language Process., Kobe, Jpn., **1**, 45-48.

Gauffin, J., and J. Sundberg (**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. of Speech and Hearing Research, **32**, 556-565.

Gobl, C. (**1988**). “Voice source dynamics in connected speech,” Speech Transmiss. Lab. Q. Prog. Stat. Rep., **1**, 123-159.

Gobl, C., and Ní Chasaide, A. (**1988**). “The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study,” Speech Transmiss. Lab. Q. Prog. Stat. Rep., **2-3**, 23-39.

Hertegård, S. (**1994**). “Vocal fold vibrations as studied with flow inverse filtering,” Ph.D. thesis, Univ. of Stockholm.

Hertegård, S., and Gauffin, J. (**1992**). “Acoustic properties of the Rothenberg mask,” Speech Transmiss. Lab. Q. Prog. Stat. Rep., **2-3**, 9-18.

Holmberg (**1993**). “Aerodynamic measurements of normal voice,” Ph.D. thesis, Univ. of Stockholm.

Holmberg, E.B., Hillman, R.E., Perkell, J.S., and Gress, C. (**1994**). “Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings,” J. Speech Hear. Res. **37**, 484-495.

Hong, S., Kang, S., and Ann, S. (**1994**). “Voice parameter estimation using sequential SVD and wave shaping filter bank,” Proc. Int. Conf. Spoken Language Process., Yokohama, Jpn., **3**, 1059-1062.

Jansen, J., Cranen, B., and Boves, L. (**1991**). “Modelling of source characteristics of speech sounds by means of the LF-model,” Proceedings of Eurospeech, Genova, Italy, **1**, 259-262.

Karlsson, I. (**1990**). “Voice source dynamics of female speakers,” Proc. Int. Conf. Spoken Language Process., Kobe, Jpn., **1**, 69-72.

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, Univ. 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,” Proc. Int. Conf. on Acoustic Speech Signal Process., San Francisco, USA, **2**, 13-16.

McGowan (**1994**). “Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests,” Speech Communication **14**, 19-48.

Milenkovic (**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,” The Computer Journal, **7**, 308-313.

Ní Chasaide, A., and Gobl, C. (**1990**). “Linguistic and paralinguistic variation in the voice source,” Proc. Int. Conf. Spoken Language Process., Kobe, Jpn., **1**, 85-88.

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.

Riegelsberger, E.L., and Krisnamurthy, A.K. (**1993**). “Glottal source estimation: methods of applying the LF-model to inverse filtering,” Proc. Int. Conf. on Acoustic Speech Signal Process, Minneapolis, USA, **2**, 542-545.

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

Rothenberg, M. (**1973**). “A new inverse filtering technique for deriving the glottal airflow during voicing,” J. Acoust. Soc. Am. **53**, 1632-1645.

Rothenberg, M. (**1977**). “Measurement of airflow in speech,” J. of Speech and Hearing Research, **20**, 155-176.

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

Schoentgen, J. (**1995**). “Dynamic Models of the Glottal Pulse,” in *Levels in Speech Communication: Relations and Interactions*, a tribute to Max Wajskop, edited by C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen (Elsevier, Amsterdam), 249-266.

Strik, H. (**1994**). “Physiological control and behaviour of the voice source in the production of prosody,” Ph.D. dissertation, Univ. of Nijmegen.

Strik, H. (**1996**a). “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. (**1996b**). “Testing two automatic methods for estimation of voice source parameters,” in *Proceedings of the Department of Language and Speech*, edited by H. Strik, N. Oostdijk, C. Cucchiarini, & P.A. Coppen, Vol. 19, pp. 105-127, Nijmegen, The Netherlands.

Strik, H., and Boves, L. (**1992a**). “Control of fundamental frequency, intensity and voice quality in speech,” Journal of Phonetics **20**, 15-25.

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

Strik, H., and Boves, L. (**1994**). “Automatic estimation of voice source parameters,” Proc. Int. Conf. Spoken Language Process., Yokohama, Jpn., **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, **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, **1**, 121-124.

Strube, H.W. (**1974**). “Determination of the instant of glottal closure from the speech wave,” J. Acoust. Soc. Am. **56**, 1625-1629.

Sundberg, J., and J. Gauffin (**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. London: Academic Press, 301-320.