Comments on "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering" [J. Acoust. Soc. Am. 98, 763–767 (1995)]

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

(Received 12 December 1995; accepted for publication 7 May 1996)

In the subject paper [Alku and Vilkman, J. Acoust. Soc. Am. **98**, 763–767 (1995); hence AV], AV describe the results of their research on the effect of bandwidth on estimated voice source parameters. They found that reducing the bandwidth (by low-pass filtering the glottal flow signals) leads to a distortion of the estimated parameters. Although I do agree that low-pass filtering influences the estimate of the voice source parameters, I do not agree with some of their conclusions, explanations, and recommendations. Furthermore, the method they used does not seem to be optimal for the purpose of their research. These matters are discussed in this Letter. © *1996 Acoustical Society of America.*

PACS numbers: 43.70.Aj [AL]

INTRODUCTION

In their paper, Alku and Vilkman (1995) study the effect of the bandwidth of the glottal flow signals on estimated voice source parameters. In order to study this effect, AV made some choices regarding the research method. Because these choices are important for the final results, I will discuss their choices of voice source parameters, the method to estimate these voice source parameters, the low-pass filter used to reduce the signal's bandwidth (in Sec. I), and the evaluation method (in Sec. II). I will argue that their choices are not always optimal and that there are alternatives which probably have fewer of the drawbacks mentioned in Secs. I and II. Furthermore, in Sec. III it is argued that studies in which the acoustic signal is measured by means of a microphone only, should be treated separately from studies in which oral airflow is measured by means of a Rothenberg mask (Rothenberg, 1973). To make it easier for the reader to compare my comments with the article by AV, I will utilize the terms used by AV as much as possible in this Letter.

I. DATA ANALYSIS

A. Voice source parameters

I will start this section by giving a short description of the method used by AV to estimate voice source parameters. First, the inverse filter signals U_g (estimate of the glottal flow) and dU_g (derivative of U_g) are calculated. Next, some parameters are estimated from U_g: difference between the maximum and minimum flow (A_{ac}), the moment of the onset of glottal opening (t_o), the moment of maximal glottal opening (t_m), and the moment of the end of glottal closure (t_c); and other parameters are estimated from dU_g: the minimum of dU_g (A_{min}), the moment of minimum dU_g (t_{dm}), and the moment when dU_g returns to zero level (t_{dz}) (for a definition of these parameters see also Figs. 1 and 2 of AV). In turn, the time points are used to calculate the following parameters: opening interval: $t_{01}=t_m - t_o$, closing interval: $t_{02}=t_c - t_m$, return phase: $t_{\mathrm{ret}}=t_{dz} - t_{dm}$, open quotient: $\mathrm{OQ}=(t_{01} + t_{02})/T$, speed quotient: $\mathrm{SQ}=t_{01}/t_{02}$, and closing quotient: $\mathrm{CQ}=t_{02}/T$ (T is the length of the pitch period). As these last six parameters are all derived from estimated time points, they will be called derived time parameters.

To evaluate their results, AV choose to use the parameters OQ, SQ, CQ, t_{ret}, A_{min}, and A_{ac}. Consequently, all time-based parameters used for evaluation are derived time parameters. This choice of parameters has an important drawback: whenever there is a change in a derived time parameter, it is difficult to determine how this change came about. For instance, $\mathrm{SQ}=(t_m-t_o)/(t_c-t_m)$ and thus an increase in SQ could be the result of a larger t_m, 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 estimations remain constant. It is always possible that changes in the estimations cancel each other out. Therefore, it is probably better to study the effect of bandwidth on the time points themselves. This makes it easier to evaluate and explain the results. If necessary, these time points can then be used to calculate any desired parameter.

Let us first examine the estimates of t_o. A slow increase in U_g just after t_o is often observed in practice. In such a case AV define t_o as ''the first sample whose amplitude was at least 5% of the difference between the amplitude at t_m and the amplitude t_c.'' In other cases t_o is defined as ''the time after glottal closure when the flow showed a clear increase.'' There are two problems with this definition of t_o. First, ''a clear increase'' is a rather vague description. The reader might look at Fig. 2 of AV and try to decide where the exact position of the clear increase is. And second, depending on the amount of increase, t_o will be determined by one of the definitions stated above. One can easily observe that the values for t_o obtained with these two definitions can be very different. Therefore, this definition (or rather the two defini-

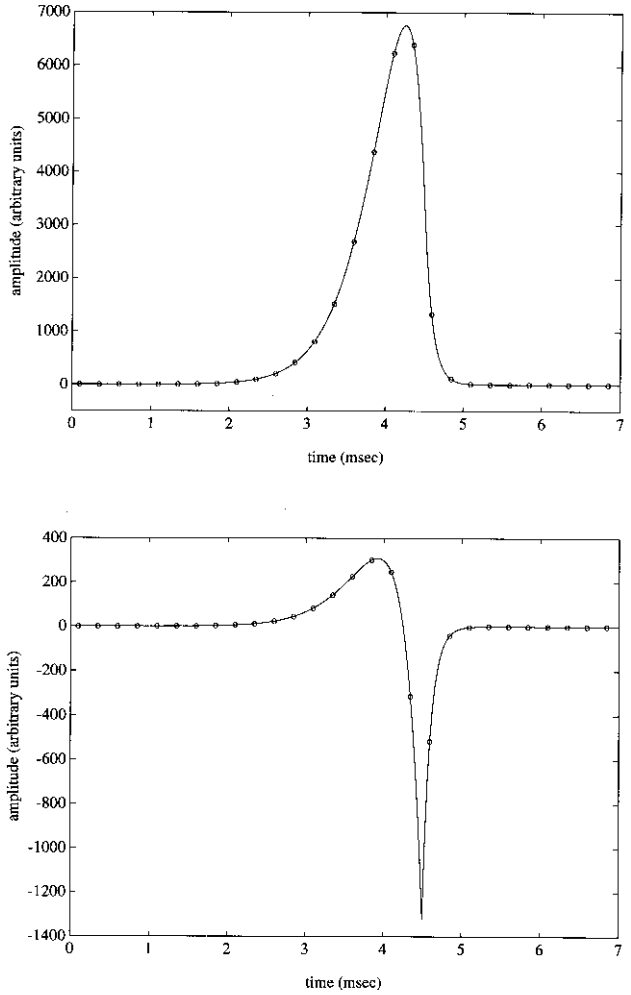

FIG. 1. An example of U_g (top) and dU_g (bottom) for pressed phonation. Shown are a time-continuous version of the signals (solid line), and a sampled version for a sampling frequency of 4 kHz (○).

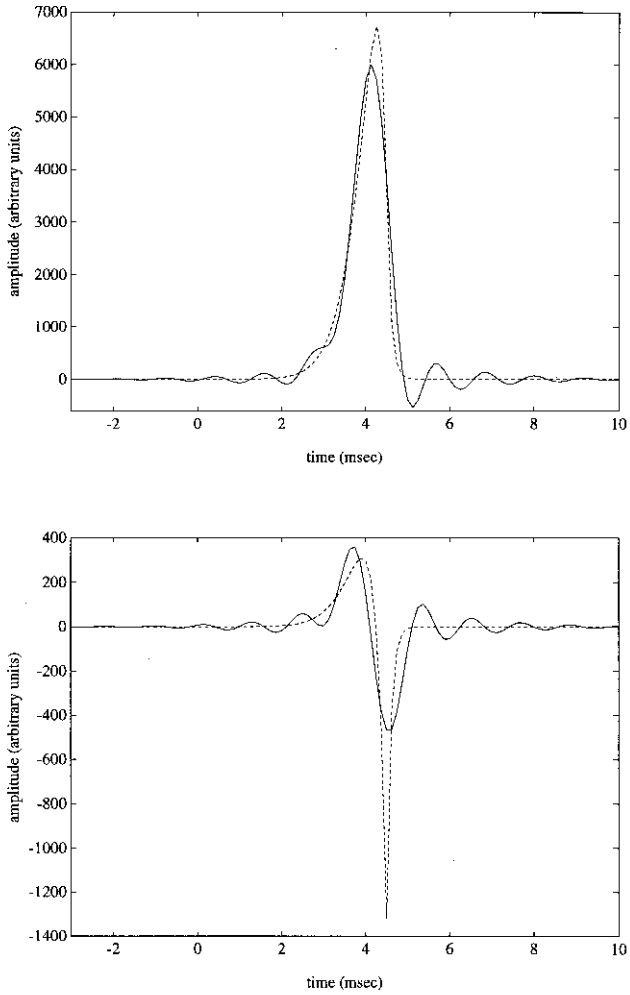

FIG. 2. An example of U_g (top) and dU_g (bottom) for pressed phonation. Shown are the signals before (dashed) and after (solid) low-pass filtering.

tions) of t_o, will yield large errors in the estimations of t_o.

In order to illustrate other disadvantages of the method used by AV, an example of a flow pulse and its derivative are shown in Fig. 1. It concerns a pulse calculated by using the analytical expressions for the LF model (Fant *et al.*, 1985). The values used to calculate this pulse are based on the values given by AV for a pressed pulse. In Fig. 2 the same pulse is shown, before and after low-pass filtering. For low-pass filtering a standard linear phase FIR-filter matching the specifications given in AV is used (i.e., the cutoff frequency is 1 kHz, and the attenuation in the stop band was more than 70 dB).

The signals drawn in Fig. 1 are idealized flow signals. In practice the inverse filter results always contain some disturbances, like, e.g., noise, formant ripple, carry-over ripple, and disturbances due to low- and/or high-pass filtering (which can lead to phase distortion and a ripple in the signal). The fact that the inverse filtered signals contain disturbances can, e.g., be seen in Figs. 1 and 2 of AV. These disturbances will have an influence on the estimated voice source parameters. For instance, in Figs. 1 and 2 of AV, and Fig. 2 of the current article one can see that these disturbances will influence the estimates of both t_o and t_c to a large extent.

Figure 1 is used to explain another disadvantage of AV's parameter-estimation method. This figure shows a time-continuous version of a synthesized flow pulse (solid line) and a sampled version of this flow pulse (symbols ''○''). AV used sampled versions of glottal flow signals to estimate voice source parameters. Their estimates are the positions and values of specific samples, e.g., a zero crossing, maximum or minimum. Consequently, in AV's method the estimates are restricted to positions and values of samples. However, due to the limited time resolution, the signal samples need not coincide with the most relevant time instants, which in turn gives rise to errors in the parameter estimates (see Fig. 1). This sampling error will be larger for smaller values of the sampling frequency. Therefore, sampling frequency also affects the estimates. On average, the error will be smaller for A_{ac} and t_m than for $A_{\min}$ and t_{dm}. The reason is that the signal changes more rapidly around t_{dm}. The sampling error is largest for those parts of the pulse in which the signal varies quickly, i.e., the high-frequency parts. Analogously, the average sampling error will be larger for pressed

pulses than for breathy ones, because for the former the signal changes more quickly.

In this section the parameter-estimation method used by AV and its drawbacks have been described. An alternative method would be to fit a voice source model to the data (Strik *et al.*, 1993; Strik and Boves, 1994). Given in Fig. 1 is the fit through the samples. However, because the fit and the original signal are almost identical, the two signals overlap. Consequently, the estimated parameters resulting from this fit differ only slightly from the values used to synthesize the sampled signal. In Strik and Boves (1994) it was shown that with this fit method it is possible to obtain good estimates and positions and amplitudes of time points lying between samples. Furthermore, in this method the estimates of the parameters are based on the signal for the whole pitch period, and are therefore more robust.

B. Low-pass filtering

In this section low-pass filtering will be considered in more detail. AV study the effect of bandwidth on the estimated voice source parameters by low-pass filtering the flow signals. For low-pass filtering AV use a standard linear phase FIR filter whose attenuation in the stop band was more than 70 dB. Using such a filter will bring about a ripple in the signal. An example of such a ripple can be seen in Fig. 2, and also in Fig. 2(c) and 2(d) of AV. This ripple will affect the estimates (see Fig. 2) and will lead to an error in the estimated voice source parameters.

To low-pass filter the signal in Fig. 2 a standard linear phase FIR filter with a cutoff frequency of 1 kHz was used (just as was done by AV). If the cutoff frequency is higher, the ripple will be smaller and, consequently, the error will be smaller too. However, the error in the estimates does not only depend on the cutoff frequency, but also on the type of low-pass filter used. A standard linear phase FIR filter has a large ripple in its impulse response, but there are other types of low-pass filters in which the ripple in the impulse response is smaller or totally absent. An example of the latter is a convolution with a Blackman window. The experiments in Strik *et al.* (1993) revealed that this type of filter usually produces better results than other types of filters.

The general conclusion of AV is that bandwidth affects the estimates. Although it is true that low-pass filtering influences the estimates (Strik *et al.*, 1992; Strik *et al.*, 1993; Perkell *et al.*, 1994), this conclusion is not complete because besides the bandwidth of the low-pass filter many other factors play a role. Above some of these factors were discussed, i.e., the type of low-pass filter, the method used for parameter estimation, the sampling frequency, and the frequency contents of the part of the flow signal under study. Furthermore, low-pass filtering can also reduce the error in the estimates, certainly if sample-based estimation methods (like the one used by AV) are used. This can easily be seen in Fig. 1. Imagine that these pulses are not clean, but contain some disturbances, like, e.g., noise. It is obvious that these disturbances will affect the position of zero crossings and extrema, and also the values of these extrema. By using an appropriate low-pass filter the effect of the disturbances on the estimates can be reduced. However, in that case one should take care to use a filter that does not disturb the signal too much. In any case, the low-pass filter (even a very good one) will always disturb the signal to some extent. To conclude, low-pass filtering can decrease the error in the estimates by reducing the effect of the disturbances, on the one hand, but it can increase the error by altering the shape of the pulses, on the other.

To end this section, I will examine the conclusion of AV that the effect of low-pass filtering was largest for the parameters calculated from dU_g, and their explanation of this finding. The conclusion was based on their results that the distortions in $A_{\min}$ and t_{ret} were larger than those in A_{ac}, OQ, SQ, and CQ. However, the three time parameters used to calculate OQ, SQ, and CQ (i.e., t_o, t_m, and t_c) can also be derived from dU_g, instead of U_g. Although in that case the calculated values would be slightly different, the magnitude of the distortions is likely to be similar, and the effect of low-pass filtering on OQ, SQ, and CQ will be small regardless of whether they are derived from dU_g or U_g. Therefore, their conclusion that the distortion due to low-pass filtering is larger for parameters calculated from dU_g than for those calculated from U_g is true for the parameters (and the definitions of these parameters) they used, but not in general.

The explanation offered by AV for the finding that the distortion is largest for the parameters calculated from dU_g is that ''this is natural since differentiation corresponds to high-pass filtering'' (p. 766). Indeed, the frequency contents of a signal and the magnitude of the distortions due to low-pass filtering are not independent. In general, the distortions of the parameters will be largest for the high-frequency parts of the flow signals, both between and within pulses. Between pulses because the distortion for pressed pulses will be larger than for breathy pulses (as shown by AV), and within pulses because the distortion will be larger for the high-frequency parts of the pulses (generally around the moment of excitation) than for the other parts (as was also shown by AV). Therefore, the conclusion is that the distortions are larger for the high-frequency parts of the flow signals, and not that the distortions of the estimates from dU_g are larger. Furthermore, as argued above, some parameters can be defined in both U_g and dU_g and for both definitions the distortions will be similar. Thus, the explanation given by AV does not seem to be plausible.

II. EVALUATION METHOD

In the previous section it was argued that parameters estimated with the method used by AV are likely to contain substantial errors. With the data presented in AV it is not possible to determine what the magnitude of the estimation error is. The reason is that the standard deviations presented in their Tables I and II are the result of a combination of these estimation errors and the variation of the parameters (both within and between the four subjects).

One can observe that the standard deviations in their Tables I and II are fairly large, especially for the parameters A_{ac}, $A_{\min}$, and t_{ret}, and for all parameters for pressed voice. In order to get an idea of the significance of the distortions they found, the standard deviations presented in their Table I

TABLE I. Standard deviations of the extracted parameters for the male subjects, expressed in percentages of the mean. The values are based on the values given in Table I of AV.

Voice type	OQ	SQ	CQ	t_{ret}	A_{min}	A_{ac}
Breathy	3.2	21.4	11.4	48.2	63.0	58.8
Normal	8.0	10.3	10.7	63.5	76.7	58.5
Pressed	31.7	24.7	21.0	67.9	73.6	72.1

are converted to percentages of the mean (see Table I). This makes it easier to compare these results with those of Table III in AV.

A comparison of these values with those of their Table III reveals that for the four male subjects the distortion (in Table III) is larger than the standard deviation (in Table I) in only two cases, viz. for t_{ret} if the bandwidth is 1 kHz and the voice type is normal or pressed. Analogously, for the female subjects the distortion is larger than the standard deviation in only one case, viz., for t_{ret} if the bandwidth is 1 kHz and the voice type is normal. Therefore, it seems that their method to study the effect of bandwidth on estimated parameters is not very sensitive.

To conclude this section, I will present a method which has fewer of the drawbacks mentioned above. The starting point of this method would be a representative database of synthesized flow pulses with known parameters. Since in this case the input parameters are known, and do not contain any estimation error, it can be determined what the estimation error is without low-pass filtering. This can simply be done by comparing the estimated parameters (without low-pass filtering) with the input parameters. Finally, an estimation can also be done with low-pass filtering. The distortions found for low-pass filtering can be compared with the intrinsic estimation error of the method, in order to judge whether the distortions found are significant.

III. TWO TYPES OF STUDIES

In their introduction AV mention several studies on inverse filtering in which different bandwidths are used. This observation was the starting point of their research. Later in their introduction they mention that all studies in which the bandwidth was smaller than 4 kHz are studies in which the oral airflow (recorded by means of a Rothenberg mask) was used, and that in the studies in which the speech pressure waveform was used the bandwidth was larger than 4 kHz. Further on in their article they do not distinguish these two types of studies any more. They conclude that bandwidth affects the estimates, and recommend the use of a bandwidth of at least 4 kHz. This recommendation makes sense for the studies based on the speech pressure waveform, but it does not seem to make sense for the studies based on the oral airflow. First of all, because it is known that the frequency response of the Rothenberg mask is only flat up to about 1 or 2 kHz (see, e.g., Hertegård and Gauffin, 1992). Second, because the flow signal has a slope of about −12 dB/oct on average, the dynamic range of the recording equipment generally does not allow for a much wider band. Therefore, the two types of studies should be treated separately.

In studies in which the speech pressure waveform is recorded by means of a microphone it seems advisable to use a bandwidth of at least 4 kHz. Apparently, this was done in all studies of this type mentioned by AV. I would like to repeat here that also in this case low-pass filtering can reduce the error in the estimates, especially if sample-based estimation methods are used (as AV did). However, in this case one should choose a low-pass filter which does not disturb the signal too much itself.

On the other hand there are the studies in which the oral airflow is measured by means of a Rothenberg mask. This technique is usually adopted by researchers who want to measure dc flow as well. In doing so they know they have to cope with the limitations of the Rothenberg mask. For this type of studies it is not sufficient to simply recommend the use of a bandwidth larger than 4 kHz. The question is rather, what kind of signal analysis should be used given the limitations of the Rothenberg mask. This has to be studied.

ACKNOWLEDGMENTS

I would like to thank Loe Boves and Bert Cranen for their helpful comments and suggestions.

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

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.

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

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.

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

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, Vol. 1, pp. 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. on Spoken Language Processing, Banff, Canada **1**, 121–124.