Normalized amplitude quotient for parametrization of the glottal flow[a)]

Paavo Alku[b)] and Tom Bäckström
Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Box 3000, Fin-02015 TKK, Finland

Erkki Vilkman
University of Oulu, Department of Otolaryngology and Phoniatrics, Oulu, Finland and Helsinki University Central Hospital, Box 220, Fin-00029 HUS, Finland

(Received 23 October 2001; accepted for publication 2 May 2002)

Normalized amplitude quotient (NAQ) is presented as a method to parametrize the glottal closing phase using two amplitude-domain measurements from waveforms estimated by inverse filtering. In this technique, the ratio between the amplitude of the ac flow and the negative peak amplitude of the flow derivative is first computed using the concept of equivalent rectangular pulse, a hypothetical signal located at the instant of the main excitation of the vocal tract. This ratio is then normalized with respect to the length of the fundamental period. Comparison between NAQ and its counterpart among the conventional time-domain parameters, the closing quotient, shows that the proposed parameter is more robust against distortion such as measurement noise that make the extraction of conventional time-based parameters of the glottal flow problematic. Experiments with breathy, normal, and pressed vowels indicate that NAQ is also able to separate the type of phonation effectively. © *2002 Acoustical Society of America.* [DOI: 10.1121/1.1490365]

PACS numbers: 43.70.Jt, 43.70.Bk, 43.70.Gr, 43.72.Ar [AL]

I. INTRODUCTION

Inverse filtering provides a noninvasive method to estimate the excitation of voiced speech, the glottal volume velocity waveform. Inverse filtering studies typically involve two stages. In the first one, an estimate of the glottal flow waveform is computed either from the oral flow using the pneumotachographic mask (also called the Rothenberg's mask) (Rothenberg, 1973) or from the speech pressure waveform (e.g., Wong *et al.*, 1979). In the second stage, the obtained estimates of the glottal flow or its derivative are parametrized by expressing their most important features using few numerical values.

Parametrization of the voice source has been the target of intensive research during the past few decades. This, in turn, has resulted in a large variety of methods to quantify the waveforms given by inverse filtering. One of the most widely used approaches to parametrize the voice source is to apply time-based parameters. This corresponds to quantifying the glottal flow using certain quotients between the closed phase, the opening phase, and the closing phase of the glottal volume velocity waveform (e.g., Holmberg *et al.*, 1988). Time-based measures have also been computed using the first derivative of the glottal flow by applying, for example, the time difference between the beginning of the closing phase and the instant of the maximal negative peak (Sundberg *et al.*, 1993). If inverse filtering is based on recording the oral flow using the properly calibrated Rothenberg mask, it is possible to parametrize voice production by measuring absolute amplitude domain values (e.g., ac flow, minimum flow, negative peak amplitude of the differentiated flow) (Holmberg *et al.*, 1988; Hertegård *et al.*, 1992; Sundberg *et al.*, 1999). Moreover, methods have been developed to quantify the voice source in the frequency domain. These techniques are typically based on measuring the decay of the voice source spectrum either from the spectral harmonics (Howell and Williams, 1992; Childers and Lee, 1991) or from the pitch-synchronously computed spectrum (Alku *et al.*, 1997). Finally, one category of voice source parametrization methods is represented by techniques that fit certain predefined mathematical functions using, for example, the Liljencrants–Fant (LF) model (Fant *et al.*, 1985) to the glottal waveform obtained by inverse filtering. In these methods, quantification of the voice source corresponds to determining the optimal parameter values of the underlying mathematical functions so that the glottal waveform given by inverse filtering is matched by its synthetic model (Carlson *et al.*, 1989; Strik and Boves, 1992).

Among the quantification methods listed above, the use of time-based parameters is one of the most prevalent. The three most commonly used time-based parameters are: (1) open quotient (OQ), which is the ratio between the open phase of the glottal pulse and the length of the fundamental period; (2) speed quotient (SQ), which is the ratio between the glottal opening and closing phases; and (3) closing quotient (CQ), which is defined as the ratio between the glottal closing phase and the length of the fundamental period. The prevalence of these time-based parameters as a method to quantify the voice source comes from the fact that they can be defined without knowing the absolute flow values of the glottal volume velocity waveform. In other words, applying

[a)]Portions of this work were presented in "Normalized amplitude quotient for parameterization of the glottal flow," Proceedings of the 5th International Workshop, Advances in Quantitative Laryngoscopy, Voice and Speech Research, Groningen, The Netherlands, April 2001.

[b)]Electronic mail: paavo.alku@hut.fi

the time-based parameters does not require the use of the Rothenberg mask in the inverse filtering stage. Moreover, computation of the time-based parameters is, at least in principle, straightforward because their values can be determined from simple time-length measurements computed directly from the waveforms given by inverse filtering. However, application of the time-based parameters in practice has shown, unfortunately, that accurate computation of their values is problematic (Holmberg *et al.*, 1988; Dromey *et al.*, 1992). This is due to the fact that time instants of glottal opening and closure are sometimes difficult to extract exactly due to formant ripple and noise that is present in the glottal waveforms given by inverse filtering. Even in the absence of formant ripple, measuring OQ and SQ is difficult because of the gradual opening of the vocal folds. Due to these problems, computation of the time-based parameters is sometimes performed by replacing the true time instants of glottal opening and closure by the time instants when the glottal flow crosses a level which is set to a certain ratio (e.g., 50%) of the difference between the maximum and minimum amplitude of the glottal cycle (Dromey *et al.*, 1992).

Among the three time-based parameters mentioned above, the use of CQ is justified for two reasons. First, CQ constitutes a measure that is affected by changes of the glottal pulse during its closing phase. Closing phase, in turn, corresponds to the portion of the glottal cycle during which the main excitation of the vocal tract occurs (Fant, 1993). Therefore, the value of CQ reflects changes that occur in the glottal source when vocal intensity or phonation type is altered: the glottal closing phase typically decreases when intensity is increased or when phonation is changed from soft to pressed (Monsen and Engebretson, 1977; Alku and Vilkman, 1996a; Sulter and Wit, 1996). Second, computation of CQ does not require determining the time instant of glottal opening, which is typically much more difficult to extract accurately than the time instant of glottal closure. Hence, when parametrization of voice production is computed from glottal flows distorted by formant ripple or noise, CQ typically yields more robust results than OQ or SQ.

In studying quantification of the glottal flow with the time-based parameters, it has become evident that these are computed by measuring the *time lengths* between corresponding events (i.e., glottal opening and closure as well as the instant of the maximal flow). However, it is also possible to measure time-domain features of the glottal closing phase using the *amplitude-domain* values extracted from the glottal flow and its first derivative. This is based on the voice source parametrization schemes developed independently and in parallel by Fant (Fant and Lin, 1988; Fant *et al.*, 1994; Fant, 1995, 1997) and Alku and Vilkman (1996a, 1996b). In these studies, the application of the ratio between the amplitude of the ac flow and the negative peak amplitude of the flow derivative has been analyzed in parametrization of the glottal source. This ratio was shown by Fant and his co-authors to yield ''a measure of effective decay time of the glottal flow pulse'' and it was used as a method to reduce the number of the LF parameters in modeling of the glottal source (Fant *et al.*, 1994; Fant, 1995, 1997).

Based on these previous studies, the current survey introduces a time-domain voice source parameter, the normalized amplitude quotient (NAQ), which is closely related to CQ. The goal of the study is to analyze, first, whether NAQ provides a more robust method for parametrizing the time-domain features of the glottal flow than CQ. Second, our aim is to analyze how NAQ behaves in the parametrization of glottal flows of different phonation types.

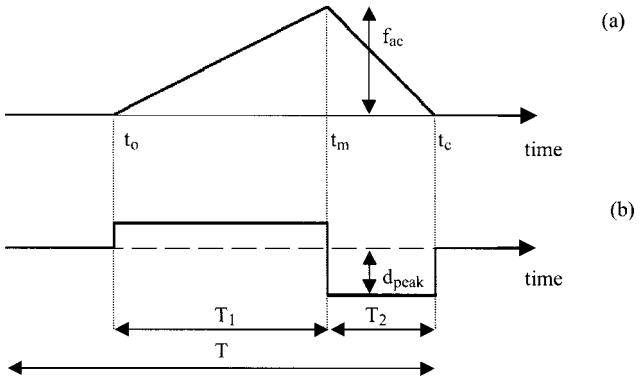

FIG. 1. A triangular-shaped glottal flow pulse (a) and its first derivative (b). Amplitude values shown in the graphs: ac flow (f_{ac}), negative peak amplitude of the differentiated flow (d_{peak}). Time values shown in the graphs: lengths of the fundamental period (T), glottal opening phase (T_1), and the glottal closing phase (T_2), instants of glottal opening (t_o), maximal glottal flow (t_m), and glottal closure (t_c).

II. MATERIALS AND METHODS

A. Normalized amplitude quotient

In order to derive the normalized amplitude quotient, let us start from a simplified model of the glottal flow represented by a triangular pulse during the glottal open phase and a zero flow during the closed phase [Fig. 1(a)]. The length of the opening phase, the closing phase, and the fundamental period is denoted by T_1, T_2, and T, respectively. The only amplitude domain value needed to define this simplified glottal pulse is the maximum value of the flow, which is denoted by f_{ac}. The first derivative of the triangular-shaped glottal pulse is given by two rectangular pulses shown in Fig. 1(b). The first of these pulses is positive and it lies between the time instant of glottal opening (t_o) and the instant of the maximum flow (t_m). The second rectangular pulse is negative; it starts at t_m and ends at the instant of glottal closure (t_c). It is worth noticing that the areas of both of the two rectangular pulses are equal to f_{ac}. This derives from the fact that the integral of the signal shown in Fig. 1(b) is the triangular pulse shown in Fig. 1(a), which starts from the zero level at glottal opening, reaches its maximum value f_{ac} at time instant t_m, and returns to the zero level at glottal closure. Due to the rectangular shape of the flow derivative, the following equation holds true during the glottal closing phase:

$$A_2 = d_{peak} \cdot T_2 = f_{ac} \rightarrow T_2 = \frac{f_{ac}}{d_{peak}} = \mathrm{AQ}, \tag{1}$$

where A_2 denotes the area of the rectangular pulse between t_m and t_c in Fig. 1(b) and AQ denotes the ratio between f_{ac} and d_{peak}. Equation (1) yields, for the simplified triangular

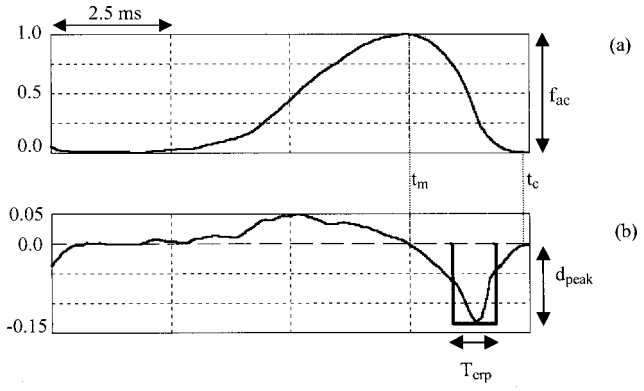

FIG. 2. A glottal flow pulse (a) and its first derivative (b) computed from natural speech by inverse filtering. Amplitude values shown in the graphs: ac flow (f_{ac}), negative peak amplitude of the differentiated flow (d_{peak}). Time values shown in the graphs: length of the equivalent rectangular pulse (T_{erp}), instants of maximal glottal flow (t_m), and glottal closure (t_c).

glottal pulse, the exact time length of the glottal closing phase as a ratio of two amplitude values, the first of which is the flow maximum and the second of which is the negative peak amplitude of the flow derivative. The value of the closing quotient can now be obtained for the triangular glottal pulse using Eq. (1) as follows:

$$\mathrm{CQ}=\frac{T_2}{T}=\frac{f_{\mathrm{ac}}}{d_{\mathrm{peak}}\cdot T}=\frac{\mathrm{AQ}}{T}. \tag{2}$$

Obviously, glottal pulses of a triangular shape do not exist in real human speech production. This can be demonstrated by comparing the glottal flow [Fig. 2(a)] of a natural vowel, obtained by inverse filtering, to its triangular-shaped artificial counterpart [Fig. 1(a)]. In particular, the shape of the differentiated glottal flow computed from the real utterance [Fig. 2(b)] is considerably different from its rectangular-shaped counterpart shown in Fig. 1(b). However, it is still possible to use the amplitude-based quotient presented above to obtain a CQ-related measure that needs no extraction of the time instant of glottal closure. Computation of the ratio $f_{\mathrm{ac}}/d_{\mathrm{peak}}$, which yields an exact time length of the glottal closing phase only for the triangular flow pulse, can be considered in the case of a natural glottal pulse as follows [see Fig. 2(b)]. An *equivalent rectangular pulse*, the height of which equals d_{peak}, is set at the instant of the negative peak of the flow derivative. The time length of this pulse is initially infinitesimal. The length of the rectangular pulse is increased until its area becomes equal to f_{ac}. (Similar to the case of the triangular glottal pulse, the value of f_{ac} equals the area computed from the flow derivative between time instants of the maximal flow and glottal closure for the natural glottal pulse as well.) When the area of the equivalent rectangular pulse is equal to f_{ac}, the length of the pulse is denoted by T_{erp} as shown in Fig. 2(b). In this case, the following equation holds:[1]

$$A_{\mathrm{erp}}=T_{\mathrm{erp}}\cdot d_{\mathrm{peak}}=f_{\mathrm{ac}}\rightarrow T_{\mathrm{erp}}=\frac{f_{\mathrm{ac}}}{d_{\mathrm{peak}}}=\mathrm{AQ}, \tag{3}$$

where A_{erp} denotes the area of the equivalent rectangular pulse and AQ denotes the ratio between f_{ac} and d_{peak}. Hence, in the case of the natural glottal pulse, computation

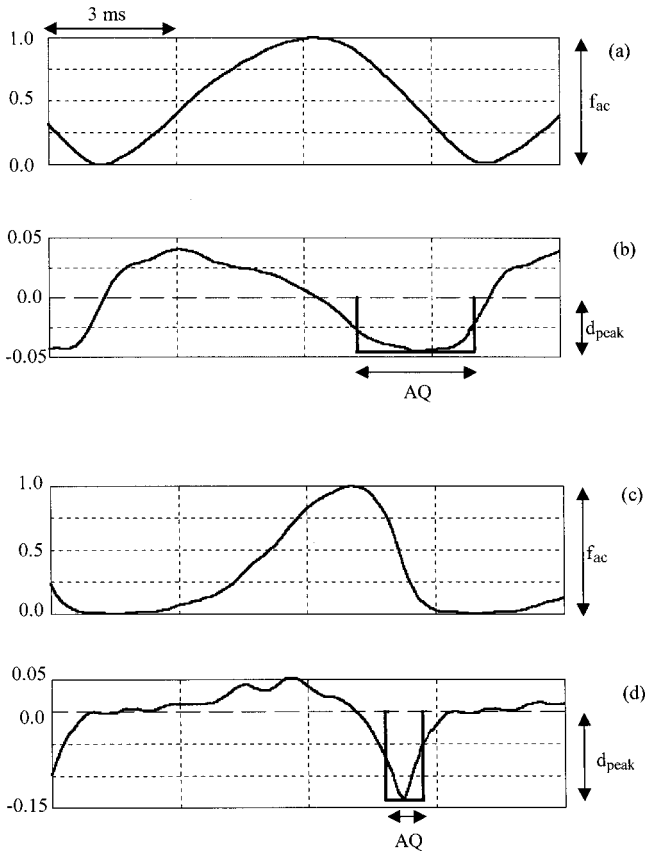

FIG. 3. Glottal flow (a) and its first derivative (b) estimated by inverse filtering in breathy phonation, glottal flow (c) and its first derivative (d) estimated by inverse filtering in pressed phonation. Amplitude values shown in the graphs: ac flow (f_{ac}), negative peak amplitude of the differentiated flow (d_{peak}). According to Eq. (3), the length of the equivalent rectangular pulse equals AQ, i.e., the ratio between f_{ac} and d_{peak}, shown in the graphs by arrows.

of the amplitude ratio $f_{\mathrm{ac}}/d_{\mathrm{peak}}$ corresponds to adjusting the time length of a rectangular pulse, whose height equals d_{peak} and whose center point[2] is located at the instant of the negative peak of the flow derivative, until the area of the pulse becomes equal to f_{ac}.

Equation (3) yields an amplitude-domain quotient to quantify the closing phase of the glottal flow that differs in two ways from measurements used in the conventional time-based parameters of the glottal flow. First, the value of AQ does not require extraction of the time instant of glottal closure. Second, AQ yields a time length, which does not correspond simply to the entire length of the glottal closing phase; rather, it represents a measure that reflects characteristics of the flow derivative in the vicinity of its negative peak. This is demonstrated in Fig. 3, which shows how the length of the equivalent rectangular pulse, i.e., AQ defined in Eq. (3), is large when the glottal flow derivative is smooth [Fig. 3(a)], which occurs in breathy phonation. However, the time length of the equivalent rectangular pulse is small in the case of a rapidly fluctuating derivative [Fig. 3(b)] corresponding to the pressed phonation type. It is worth noticing that Eq. (3) yields a value that equals the true length of the glottal closing phase only in the case of the triangular flow pulse. For natural glottal flows, however, the value given by

Eq. (3) is always smaller than the length of the glottal closing phase. The equivalent rectangular pulse has been applied previously in different forms in other fields of science. In telecommunications, frequency characteristics of filters are quantified using the concept of noise-equivalent bandwidth (Carlson, 1986). This corresponds to defining an ideal rectangular pulse for a given filter in the frequency domain that would pass as much white-noise power as the filter in question. A similar approach has also been used in psychoacoustics in quantifying auditory filters using the approach of the equivalent rectangular bandwidth (ERB) (Moore, 1982).

The time-length measure given in Eq. (3) can be normalized with respect to the length of the fundamental period in a similar manner as is done in the computation of CQ. This finally yields the following equation for normalized amplitude quotient (NAQ):

$$\mathrm{NAQ}=\frac{\mathrm{AQ}}{T}=\frac{f_{\mathrm{ac}}}{d_{\mathrm{peak}}\cdot T}. \tag{4}$$

B. Speech material

The performance of the proposed voice source parametrization method was evaluated by collecting speech data from five female and five male speakers. None of the subjects had a history of voice or hearing disorders. The voices were also perceptually within normal limits as judged by a phoniatrician. The age of the subjects varied between 29 and 52 years for females and between 32 and 47 years for males. The speakers were asked to produce a sustained /a/ vowel using breathy, normal, and pressed phonation types. The pitch was kept as constant as possible throughout the recording. The length of the pronunciation was 2 s. Subjects were allowed to use their natural fundamental frequency and intensity level during the recording. All the speakers were first trained to produce the vowel with the three different phonation types by mimicking a qualified instructor. During the recording, voice quality was assessed by a phoniatrician who asked the subject to repeat the speaking task until phonation was satisfactory. Recording of the utterances was performed in an anechoic chamber using a condenser microphone (Brüel & Kjær 4133 together with preamplifier Brüel & Kjær 2636), held 40 cm from the lips of the speaker. Speech data were saved onto a digital tape (DAT recorder TEAC RD-200T) using a sampling frequency of 22 050 Hz and a resolution of 16 bits.

After recording the speech signals an informal listening test was made in order to verify that the voices belonged to the three different phonation types. Three phonation types of each speaker were played to a panel in random order. Each panelist was asked to mark the order of the phonation type. The panel consisted of five members, all experienced in voice research. The experiment proved conclusive: all the members of the panel sorted the phonation types of each speaker correctly. The judgments of the panel were thus in line with those of the phoniatrician. Hence, we can conclude that all the subjects succeeded in producing the phonation types as required.

C. Inverse filtering

The glottal volume velocity waveforms were estimated using an inverse filtering technique that is described in detail in Alku and Vilkman (1994). This inverse filtering technique applies the acoustic speech pressure waveform that has been recorded in a free field for estimation of the voice source, i.e., no flow mask is required. The method developed is based on modeling the vocal tract transfer function with an all-pole filter which is determined using a sophisticated algorithm called discrete all-pole modeling (DAP) (El-Jaroudi and Makhoul, 1991). In comparison to the conventional linear predictive coding (LPC), which is usually applied in automatic inverse filtering, the DAP technique has been shown to yield more accurate estimates of formants, especially for high-pitched voices (El-Jaroudi and Makhoul, 1991). Hence, the estimated glottal airflow waveforms are less distorted by formant ripples (Alku and Vilkman, 1994).

The sampling frequency of the signals was first decreased from the original value of 22.050 kHz to 8.0 kHz. In order to avoid aliasing, all the signals were low-pass filtered before the downsampling with a linear phase FIR filter, that had its cutoff frequency at 4.0 kHz. Signals were then high-pass filtered with a linear phase FIR filter using a cutoff frequency of 50.0 Hz in order to remove any possible low-frequency air-pressure variations picked up during the recordings. Inverse filtering was computed by modeling the vocal tract transfer function with an all-pole filter, the order[3] of which was varied between 8 and 12. Glottal flows were estimated using a block length of 32 ms together with Hamming windowing. The position of the analysis window was initially set to the middle of the utterance. However, if the estimated glottal waveform showed evidence of formant ripple (i.e., there was an oscillating component present in the closed phase), the position of the analysis window was varied in order to find a setting that yielded a waveform with a smaller amplitude of such distortion.

III. RESULTS

The behavior of NAQ was analyzed in two parts. In the first part, robustness of the parameter was tested and compared to that of CQ. In this experiment, problematic conditions in extracting data values from glottal flows were simulated by degrading the waveforms given by inverse filtering with additive noise. The goal of the second part was to find out what kind of values NAQ typically yields when parametrizing glottal flows of three different phonation types.

A. Robustness of NAQ

One of the most important reasons behind the development of NAQ was the fact that extraction of time-based measures from glottal waveforms estimated by inverse filtering is often problematic due to noise and formant ripple that is present in the waveforms computed from natural speech. In order to evaluate the performance of NAQ in problematic conditions, we generated test material using the glottal flows inverse filtered from the voices of the ten speakers. A single cycle was cut from each glottal flow waveform. By concatenating ten such cycles, a pulse form with no jitter in the

length of the fundamental period was computed for each speaker and each phonation type, yielding a total of 30 regular glottal pulse forms. [In real speech, even when this is produced using sustained phonation, the characteristics of a single glottal pulse vary between consecutive cycles. Therefore, it is common practice to average measurements extracted from a single cycle over four to six periods (e.g., Holmberg *et al.*, 1988). However, in the current experiment a larger number of cycles was acquired for statistical analyses and, hence, ten periods were used in the averaging. The glottal excitation was synthesized by concatenating identical single pulses so that the parametrization showed no cycle-to-cycle variation.] Glottal pulse forms were then distorted by additive noise (zero mean, Gaussian distribution) in order to simulate suboptimal conditions in the parametrization of the waveforms. The amount of noise was quantified by measuring the signal-to-noise ratio (SNR) from the degraded glottal pulse forms. Eight different noise conditions were created corresponding to the following SNR values: infinity, i.e., no noise added, 60, 55, 50, 45, 40, 35, and 30 dB. Hence, the total number of glottal waveforms used in this first experiment was 240 (10 speakers, 3 phonation types, and 8 SNR categories).

Both CQ and NAQ were determined automatically for each individual glottal cycle using the following procedure. The time instant of the negative peak (t_{peak}) of the flow derivative was first identified during the glottal cycle. The length of the glottal closing phase was then determined as the sum of time spans of consecutive negative samples of the derivative before and after t_{peak}. The value of CQ was obtained as the ratio between this closing phase and the length of the fundamental period. In computing the value of NAQ using Eq. (4), f_{ac} was determined as the largest ac-flow value during the fundamental period. The value of d_{peak} was obtained by taking the amplitude of the derivative at t_{peak}. Finally, CQ and NAQ extracted from individual glottal cycles were averaged over the ten periods for each of the 240 glottal pulse forms. In order to analyze cycle-to-cycle variation in the quotient values extracted in noisy conditions, we also computed standard deviation of both CQ and NAQ over the ten glottal cycles in all 240 cases.

Two examples describing behavior of CQ and NAQ as a function of SNR are shown in Fig. 4 and Fig. 5 based on single-subject data from a female and a male subject, respectively. Both results were obtained from speech samples produced in normal phonation. Two characteristic differences between the quotients can be observed from the examples shown. First, the value of NAQ is clearly smaller (approximately 0.15 when SNR$=\infty$) than that of CQ (approximately 0.30 when SNR$=\infty$). Second, extraction of NAQ is more robust against noise, because its value changes less when SNR decreases.

In order to analyze the behavior of CQ and NAQ from the voices of all ten subjects, we used the following statistical measurements. First, relative changes of CQ, denoted by r_{CQ}, and NAQ, denoted by r_{NAQ}, between the clean conditions and the noisy conditions were computed as follows:

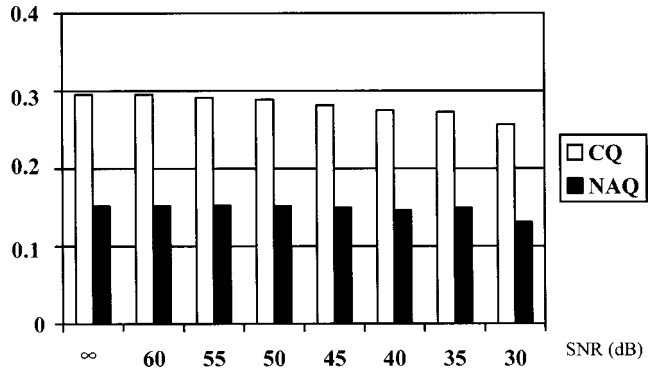

FIG. 4. Closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the *x* axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Female speaker, normal phonation.

$$r_{\text{CQ}} = 100\% \cdot \frac{|\text{CQ}_{\text{SNR}} - \text{CQ}_{\infty}|}{\text{CQ}_{\infty}}, \tag{5}$$

$$r_{\text{NAQ}} = 100\% \cdot \frac{|\text{NAQ}_{\text{SNR}} - \text{NAQ}_{\infty}|}{\text{NAQ}_{\infty}}, \tag{6}$$

where CQ_{SNR} and NAQ_{SNR} denote the closing quotient and the normalized amplitude quotient, respectively, computed with a finite SNR value, i.e., SNR$=60,55,...,30$ dB, and CQ_{∞} and NAQ_{∞} denote the closing quotient and the normalized amplitude quotient, respectively, computed in clean conditions, i.e., SNR$=\infty$. Second, in order to analyze how cycle-to-cycle variation of the quotients varies as a function of SNR, we computed the coefficient of variation (Wilks, 1962) for CQ, denoted by μ_{CQ}, and for NAQ, denoted by μ_{NAQ}, in noisy conditions as follows:

$$\mu_{\text{CQ}} = 100\% \cdot \frac{\text{s.d.}_{\text{CQ,SNR}}}{m_{\text{CQ,SNR}}}, \tag{7}$$

$$\mu_{\text{NAQ}} = 100\% \cdot \frac{\text{s.d.}_{\text{NAQ,SNR}}}{m_{\text{NAQ,SNR}}}, \tag{8}$$

where $\text{s.d.}_{\text{CQ,SNR}}$ and $m_{\text{CQ,SNR}}$ denote standard deviation and mean, respectively, of the closing quotient, and $\text{s.d.}_{\text{CQ,SNR}}$ and $m_{\text{CQ,SNR}}$ denote standard deviation and mean, respec-

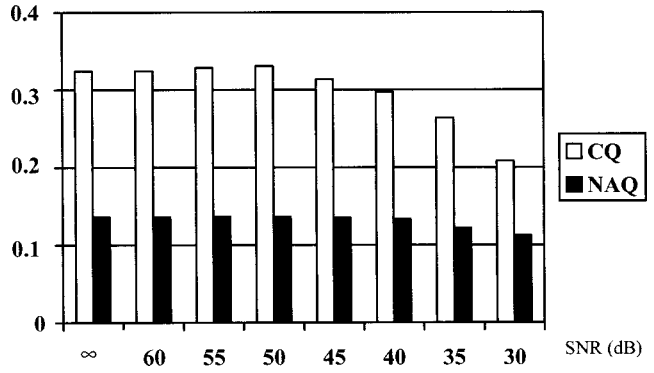

FIG. 5. Closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the *x* axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Male speaker, normal phonation.

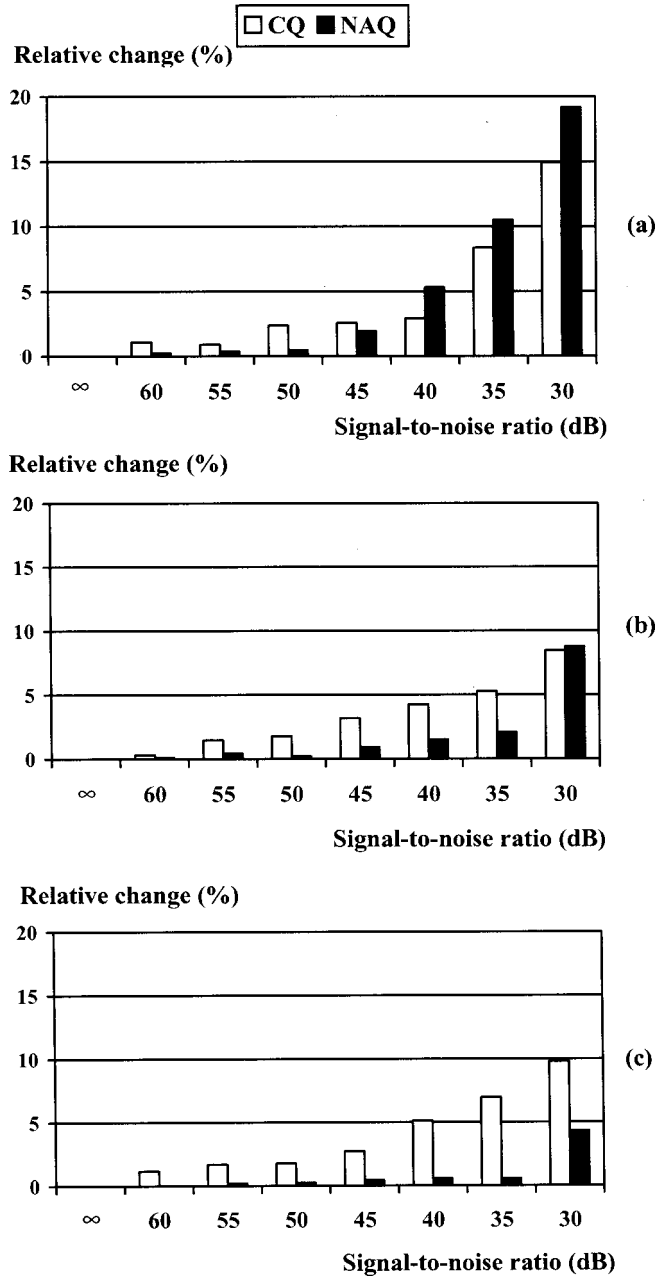

FIG. 6. Relative change of closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the x axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Relative change was determined for CQ and NAQ using Eq. (5) and Eq. (6), respectively. Female speakers ($n=5$), phonation type: breathy (a); normal (b); pressed (c).

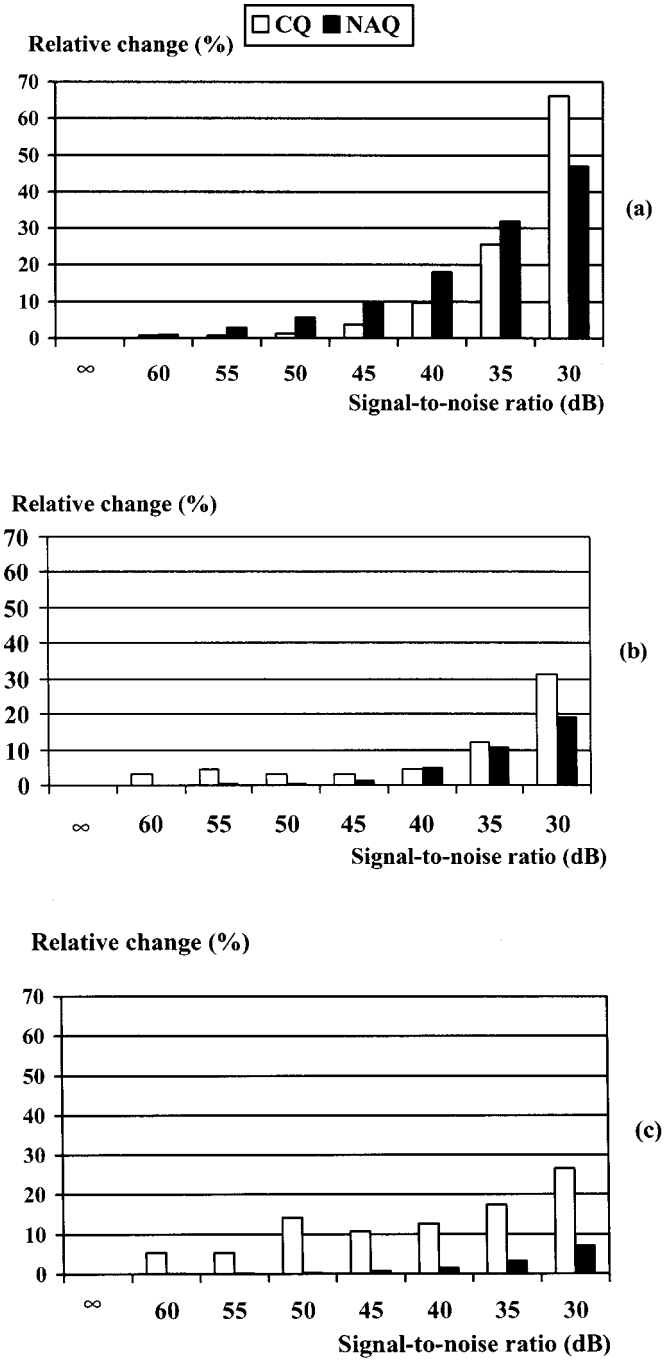

FIG. 7. Relative change of closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the x axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Relative change was determined for CQ and NAQ using Eq. (5) and Eq. (6), respectively. Male speakers ($n=5$), phonation type: breathy (a); normal (b); pressed (c).

tively, of the normalized amplitude quotient when parametrization was computed over ten glottal periods using a finite SNR value, i.e., SNR$=60,55,...,30$ dB.

The relative change of CQ and NAQ between the clean and noisy conditions is shown in Fig. 6 and Fig. 7 for female and male subjects, respectively. From these pictures it can be seen that r_{NAQ} is less than r_{CQ} in speech samples produced in normal and pressed phonation for all the values of SNR. (The sole exception is SNR value 30 dB in normal phonation of female subjects where r_{NAQ} was slightly larger than r_{CQ}.) This implies that NAQ changed less than CQ when extraction of the parameter has been distorted by adding noise to the glottal pulse form. In the case of breathy phonation, though, the result was different: the change of NAQ due to added noise was larger than that of CQ for female voices in three SNR values (40, 35, and 30 dB) and for male voices in six SNR values (60, 55, 50, 45, 40, and 35 dB).

Coefficient of variation is shown as a function of SNR in Fig. 8 and Fig. 9 for female and male subjects, respectively. Since the glottal waveforms to be parametrized were constructed by concatenating ten identical periods, it is natural that the cycle-to-cycle variation is zero for both CQ and

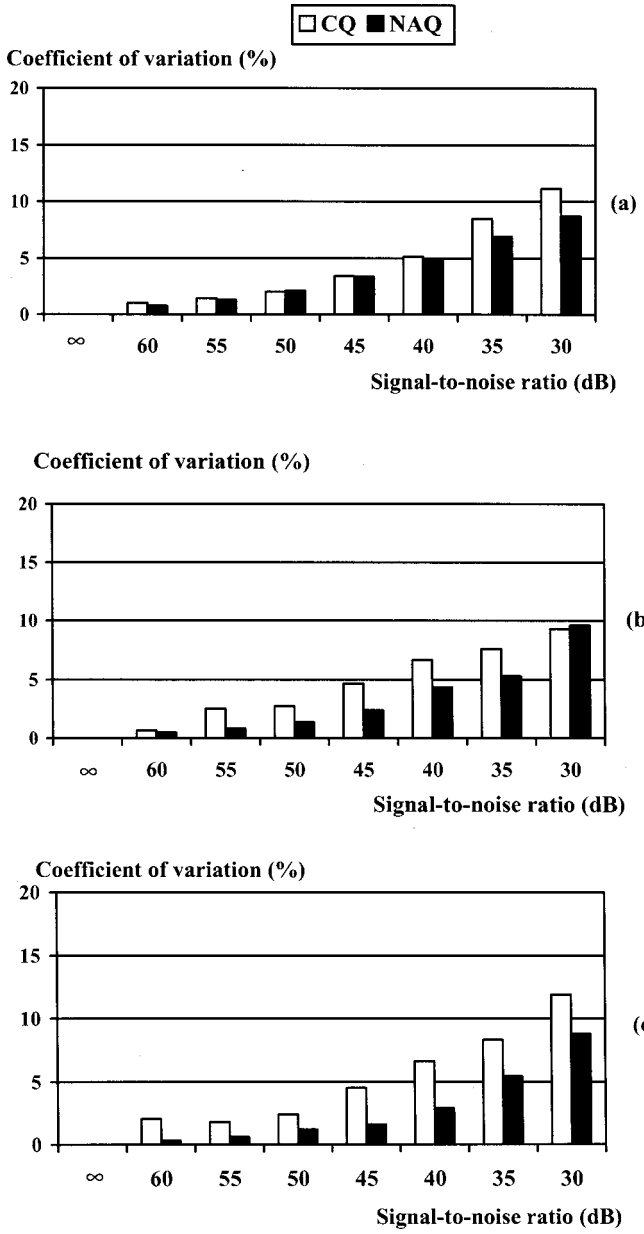

FIG. 8. Coefficient of variation for closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the x axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Coefficient of variation was determined for CQ and NAQ using Eq. (7) and Eq. (8), respectively. Female speakers ($n=5$), phonation type: breathy (a); normal (b); pressed (c).

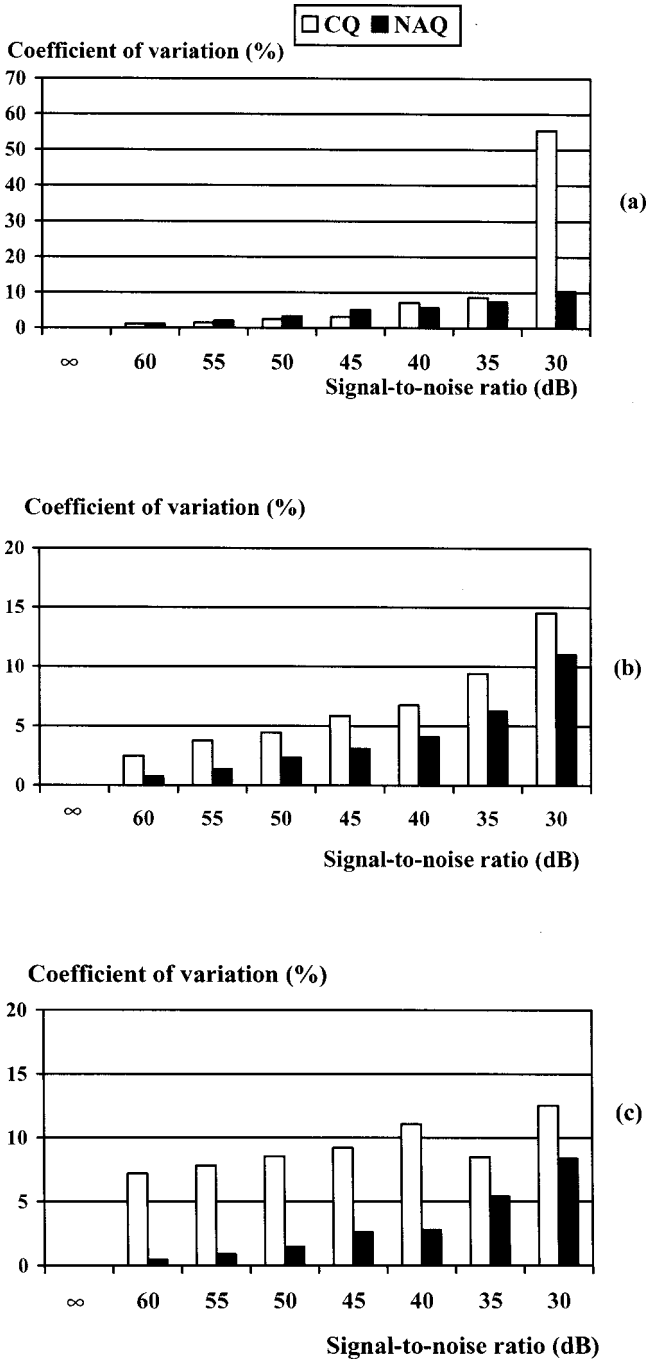

FIG. 9. Coefficient of variation for closing quotient (CQ, white bars) and normalized amplitude quotient (NAQ, black bars) as a function of signal-to-noise-ratio (SNR). The first value on the x axis (SNR$=\infty$) corresponds to the case in which no noise was added to the glottal flow. Coefficient of variation was determined for CQ and NAQ using Eq. (7) and Eq. (8), respectively. Male speakers ($n=5$), phonation type: breathy (a); normal (b); pressed (c).

NAQ when SNR equals infinity. However, when noise is added, both parameters show increased cycle-to-cycle variation. Coefficient of variation is below 10% for both parameters in almost all cases. The only major exception is CQ of breathy phonation for male voices [Fig. 9(a)] with SNR equal to 30 dB. In this case, the large value of μ_{CQ} was caused by occasional positive values of the flow derivative during the glottal closing phase due to the large amount of added noise. This, in turn, caused the extraction of the closing phase to yield values of large variation.

It can be seen that coefficient of variation is clearly larger for CQ than for NAQ. In the case of normal and pressed phonation, the value of μ_{CQ} was larger in all the SNR categories for both genders. (The sole exception is again SNR value 30 dB in normal phonation of females where μ_{NAQ} was slightly larger than μ_{CQ}.) Only in breathy phonation did μ_{NAQ} yield larger values than μ_{CQ}. This occurred in only one SNR category (SNR equal to 50 dB) for females and in three SNR categories (SNR equal to 55, 50, and 45 dB) for males.

Analysis of variance (ANOVA) of the mathematical package MATLAB (Mathworks Inc.) was also used to investi-

TABLE I. Mean, standard deviation and range of closing quotient (CQ) and normalized amplitude quotient (NAQ) in the three phonation types. Female speakers ($n=5$).

Parameter	Phonation type	Mean	Standard deviation	Range
CQ	Breathy	0.40	0.065	0.29–0.48
	Normal	0.29	0.037	0.26–0.36
	Pressed	0.26	0.036	0.22–0.32
NAQ	Breathy	0.22	0.039	0.15–0.27
	Normal	0.15	0.016	0.13–0.16
	Pressed	0.12	0.020	0.10–0.15

gate the effects of different factors on CQ and NAQ. First, one-way ANOVA with noise category as a factor was computed. Results showed a statistically significant influence of SNR on CQ [$F(7,232)=5.81$, $p<0.05$] but not on NAQ. Second, a three-way analysis of variance with the factors noise category, phonation type and gender was computed by taking into account noise categories with SNR$\geqslant$40, where effects of noise did not affect either of the two parameters too seriously. It was found that both CQ [$F(2,171)=300.8$, $p<0.05$] and NAQ [$F(2,171)=361.3$, $p<0.05$] were highly dependent statistically on the phonation type. However, there was no effect of gender either on CQ [$F(1,171)=0.88$, $p>0.05$] or NAQ [$F(1,171)=0.000\,17$, $p>0.05$]. In summary, these statistical analyses on noise-corrupted waveforms indicate that both of the parameters are able to categorize the type of phonation effectively, but NAQ is less vulnerable to distortion of the glottal pulse than CQ.

B. Behavior of NAQ in parametrization of glottal flows of different phonation types

The aim of the second part of our experiments was to analyze how NAQ behaves when it is used in parametrization of glottal flows representing three different phonation types (breathy, normal, and pressed). The data material of this part comprised all 30 glottal waveforms (10 speakers, 3 phonation types) obtained by inverse filtering. The glottal waveforms and their first derivatives were analyzed using the same extraction procedure described in Sec. III A. The values required for computation of CQ and NAQ were averaged over four consecutive glottal periods.

The results are shown in Table I and Table II for female and male voices, respectively. These data show that the mean value of NAQ decrease for both genders when phonation is changed along with the axis breathy–normal–pressed. It

TABLE II. Mean, standard deviation and range of closing quotient (CQ) and normalized amplitude quotient (NAQ) in the three phonation types. Male speakers ($n=5$).

Parameter	Phonation type	Mean	Standard deviation	Range
CQ	Breathy	0.45	0.046	0.38–0.51
	Normal	0.27	0.022	0.24–0.30
	Pressed	0.22	0.031	0.18–0.25
NAQ	Breathy	0.28	0.044	0.23–0.35
	Normal	0.13	0.022	0.11–0.17
	Pressed	0.09	0.011	0.08–0.11

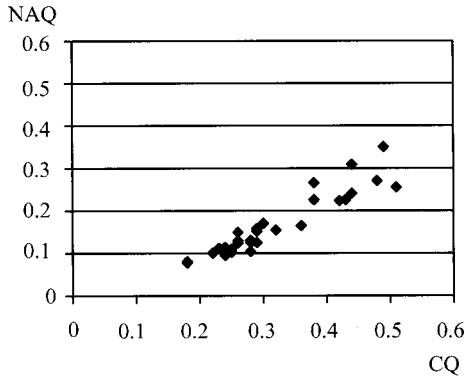

FIG. 10. Closing quotient (CQ) and normalized amplitude quotient (NAQ) expressed in the same plane for all the analyzed speech samples (10 speakers, 3 phonation types).

should be emphasized that this monotonic decrease of the parameter value occurred for all ten subjects analyzed. Statistical analyses (t-test, 95%) revealed that for male subjects the value of NAQ was significantly affected by the phonation type. The value of CQ, however, did not show a statistically significant difference between the normal and pressed male voices. For female subjects, both CQ and NAQ yielded a statistically significant difference between the breathy and normal voices, while neither of the two parameters showed a significant difference between samples produced using normal and pressed phonation.

In summary, all 30 analyzed voices are expressed in the same coordinate in Fig. 10. This figure indicates that correlation between CQ and NAQ is high (correlation coefficient equaled 0.94), even though voices from different phonation types and from both genders are pooled together. This result indicates that in the time-domain parametrization of the glottal closing phase the conventional CQ parameter can be replaced with NAQ, even though the analysis involves speech sounds of greatly different voice source characteristics. This, in turn, implies that the time-domain parametrization of the glottal closing phase can be improved by making it more straightforward and less vulnerable to distortion such as measurement noise.

IV. CONCLUSIONS

This study presents the normalized amplitude quotient (NAQ), which makes possible quantifying time-based features of the voice source from amplitude-domain measurements extracted from the glottal flow and its first derivative. The parameter is based on expressing the ratio between the amplitude of the ac flow and the negative peak amplitude of the flow derivative using the concept of equivalent rectangular pulse, a hypothetical signal with its center point located at the instant of the main excitation of the vocal tract. The ratio between the two amplitude values corresponds to the time length of the equivalent rectangular pulse when the area of the pulse is equal to the ac flow of the glottal pulse. The value of NAQ is determined by normalizing the time length of the equivalent rectangular pulse with respect to the duration of the fundamental period.

NAQ is a sequel to previous studies addressing parametrization of the glottal flow based on amplitude domain measurements (Fant and Lin, 1988; Fant *et al.*, 1994; Fant, 1995, 1997; Alku and Vilkman, 1996a, 1996b). The latter studies presents AQ as a straightforward method to parametrize glottal flows of different phonation types from two amplitude domain measures. Significantly, their studies did not discuss the fact that AQ, even though extracted from two amplitude values, yields a time-domain quantity. Moreover, there was no normalization of AQ with respect to the length of the fundamental period. Consequently, AQ values presented in Alku and Vilkman (1996a, 1996b) were different for female and male speakers. Obtaining a time-domain measure by computing the ratio between f_{ac} and d_{peak} was first presented in Fant *et al.* (1994). In their study, the ratio between the two amplitude values is called the effective declination time and it is geometrically interpreted as ''the projection on the time axis of a tangent to glottal flow at the point of excitation, limited by ordinate values of 0 and f_{ac}'' (Fant, 1997). In Fant (1995, 1997), the effective declination time is normalized by multiplying it by $F0/110$ (i.e., fundamental frequency, $F0$, of the voice divided by the approximated average fundamental frequency, 110 Hz, typical in male speech). This normalization yields a voice source parameter, denoted by R_d, that equals NAQ divided by 110.

Even though there is a close relationship between NAQ of the current study and R_d discussed in Fant (1995, 1997), the concept of the equivalent rectangular pulse is new in the parametrization of the glottal source. This concept constitutes an alternative to Fant's geometrical interpretation of the ratio between the amplitude of the ac flow and the negative peak amplitude of the flow derivative. Our new interpretation is motivated, first, because the equivalent rectangular pulse makes it easier to understand and visualize the role of the quotient between the two amplitude values rather than to deal with the ''projection on the time axis of a tangent to the glottal flow.'' Second, and more importantly, the computation of NAQ using the current approach allows for a straightforward comparison of two glottal flow pulses in terms of their time-domain features during the closing phase: the glottal pulses are transformed into the *same* simple functions (i.e., rectangular pulses) and the comparison is performed between these waveforms. Transforming waveforms of different shapes into rectangular functions has proven to be useful, for example, in telecommunications in comparing frequency characteristics of filters using noise-equivalent bandwidths (Carlson, 1986). Therefore, to make this idea known also in the time-domain parametrization of the glottal waveform is justified. Finally, we would like to point out that previous studies on R_d have not addressed widely the classification of the phonation type.

In computation of NAQ, the length of the equivalent rectangular pulse is determined using the maximum amplitude values of the flow and its derivative during one glottal cycle without requiring the extraction of the time instant of glottal closure. Therefore, computation of NAQ is straightforward, even though glottal flow waveforms contain distortion such as measurement noise that make the extraction of glottal closure complicated. Difficult conditions in data extraction were simulated in the present study by degrading glottal flows given by inverse filtering with additive Gaussian noise. The comparison between NAQ and its conventional counterpart, closing quotient (CQ), showed that the proposed new technique yielded a more robust parametrization method than CQ in normal, but especially in pressed phonation. However, in breathy phonation, in particular for male voices, the value of CQ changed less than NAQ when noise-distorted pulse forms were compared to the clean ones. (From the point of view of NAQ this is regrettable because time instants of breathy phonation are typically the most difficult to extract reliably and, hence, their extraction would require improved parametrization methods.) The reason why NAQ is more robust against noise than CQ in normal and pressed phonation types but not so much in breathy phonation is explained by the shape of glottal flow derivative at the instant (t_{peak}) of its negative peak. In normal and in pressed phonation, the derivative curve at t_{peak} is very sharp and of a large amplitude, which implies that the value of d_{peak} changes little, even though the flow waveform is somewhat distorted by noise. However, in breathy phonation the waveform of the flow derivative at d_{peak} is smooth and the value d_{peak} is low, which makes it more vulnerable to the effects of noise.

Since NAQ is computed using two amplitude values, both of which are extracted at a single time instant, it is possible for the accuracy of the quotient to deteriorate when the glottal flows are severely affected by noise. Distortion of the flow signal even at a single time instant during the glottal closing phase might, in the worst case, cause a large error in the value of the negative peak of the derivative. (This results typically in an increased value of d_{peak}, which in turn reduces the value of NAQ.) To alleviate the distortion caused by instantaneous noisy peaks of the flow derivative, it could be possible to use a procedure where a glottal flow model represented by predefined mathematical functions is first fit to the flow waveform over the entire length of the closing phase. The value of NAQ could then be computed by using the derivative of this mathematical function instead of using d_{peak} extracted at a single time instant of the original, distorted derivative.

By analyzing voices produced in three different phonation types, the study showed that NAQ values were, on average, approximately 50% smaller than the corresponding CQ values. It was also proven that there is a high correlation between NAQ and CQ. In addition, the value of NAQ showed a monotonic decrease for all ten analyzed subjects when phonation was changed from breathy to normal and then to pressed. Hence, NAQ reflects changes in the voice register. In this respect, the behavior of NAQ is similar to that of CQ. In measuring the time-domain features of the glottal closing phase, however, NAQ takes into account only the energetically decisive portion of the flow derivative in the vicinity of t_{peak}. This principle can be considered more justified than measuring the entire length of the glottal closing phase embedded in the computation of CQ if the derivative waveform is smooth (and consequently of minor importance energetically) in the beginning and in the end of the glottal closing phase.

In focusing on a subsection of the entire glottal closing phase, NAQ is similar to some of the time-domain parametrization schemes reported in previous studies (Fant, 1997; Sundberg *et al.*, 1993; Frölich *et al.*, 2001). In these studies, the parametrization of the glottal flow has been computed by defining the closing phase as a time span between the flow maximum and the negative peak of the flow derivative. Measuring this time span, which is denoted by T_{pp} in Sundberg *et al.* (1993), implies that the so-called return phase of the flow derivative has been ignored. Among these studies, the work by Sundberg *et al.* (1993) is closest to the current survey, because it also involves analyses on the mode of phonation. Unfortunately, the results obtained by Sundberg *et al.* (1993) cannot be directly compared to those of the current study, because breathy phonation was not analyzed in their experiments and they did not normalize T_{pp} with respect to the length of the fundamental period. However, Sundberg *et al.* (1993) reported that T_{pp} tended to be shorter in pressed voices than in normal voices of the same pitch. Hence, their results are in line with those of the current study, even though the parametrization methods, both of which focus on a subsection of the glottal closing phase, are different. Interestingly, it was reported by Sundberg *et al.* (1993) that in 24% of the cases T_{pp} increased when phonation was changed to pressed. Comparison of this finding to the result of the current study, which shows that there was a monotonic change of NAQ between the phonation types for all the utterances, suggests that NAQ is able to classify the phonation type more accurately than T_{pp}.

[1]It should be noticed in Eqs. (3) and (4) that the domain of AQ is time, because this quotient is defined as a ratio between a flow value and a value of the *time derivative* of the flow. If inverse filtering is based on digital signal processing, which is typical in voice source analysis today, the values of f_{ac} and d_{peak} are usually extracted from discrete-time waveforms that are expressed using integer numbers as the time variable. In this case, AQ in Eqs. (3) and (4) needs to be divided by the sampling frequency in order to express the parameter value in seconds.

[2]The resulting length of the rectangular pulse does not, of course, depend on the absolute time location of the pulse. However, to adjust the center point of the pulse to the instant of the negative peak of the flow derivative emphasizes that d_{peak} is the only amplitude value of the flow derivative needed in the computation of T_{erp}.

[3]Modeling a spectral resonance with a digital all-pole filter requires (at least) one complex conjugate pair of poles in the z domain. Since the signal bandwidth used in the current study was 4 kHz, and vowels have on average one formant per 1 kHz, the minimum order of the vocal tract filter is eight. However, using this small order of the vocal tract model sometimes results in formant ripple due to insufficient canceling of the vocal tract resonances. In general, the distortion caused by format ripple can be reduced by increasing the order of the all-pole filter. Then again, if too large a filter order is used, the glottal flow estimate can be distorted by smoothing caused by increased low-frequency amplification of the inverse filter. In order to minimize the effects of these two distortions in the current study, the order of the vocal tract all-pole filter was adjusted separately for each utterance by always starting with a filter order of eight and by increasing it to 10 or 12 when a reduction in the amount of formant ripple was required.

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

Alku, P., and Vilkman, E. (**1994**). ''Estimation of the glottal pulseform based on discrete all-pole modeling,'' in Proceedings of the International Conference on Spoken Language Processing 1994 (Yokohama), 1619–1622.

Alku, P., and Vilkman, E. (**1996a**). ''A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers,'' Folia Phoniatr Logop **48**, 240–254.

Alku, P., and Vilkman, E. (**1996b**). ''Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering,'' Speech Commun. **18**, 131–138.

Carlson, B. (**1986**). *Communication Systems* (McGraw-Hill, Singapore), pp. 177–178.

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

Childers, D. G., and Lee, C. K. (**1991**). ''Vocal quality factors: Analysis, synthesis, and perception,'' J. Acoust. Soc. Am. **90**, 2394–2410.

Dromey, C., Stathopoulos, E. T., and Sapienza, C. M. (**1992**). ''Glottal airflow and electroglottographic measures of vocal function at multiple intensities,'' J. Voice **6**, 44–54.

El-Jaroudi, A., and Makhoul, J. (**1991**). ''Discrete all-pole modeling,'' IEEE Trans. Signal Process. **39**, 411–423.

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

Fant, G. (**1995**). ''The LF-model revisited. Transformations and frequency domain analysis,'' Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm **2–3**, pp. 119–156.

Fant, G. (**1997**). ''The voice source in connected speech,'' Speech Commun. **22**, 125–139.

Fant, G., Kruckenberg, A., Liljencrants, J., and Båvegård, M. (**1994**). ''Voice source parameters in continuous speech. Transformation of LF-parameters,'' in Proceedings of the International Conference on Spoken Language Processing 1994 (Yokohama), pp. 1451–1454.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). ''A four-parameter model of glottal flow,'' Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm **4**, pp. 1–13.

Fant, G., and Lin, Q. (**1988**). ''Frequency domain interpretation and derivation of glottal flow parameters,'' Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm **2–3**, pp. 1–21.

Fröhlich, M., Michaelis, D., and Strube, H. (**2001**). ''SIM—Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals,'' J. Acoust. Soc. Am. **110**, 479–488.

Hertegård, S., Gauffin, J., and Karlsson, I. (**1992**). ''Physiological correlates of the inverse filtered flow waveform,'' J. Voice **6**, 224–234.

Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (**1988**). ''Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice,'' J. Acoust. Soc. Am. **84**, 511–529.

Howell, P., and Williams, M. (**1992**). ''Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech,'' J. Acoust. Soc. Am. **91**, 1697–1706.

Monsen, R. B., and Engebretson, A. M. (**1977**). ''Study of variations in the male and female glottal wave,'' J. Acoust. Soc. Am. **62**, 981–993.

Moore, B. C. (**1982**). *An Introduction to the Psychology of Hearing* (Academic, London), p. 82.

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

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

Sulter, A. M., and Wit, H. P. (**1996**). ''Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age,'' J. Acoust. Soc. Am. **100**, 3360–3373.

Sundberg, J., Andersson, M., and Hultqvist, C. (**1999**). ''Effects of subglottal pressure variation on professional baritone singers' voice sources,'' J. Acoust. Soc. Am. **105**, 1965–1971.

Sundberg, J., Titze, I., and Scherer, R. (**1993**). ''Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source,'' J. Voice **7**, 15–29.

Wilks, S. S. (**1962**). *Mathematical Statistics* (Wiley, New York), p. 74.

Wong, D. Y., Markel, J. D., and Gray, Jr., A. H. (**1979**). ''Least-squares glottal inverse filtering from acoustic speech waveforms,'' IEEE Trans. Acoust., Speech, Signal Process. **27**, 350–355.