Perceived loudness of speech based on the characteristics of glottal excitation source

Guruprasad Seshadri[a)]
Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India

B. Yegnanarayana
International Institute of Information Technology, Hyderabad 500032, Andhra Pradesh, India

(Received 6 June 2008; revised 10 July 2009; accepted 13 July 2009)

The impulse-like characteristic of glottal excitation in speech production is an important factor in the perception of loudness of speech signals. This characteristic is attributed to the abruptness of the closing phase in the glottal cycle. In this paper, an acoustic feature, called strength of excitation, is proposed to represent the impulse-like nature of excitation. The strength of excitation is derived from the linear prediction residual of speech signals, where the residual can be considered as an estimate of the source of excitation. Since the loudness of speech is perceived over one or more utterances of speech, it is hypothesized that the distribution of strength of excitation is indicative of the perceived loudness of speech. The distribution of strength of excitation is shown to distinguish between soft and loud utterances of speakers. The distribution can also help in discriminating between the loudness of two speakers. The loudness measure obtained using the distribution of the strength of excitation is in agreement with the subjective judgment of loudness of speech.
 [DOI: 10.1121/1.3203668]

PACS number(s): 43.71.Gv, 43.66.Cb [DOS] Pages: 2061–2071

I. INTRODUCTION

Perception of loudness of sound in human beings is defined as the magnitude of auditory sensation, which depends on the acoustic characteristics of the sound (Fletcher and Munson, 1933). Loudness of a sound is related to the distribution of spectral energy of the sound (Fletcher and Munson, 1933; Fletcher and Munson, 1937). Temporal properties of sounds such as duration, and impulsive or rhythmic nature, also affect the perceived loudness (Zwicker, 1977; Zwicker and Fastl, 1999). The problem of measurement and calculation of loudness of sounds has been studied extensively. A method for calculating the loudness of a complex tone from its frequency spectrum was proposed by Fletcher and Munson (1933). For sounds with a large number of spectral components, the loudness of one component depends on the masking effects of the other components, particularly when the components are closely spaced (Fletcher and Munson, 1937). Methods for computing the total loudness of a sound from the values of loudness due to the constituent frequency bands were suggested by Beranek *et al.* (1951) and Stevens (1956, 1961). Similarly, Zwicker and Fastl (1999) described a procedure to compute the loudness levels of pure tones at different frequencies, and to construct equal loudness contours. In the case of pure tones or noise that has a uniform distribution of energy in a given band of frequencies, Zwicker and Fastl (1999) proposed the measurement of loudness as a function of frequency separation of two pure tones, and also as a function of bandwidth of the noise.

In applying the above methods for the calculation of loudness of *speech signals*, two issues need to be taken into account. First, the measurement of loudness in those methods was based on the response of the auditory perception mechanism to sounds which were pure tones, combination of pure tones, or noises of different bandwidths. But speech sounds cannot be approximated by such signals. In particular, speech sounds cannot be modeled well by pure tones (Warren, 1973), since the short-time spectrum of speech signal has a gross envelope with a few prominent peaks (formants) around which significant energy is concentrated, and a fine structure corresponding to the fundamental frequency and its harmonics (Fant, 1960). Second, loudness is a perceptual attribute which cannot be described merely by the amount of acoustic energy or its spectral distribution. For instance, a soft voice is perceived as soft, even if the level of speech from loudspeaker is increased. Similarly, a loud voice is perceived as loud, even in the presence of some amount of ambient noise. Thus, apart from the amplitude or energy of the speech signal, the excitation source and the vocal tract system characteristics in the signal can also affect the perception of loudness (Rothenberg, 1983). Hence, there is a need to examine the perception of loudness of speech based on the production characteristics of speech signals.

Since loudness of speech is a perceptual attribute, an exact definition of loudness is elusive. On a perceptual scale, loudness of speech varies from a weak/soft voice, to a normal/modal voice, and further to a loud voice, which can extend up to shouting. This aspect of loudness is determined, to a great extent, by the physiological characteristics of speech production mechanism of the speaker. The perceived loudness depends on the nature of the speech sound, due to loading of the vocal tract system on the vocal source during

[a)]Author to whom correspondence should be addressed. Electronic mail: gurus@cse.iitm.ernet.in

the production. Loudness is also affected by the behavioral characteristics of the speaker, such as emotional state of the speaker. The behavioral characteristics can cause variations in rhythm and rate of speech. They can also result in stress on particular syllables of words, and the stress may vary for different words in a sentence/phrase. Accentuation of stressed syllables, which causes a change in the pitch pattern of the stressed syllables relative to the nonstressed syllables, is also a factor in the perception of loudness. Thus, loudness is a perceptual attribute that is governed by both physiological and the behavioral characteristics of the speaker, and is perceived over a duration of one or more utterances of speech. This paper attempts to provide a quantitative measure of the loudness of speech. The proposed measure is governed by the characteristics of the source of excitation of the speech signal, and is derived from an estimate of the source of excitation.

During production of speech, the identity of the speech sounds is governed mostly by the configuration of the vocal tract system. The size and shape of the vocal tract are dictated by the positions of the articulators. For the same configuration of the vocal tract system, loudness can be varied by varying the characteristics of glottal excitation. The characteristics of glottal excitation in speech are (a) impulse-like nature during glottal closure and (b) quasiperiodic nature of the impulse-like excitation in voiced sounds. In this paper, the impulse-like excitation is represented, both by the amplitude and the "strength" of the impulse-like excitation. The amplitude is estimated at the instant of the glottal closure, whereas the strength is based on the spread of the impulse-like excitation around the instant of the glottal closure. The notion of strength of excitation is explained as follows: Greater strength is associated with an excitation when a given amount of energy is concentrated in a short duration of time, than when the same energy is spread over a longer duration of time. For a given vocal tract system, an ideal impulse excitation can be said to have maximum strength, whereas white noise excitation of the same energy as the impulse, but spread over time, has the least strength. This paper proposes methods for estimating the amplitude and strength of excitation, and relates the perception of loudness of speech to the strength of excitation.

II. TRADITIONAL MEASURES OF LOUDNESS OF SPEECH

Measures of loudness of speech have been proposed based on the physiological characteristics of speech production and on the acoustic characteristics of speech signal. Perceptual judgments of loudness, indicated by labels such as "soft," "normal," and "loud," serve as a reference for comparison with the physiological and acoustic measures of loudness. Measurements of sound pressure level (Lane *et al.*, 1961; Ladefoged and McKinney, 1963; Allen, 1971; Orlikoff, 1991; Sulter and Wit, 1996; Holmberg *et al.*, 1988) and subglottal pressure level (Ladefoged and McKinney, 1963; Allen, 1971; Sundberg *et al.*, 2005) were observed to be strongly correlated with the perceptual judgments of loudness. Features derived from the measurements of glottal waveform have been studied for their effect on the perceived loudness of speech. Glottal volume velocity (Monsen and Engebretson, 1977), intraoral air pressure, oral airflow, sound pressure (Holmberg *et al.*, 1988), and electroglottographic (EGG) signals (Orlikoff, 1991) were obtained for male and female subjects for different levels of loudness such as soft, normal, and loud. In Holmberg *et al.* (1988), the subglottal air pressure and glottal airflow were derived from the measurements of the intraoral air pressure, oral airflow, and sound pressure. The maximum airflow declination rate (MFDR), defined as the maximum amplitude of the negative peak in the first derivative of the glottal volume velocity, was observed to be significantly lower in soft voices than in normal and loud voices (Holmberg *et al.*, 1988; Sundberg *et al.*, 2005). For each loudness condition, MFDR was observed to be highly correlated with the sound pressure level (Holmberg *et al.*, 1988). The MFDR parameter was found to increase linearly with subglottal pressure (Sundberg *et al.*, 2005). Soft voice had a more symmetrical waveform of the glottal volume velocity in general (in closing and opening phases in the glottal cycle), compared to loud voice (Monsen and Engebretson, 1977; Orlikoff, 1991). The closing portion of the glottal waveform was more abrupt for loud voices than that for normal and soft voices (Monsen and Engebretson, 1977). The less abrupt closure of the glottis in soft voices was also observed to be responsible for less energy in the high-frequency regions relative to the energy in the low-frequency regions (Holmberg *et al.*, 1988). The slope of the EGG signal in the closing phase in the glottal cycle was observed to be proportional to the amplitude of the acoustic speech signal (Orlikoff, 1991). These observations based on physical measurements of the glottal waveform provide motivation for deriving similar features from the glottal waveform estimated by inverse filtering the acoustic speech signal.

Measures of loudness of speech derived from the acoustic speech signal are based primarily on the characteristics of vibrations of the vocal folds. Glottal flow waveform estimated by inverse filtering the acoustic speech signal is parametrized to obtain such measures. Of these measures, the MFDR showed a significant increase from soft to loud levels (Sulter and Wit, 1996; Gauffin and Sundberg, 1989). Also, a strong correlation was observed between MFDR and spectral tilt (Gauffin and Sundberg, 1989). A strong correlation was also observed between the abruptness of the closing phase in the glottal cycle and the spectral tilt (Gauffin and Sundberg, 1989), where the former is related to the decrease in the rate of flow during the final part of the closing phase. Spectrum of the glottal source showed lesser roll-off for loud voice, compared to normal voice (Cairns and Hansen, 1994). Doval *et al.* (2006) observed that the maximum value of glottal excitation controlled the mid-to-high-frequency spectral slope in spectrum of the glottal flow waveform. Other parameters include open quotient (OQ) (i.e., proportion of the duration of the cycle for which the glottis is open), closed quotient (CQ) (i.e., proportion of the duration of the cycle for which the glottis is closed), closing quotient (ClQ) (i.e., proportion of the duration of the closing phase in each cycle), and speed quotient (SQ) (i.e., the time taken for the vocal folds to open, divided by the time taken for them to close). While the OQ decreased from soft to loud voices (Dromey

et al., 1992), the CQ was observed to increase from soft to loud voices (Sulter and Wit, 1996). The SQ showed an increase from soft to normal voices, and a decrease from normal to loud voices (Dromey *et al.*, 1992; Sulter and Wit, 1996). The ClQ was observed to be lowest for normal voice, and increased for both soft and loud voices (Sulter and Wit, 1996; Bäckström *et al.*, 2002). Cummings and Clements (1995) observed that the closing slope of the glottal waveform was significantly higher for loud voice compared to normal and soft voices. Also, the closing duration was significantly smaller for loud voice compared to normal and soft voices. By contrast, the opening slope and the opening duration did not show specific trends or significant differences among soft, normal, and loud voices. Bäckström *et al.* (2002) defined a parameter called amplitude quotient (AQ) as the ratio of the maximum amplitude of the glottal flow and the negative peak of the differentiated glottal flow. Normalized AQ, defined as the AQ normalized by the period of vibration, was observed to decrease with increase in vocal intensity (represented by sound pressure level). Alku *et al.* (2006) observed the variation in AQ as a function of MFDR, by varying the vocal intensity from "very soft" to "extremely loud." The AQ-MFDR curve showed a rapidly decreasing trend in the soft-normal range, followed by convergence toward a horizontal line for higher levels of loudness.

Configuration of the vocal tract may also undergo changes during the production of loud voices. Schulman (1989) observed that the patterns of movement of lips and jaws in loud speech (measured by displacement, velocity and relative timing associated with the movement) were amplified compared to normal speech. Spectral features derived from the speech signal, such as spectral tilt, changes in the formant frequencies and their bandwidths, and richness of the short-time spectrum as indicated by harmonicity of the spectrum, have been proposed as measures of loudness of speech. Ternström *et al.* (2006) defined a feature called "spectrum balance," as the energy in the high-frequency band (2–6 kHz) relative to that in the low-frequency band (0.1–1 kHz). This feature, when averaged across several segments of similar vowels, increased from soft to loud voice, but the rate of increase slowed down, or even stopped altogether, at very high levels of loudness. Very loud speech is mostly accompanied by a relative increase in the low-frequency energy, in the form of a sharper spectral peak at the first formant. Sundberg and Nordenberg (2006) defined alpha measure as the ratio of spectrum intensity above and below 1000 Hz, which was observed to increase linearly with sound energy level, corresponding to the increasing levels of loudness. Cairns and Hansen (1994) observed significant shifts in the formant frequencies and their bandwidths for loud voice, compared to normal voice. Gramming and Sundberg (1988) observed that the fundamental frequency was the strongest spectral component in soft voice, while it was typically a harmonic of the fundamental in loud voice. Moreover, the spectrum in loud voice was harmonically richer (as measured by the number of harmonics of the fundamental in a given frequency band, and their spectral intensities), compared to soft voice which had very few harmonics. For the vowel /a/, the first formant frequency was generally observed to be lower in the soft voice, compared to that in the loud voice.

Production of loudness in speech is also associated with vocal effort of the speaker. While the term vocal effort is not defined, Pickett (1956) described the range of vocal effort from "weakest voiced whisper" to "loudest possible shout." Allen (1971), and Glave and Rietveld (1975) observed that an increase in the vocal effort resulted in a corresponding increase in the perceived loudness. Glave and Rietveld (1975) observed that, between the sounds produced with effort and those produced without effort, a constant difference existed in the perceived loudness and also in the loudness calculated based on Zwicker's model. Traunmüller and Eriksson (2000) defined vocal effort in terms of the distance from the speaker as estimated by a group of listeners for a given utterance, in the context of communication over a range of distances. In general, increased vocal effort results in an increase in the energy level, the spectral emphasis (an acoustic feature reflecting the relative intensity in the higher frequency bands), the fundamental frequency, and the first formant (Traunmüller and Eriksson, 2000). Liénard and Benedetto (1999) observed that the fundamental frequency and the first formant were highly correlated with the vocal effort, while the second and third formants did not vary significantly. Also, the spectral emphasis and the amplitudes of the first three formants increased significantly with increase in the vocal effort.

The fundamental frequency of glottal vibration also reflects the variations in loudness. Studies by Harris and Weiss (1964), Lieberman *et al.* (1969), and Monsen and Engebretson (1977) have shown that, in general, there is an increase in the fundamental frequency for loud speech when compared to soft and normal speech. Holmberg *et al.* (1988) too observed that, in general, loud voice was produced with a higher fundamental frequency than that of the normal voice, whereas the fundamental frequency in a soft voice was either higher or lower compared to that in a normal voice. Alku *et al.* (2002) argued that speakers, while producing loud voice, increased the fundamental frequency to increase the number of glottal closures per unit time. This increased rapid fluctuations in the speech pressure waveform, thereby increasing the vocal intensity. Loud speech is also accompanied by an increase in the durations of vowels, diphthongs, and words (Cairns and Hansen, 1994; Traunmüller and Eriksson, 2000). However, not all increase in the fundamental frequency can be associated with an increase in loudness. For instance, speakers can keep the pitch steady and yet produce varying degrees of vocal loudness (Sundberg *et al.*, 2005). The patterns of vibrations of the vocal folds also reflect other features such as rhythm and rate of speech, and the accentuation of stressed syllables (Johnstone and Scherer, 1999; Ladd *et al.*, 1994). Thus, change in the fundamental frequency is an effect of the change in loudness rather than a cause of it.

Some of the acoustic features described above have also been used to characterize the labels of voice quality (Laver, 1994), such as creakiness, breathiness (Klatt and Klatt, 1990; Childers and Lee, 1991), falsetto (Childers and Lee, 1991),

hoarseness, and roughness (Eskenazi *et al.*, 1990). These labels of voice quality have been defined based on the configurations of the laryngeal system (Laver, 1994). Similarly, labels of voice quality such as tense and lax are described according to the degree of muscular tension in the laryngeal and supralaryngeal systems, while nasality is described based on the articulatory settings (Laver, 1994). By contrast, loudness cannot be characterized on a physiological basis alone. Moreover, a degree of perceived loudness can be associated with all the above labels of voice quality. For instance, tense voice sounds intrinsically louder than lax voice. Also, loudness is not an exclusive feature, in the sense that each voice quality can be realized with varying degrees of loudness. Thus, loudness can be viewed as an underlying feature that can be varied independently of the voice quality.

In summary, features of the glottal vibration play an important role in the production of vocal loudness. Two features of the glottal vibration are significant, namely, the amplitude of the negative peak of the differentiated glottal flow and the abruptness of the closing phase in the glottal cycle. These features are reflected in other acoustic measures such as sound energy level, spectral tilt, harmonic richness of the short-time spectrum, and, to an extent, in the sharpness of the formant peaks. However, the short duration of the impulse-like excitation in time is not captured well in the spectrum. Moreover, the estimation of the features of the glottal wave is dependent on the method of parametrization of the glottal flow waveform and the accuracy of the parameters. In view of this, an *acoustic* feature called *strength of excitation* is proposed in this paper, which can be derived from the inverse filtered signal. The motivation for deriving such a feature stems from the abruptness of the glottal closure, as illustrated in Sec. III. Computation of the proposed feature of strength of excitation is described in Sec. IV. Studies described in Sec. V show that the distribution of strength of excitation is related to the perception of loudness.

III. SIGNIFICANCE OF ABRUPTNESS OF GLOTTAL CLOSURE

A. Speech material

Speech utterances of a male speaker, spoken with varying levels of loudness such as soft, normal, and loud, were chosen from VOQUAL'03 database (d'Alessandro and Scherer, 2003). The following sentence was uttered by the speaker five times in each level of loudness: She has left for a great party today. The speech signals and the corresponding electroglottograph signals were sampled at 44.1 kHz, and both the signals were synchronized in time.

B. Measure of abruptness of glottal closure

Figure 1 shows segments of speech signals within one pitch period, and the differentiated EGG (DEGG) signals, corresponding to soft, normal, and loud utterances. The segments are shown for the vowel /a/ in the word "party." It is observed from the DEGG signals in Figs. 1(b), 1(d), and 1(f) that the abruptness of the glottal closure increases from soft to loud utterances. The abruptness of the glottal closure is reflected in the rate of decay of the DEGG signal from

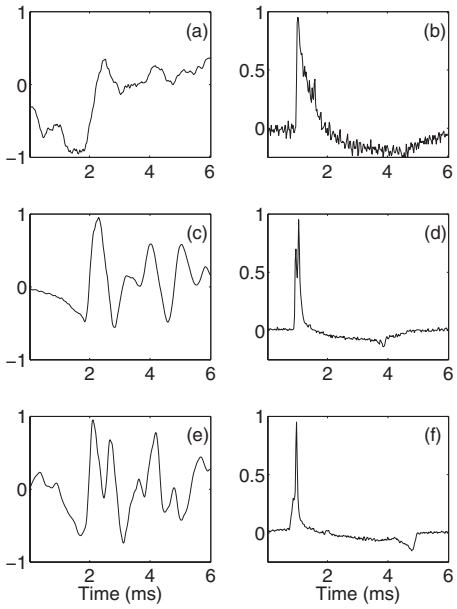

FIG. 1. Illustration of the abruptness of the glottal closure for soft, normal, and loud utterances. Speech segments within one pitch period are shown in (a), (c), and (e), which belong to soft, normal, and loud utterances, respectively. The segments correspond to the vowel /a/ in the word party in the sentence: She has left for a great party today. (b), (d), and (f) show the DEGG signals corresponding to (a), (c), and (e), respectively.

around the instant of the glottal closure. It is necessary to quantify the abruptness of the glottal closure to study its relationship with perceived loudness. First, the instants of glottal closure are estimated from the speech signal using a method described by Smits and Yegnanarayana (1995), which is based on the properties of minimum phase signals and group delay functions. A segment of 1 ms following the instant of glottal closure is considered from the DEGG signal. This segment is normalized by dividing the samples in the segment by the amplitude of the largest sample. The segment is approximated by a decaying exponential of the form $g(t)=e^{-t/\tau}$. Here the parameter τ denotes the time constant, and t denotes time. Let the samples of the segment of DEGG be denoted by $x[i]$, $i=0,1,2,\ldots,N-1$. Let t_i denote the time instant corresponding to the ith sample. It is assumed that $t_0=0$. Then, the parameter τ is estimated using the method of least squares as follows:

$$\tau^* = \arg\min_{\tau} \sum_{i=0}^{N-1} \|x[i] - e^{-t_i/\tau}\|^2. \quad (1)$$

The time constant τ^* indicates the abruptness of the glottal closure. An abrupt closure of the glottis corresponds to a faster decay of the exponential, resulting in a smaller value of the time constant τ^*. A relatively gradual closure of the glottis corresponds to a slower decay of the exponential, resulting in a larger value of the time constant τ^*.

C. Results

The values of the time constant τ^* are computed from the EGG signals of the soft, normal, and loud utterances. The distribution of the parameter τ^* is shown in Fig. 2 for the three levels of loudness. The distribution shows significant

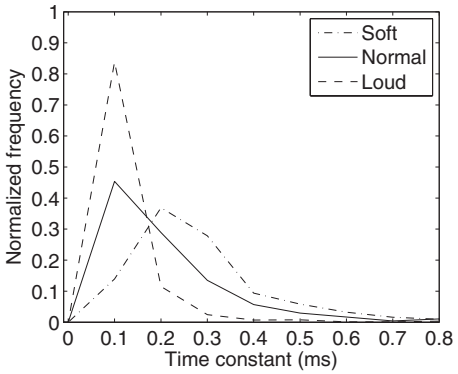

FIG. 2. Distribution of the time constant τ^* for soft, normal, and loud utterances.

variation between the soft and the loud utterances. The distribution of τ^* for the normal utterances overlaps considerably with those of both soft and the loud utterances, indicating that natural speech consists of segments of varying levels of loudness. In practice, only the speech signals are available, and not the EGG signals. Hence it is necessary to derive a measure of abruptness of glottal closure from the acoustic speech signal.

IV. MEASURES OF AMPLITUDE AND STRENGTH OF EXCITATION

To derive the features of amplitude and strength of excitation from the source of excitation of speech signal, a representation of the source of excitation is discussed in Sec. IV A. Methods for estimating the amplitude and the strength of excitation are described in Secs. IV B and IV C, respectively.

A. Representation of source of excitation

In order to characterize the impulse-like nature of excitation, an estimate of the source of excitation needs to be derived from the speech signal. Linear prediction (LP) residual can be used to approximate the source of excitation (Makhoul, 1975). LP residual is obtained by passing the speech signal through the inverse filter estimated during the LP analysis. Figures 3(a) and 3(b) show a segment of speech signal and its LP residual, respectively. LP analysis was performed on overlapped segments of speech signal (size of frame=25 ms, frame shift=5 ms, LP order=10, and sampling frequency=8 kHz). The prediction error in each glottal cycle is usually large around the instant where impulse-like excitation takes place. This happens around the instant of glottal closure for each glottal cycle due to abruptness of the closure. This is manifested as large amplitude fluctuations (both positive and negative) in the LP residual. The detection of these regions of large error in the LP residual is difficult because of the amplitude values with either polarity occurring around the instants of glottal closure. This difficulty can be overcome by using the Hilbert envelope of the LP residual (Ananthapadmanabha and Yegnanarayana, 1979). The Hilbert envelope $r[n]$ of the LP residual $e[n]$ is given by

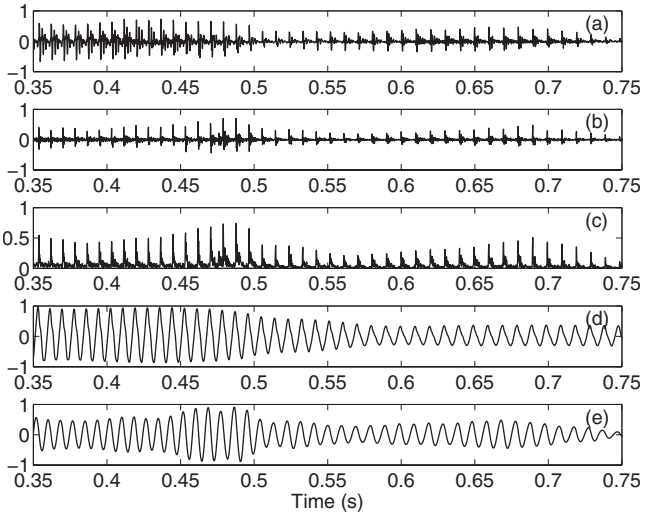

FIG. 3. (a) A segment of speech signal, (b) its LP residual, (c) Hilbert envelope of the LP residual, (d) filtered signal $y_s[n]$ derived from the speech signal, and (e) filtered signal $y_r[n]$ derived from the Hilbert envelope of the LP residual.

$$r[n] = \sqrt{e^2[n] + e_H^2[n]}, \tag{2}$$

where $e_H[n]$ denotes the Hilbert transform of $e[n]$. The Hilbert transform $e_H[n]$ of the signal $e[n]$ is given by

$$e_H[n] = \mathrm{IFT}(E_H(\omega)), \tag{3}$$

where IFT denotes the inverse Fourier transform, and $E_H(\omega)$ is given by (Oppenheim and Schafer, 1975)

$$E_H(\omega) = \begin{cases} +jE(\omega), & \omega \le 0 \\ -jE(\omega), & \omega > 0. \end{cases} \tag{4}$$

Here $E(\omega)$ denotes the Fourier transform of the signal $e[n]$. The impulse-like nature of excitation can be observed clearly from the Hilbert envelope of the LP residual, as shown in Fig. 3(c). The amplitude of the excitation can be estimated by detecting the instants of the glottal closure, and then measuring the peaks in the Hilbert envelope of LP residual around the instants. Another approach for estimating the amplitude of the excitation is proposed in Sec. IV B.

B. Estimation of amplitude of impulse-like excitation

1. Computation of filtered signal

The impulse-like excitation is due to abruptness of the glottal closure in each cycle. The characteristics of the sequence of impulse-like excitations are reflected across all the frequencies in the speech signal including 0 Hz (hereafter referred to as zero frequency). Filtering the speech signal through a resonator located at zero frequency helps in emphasizing the characteristics of excitation (Murty and Yegnanarayana, 2008). The system function of such a resonator is given by

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}, \tag{5}$$

where $a_1 = -2$ and $a_2 = 1$. The above resonator de-emphasizes the characteristics of the vocal tract, since the resonances of the latter are located at much higher frequencies than the

zero frequency. A cascade of two such resonators, given by the system function $G(z)=H(z)H(z)$, is used to reduce the effect of all the resonances of the vocal tract. Let $s[n]$ denote the input speech signal. Then the output $x_s[n]$ of the cascade of resonators is given by

$$x_s[n] = s[n] \star g[n], \tag{6}$$

where $g[n]$ is the impulse response of the system function $G(z)$ and $\star$ denotes convolution operation. The output $x_s[n]$ contains mainly the features of glottal vibrations. Filtering the signal $s[n]$ through the cascade of resonators causes the output $x_s[n]$ to grow as a polynomial function of time. This trend in $x_s[n]$ is removed by subtracting the average of sample values over a window of 10 ms (approximately 0.5–1.5 times the estimated pitch period). The resulting trend-removed signal $y_s[n]$ is given by (Murty and Yegnanarayana, 2008)

$$y_s[n] = x_s[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} x_s[n+k], \tag{7}$$

where $2N+1$ is the size (in samples) of the window. The signal $y_s[n]$ is called the *filtered signal*, an example of which is shown in Fig. 3(d) for the segment of voiced speech in Fig. 3(a).

2. Slope of positive-to-negative zero crossings

Murty and Yegnanarayana (2008) observed that the locations of positive-to-negative zero crossings (PNZCs) of the filtered signal $y_s[n]$ provide an accurate estimate of the instants of glottal closure. It is observed that the filtered signal $y_s[n]$ is free of the characteristics of the vocal tract system. The filtered signal can also be derived from the Hilbert envelope $r[n]$ of the LP residual, instead of the speech signal $s[n]$. Let $y_r[n]$ denote the filtered signal derived from $r[n]$. Figure 3(e) shows $y_r[n]$ for the segment of voiced speech in Fig. 3(a). We note from Figs. 3(d) and 3(e) that the locations of the PNZCs derived from $y_r[n]$ are nearly the same as those derived from $y_s[n]$.

A strong peak in the Hilbert envelope $r[n]$ has a corresponding PNZC in the filtered signal $y_r[n]$. It is observed from Fig. 3(e) that the slope of $y_r[n]$ at a PNZC is proportional to the amplitude of the corresponding peak in the Hilbert envelope $r[n]$. The slope of $y_r[n]$ at a PNZC is estimated by considering a region of 0.125 ms on either side of the PNZC, by assuming $y_r[n]$ to be linear in the vicinity of each PNZC. To observe the relationship between the slope of $y_r[n]$ at a PNZC and the amplitude of the corresponding peak in $r[n]$, speech signals collected from 50 female and 50 male speakers of TIMIT database (Garofalo *et al.*, 1993) were processed. For each speaker, ten spoken utterances were used, whose durations ranged from 2 to 5 s. Only voiced segments were processed. The scatter plots in Fig. 4 illustrate the linear dependence of the amplitude of the peak of $r[n]$ and the slope of the corresponding PNZC in $y_r[n]$. Both these quantities, which form an ordered pair in the scatter plots, are associated with an instant of glottal closure. Thus the number of points in the scatter plot shown in Fig. 4(a) [Fig. 4(b)] denotes the number of glottal closures in 500 utterances [50 female (male) speakers $\times$ 10 utterances per speaker]. For instance, if we assume that each of the 500 analyzed utterances contains 1 s of voiced speech, with an average pitch period of 4 ms (8 ms) for the female (male) speakers, then the number of glottal closures amounts to 125 000 (62 500). The actual number of points in the scatter plot shown in Fig. 4(a) is 156 264, while that in the scatter plot shown in Fig. 4(b) is 69 359. The ordinate in Fig. 4 shows only the magnitude of the slope of $y_r[n]$ at a PNZC, although the slope itself is negative. The values of the correlation coefficient computed from the sets of points in Figs. 4(a) and 4(b) are 0.92 and 0.94, respectively. The values of the correlation coefficient computed for different speakers ranged from 0.90 to 0.98. Note the approximate *linear* relation between the amplitude of excitation and the slope at PNZC, even though the gross slopes of the lines are different for female and male speakers due to differences in their average pitch periods.

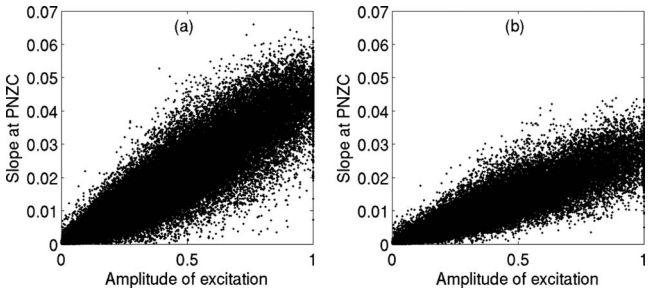

FIG. 4. Scatter plots to illustrate the linear dependence of the amplitude of excitation and the slope of PNZC in the filtered signal $y_r[n]$ for (a) 50 female speakers and (b) 50 male speakers.

C. A measure of strength of excitation

Figure 5 shows segments of voiced speech, chosen from the utterances of soft, normal, and loud voices. The impulse-like excitation, as observed from the LP residuals of the

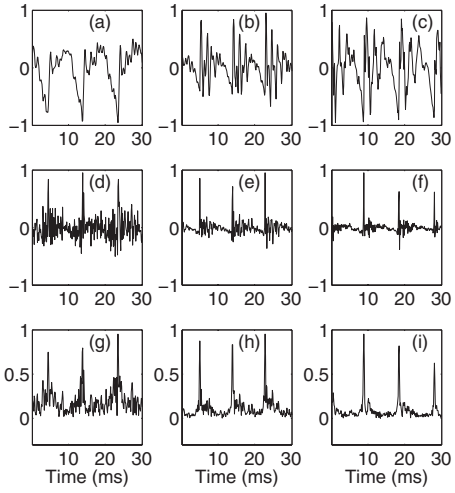

FIG. 5. Illustration of the nature of excitation in soft, normal, and loud utterances. Speech segments in (a)–(c) belong to soft, normal, and loud utterances, respectively. The segments correspond to the vowel /a/ in the word party in the sentence: She has left for a great party today. (d)–(f) show the LP residual for the signals in (a)–(c), respectively, while the figures (g)–(i) show the Hilbert envelopes of the corresponding LP residuals.

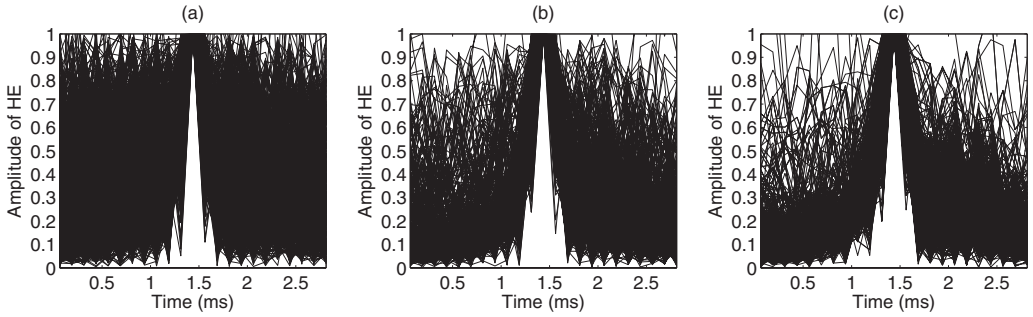

FIG. 6. Segments of Hilbert envelope of LP residual in the vicinity of impulse-like excitations for (a) soft, (b) normal, and (c) loud utterances.

speech segments, is more spread out in time for the soft utterances [Fig. 5(d)], compared to normal [Fig. 5(e)] and loud [Fig. 5(f)] utterances. The impulse-like nature of the glottal excitation can be observed clearly from the Hilbert envelope of the LP residual, shown in Figs. 5(g)–5(i) for soft, normal, and loud utterances, respectively. A measure of the strength of excitation can be derived from a short segment (1–3 ms) around the instants of impulse-like excitation. Figure 6(a) shows superimposed segments of the Hilbert envelopes around these instants derived from soft utterances. Each segment has a duration of 3 ms, and the location of the peak in the Hilbert envelope of the LP residual is at the center of the segment. Each segment is normalized by dividing the samples of the segment by the largest amplitude in the segment. All the segments are derived from the voiced regions of five soft utterances of a male speaker in the VOQUAL'03 database and are superimposed, as shown in Fig. 6(a). Similar plots are obtained for normal and loud utterances, as shown in Figs. 6(b) and 6(c), respectively. These plots show that for soft utterances, the Hilbert envelope of the LP residual is spread out more uniformly on either side of the instant of impulse-like excitation. This indicates that the impulses around the instants of impulse-like excitation are not sharp in these cases. The impulses are much sharper for loud utterances than for soft or even normal utterances.

To represent the sharpness in the Hilbert envelope of the LP residual, a feature called strength of excitation is defined as $\eta=\sigma/\mu$, where μ denotes the mean of the samples of the Hilbert envelope of the LP residual in a segment around the instant of impulse-like excitation, and σ denotes the standard deviation of the samples. For a segment of length N consisting of an ideal impulse (in discrete-time domain) of amplitude a, $\eta=\sqrt{N}$. For a segment of length N consisting of samples of equal amplitude $a/\sqrt{N}$, $\eta=0$. The segment in this case has the same energy as that of the ideal impulse of amplitude a. This case represents the maximum deviation from an ideal impulse. Thus the value of η lies between 0 and $\sqrt{N}$ for any segment, irrespective of the amplitudes of the samples in the segment. A higher value of η indicates greater strength of excitation. In general, a segment having impulse-like characteristics in excitation, as in the case of a loud voice, has a smaller value of μ and a larger value of σ, resulting in a larger value of η. By contrast, a soft voice with greater spread around the center has a larger value of μ and a smaller value of σ, resulting in a smaller value of η.

The strength η of excitation is computed for soft, normal, and loud utterances of the male speaker in the VOQUAL'03 database. Figure 7 shows the distribution of η for the three types of utterances. The plot indicates that for soft voice, the distribution has greater concentration of lower values of η, whereas for loud voice, the distribution is concentrated around larger values of η. Normal and loud utterances can also be distinguished by comparison, since the proportion of larger values of η is higher in loud utterances compared to that in the normal utterances. Also, the discrimination between soft and loud utterances in the distribution of η is comparable to that based on the distribution of the time constant parameter derived from DEGG signals (Fig. 2). Therefore, the distribution of the strength of excitation can be used as a measure of the perceived loudness of a given speech signal. The distribution of η can be used to identify soft and loud segments in the speech of a given speaker. The distribution can also help in inferring some gross speaker-specific characteristics, as discussed in the next section.

V. EVALUATION OF THE EFFECTIVENESS OF STRENGTH OF EXCITATION

In this section, the ability of the distribution of the strength (η) of excitation to distinguish between the levels of loudness within individual speakers is examined. The distribution is also used for comparing the loudness of speech from two different speakers. It is examined whether the distribution of η is in agreement with the subjective judgment of loudness of speech.

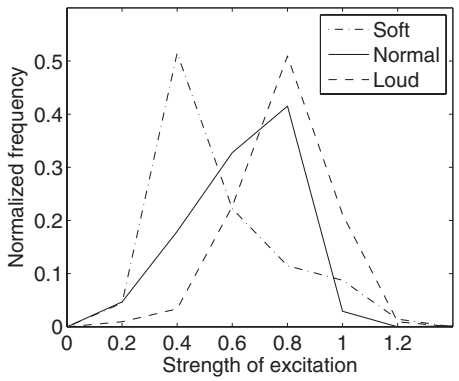

FIG. 7. Distribution of the strength (η) of excitation for soft, normal, and loud utterances.

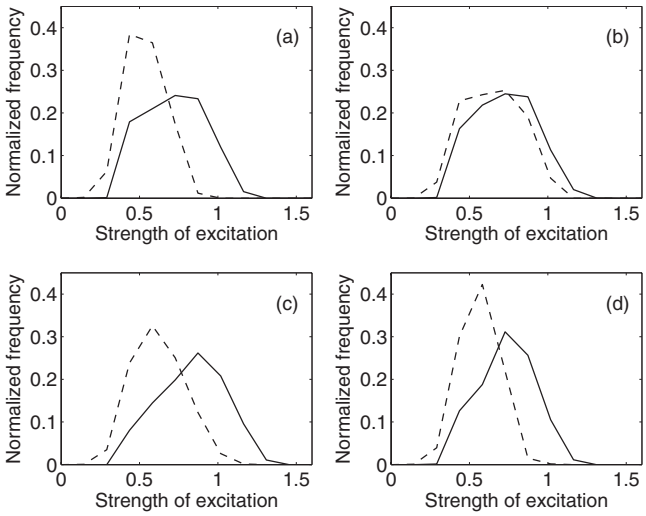

FIG. 8. Distribution of the strength (η) of excitation for four speakers. In each case, the broken and the solid lines correspond to soft and loud utterances, respectively. (a) and (b) correspond to two female speakers, while (c) and (d) correspond to two male speakers.

A. Speaker-specific nature of loudness

1. Speech material

Speech signals were collected from 44 speakers (13 female speakers and 31 male speakers) in two levels of loudness, namely, soft and loud. The speakers were undergraduate and graduate students, aged between 17 and 26 years. All the speakers spoke Indian English, and the native language of each speaker was one among Telugu, Hindi, Kannada, Tamil, Marathi, and Oriya. The speakers were guided to listen to the soft and the loud utterances of the VOQUAL'03 database, so as to help them produce the two levels of loudness while maintaining naturalness of their speech. Each speaker uttered 20 sentences in soft level and in loud level. The durations of these sentences ranged from 1 to 5 s. The speech signals were sampled at 8 kHz.

2. Results

The collected speech signals were analyzed, and the distribution of the strength (η) of excitation was computed for the soft and loud utterances of each speaker. Figure 8 shows the distribution of η for two female and two male speakers, chosen at random from among the 44 speakers for illustration. It is observed that the distribution of η does discriminate between the soft and loud utterances of the speakers. The degree of discrimination, or separation between the distributions of η for soft and loud utterances, is speaker dependent. For instance, the separation between the distributions is less in Fig. 8(b), compared to that in Figs. 8(a), 8(c), and 8(d). Some speaker-specific characteristics can be inferred from these figures. For instance, the speaker in Fig. 8(b) is not able to produce utterances which are significantly louder than the soft utterances. The distribution of η may also indicate the range of loudness that can be produced by a speaker. The plots in Fig. 8, derived from the speech of four speakers, are indicative of the general trend. Any other set of four speakers is equally suitable for illustration.

The distribution of η for a given loudness level of a speaker can be approximated by a Gaussian probability density function. A measure of distance between two distributions is the Kullback–Leibler (KL) divergence (Kullback, 1968). When both the distributions are described by univariate Gaussian probability density functions, the KL divergence is given by (Cover and Thomas, 1991)

$$d_{KL}(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left\{\frac{\sigma_{\mathcal{A}}^2}{\sigma_{\mathcal{B}}^2} + \frac{\sigma_{\mathcal{B}}^2}{\sigma_{\mathcal{A}}^2}\right\} - 1 + \frac{1}{2}\{\mu_{\mathcal{A}} - \mu_{\mathcal{B}}\}^2\left\{\frac{1}{\sigma_{\mathcal{A}}^2} + \frac{1}{\sigma_{\mathcal{B}}^2}\right\}, \tag{8}$$

where $\mu_{\mathcal{A}}$ and $\sigma_{\mathcal{A}}$ denote the mean and the standard deviation, respectively, of the samples in set $\mathcal{A}$, while $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ denote the corresponding quantities for the samples in set $\mathcal{B}$. Also computed is $|\mu_{\mathcal{A}}-\mu_{\mathcal{B}}|$, which is the absolute value of the difference of the mean values $\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$. In this study, the samples in sets $\mathcal{A}$ and $\mathcal{B}$ are the values of the strength (η) of excitation. Let us consider the following two cases: (a) When the values of η in both $\mathcal{A}$ and $\mathcal{B}$ are derived from the soft utterances of a speaker, $d_{KL}(\mathcal{A},\mathcal{B})$ and $|\mu_{\mathcal{A}}-\mu_{\mathcal{B}}|$ are small. (b) A similar behavior is expected when the values of η in both $\mathcal{A}$ and $\mathcal{B}$ are derived from the loud utterances of the speaker. If we denote soft and loud as two classes of loudness, then the above two cases represent intra-class comparisons. By contrast, inter-class comparisons are those where the values of η in $\mathcal{A}$ and $\mathcal{B}$ are derived from the soft (loud) and the loud (soft) utterances, respectively, of a speaker. Both $d_{KL}(\mathcal{A},\mathcal{B})$ and $|\mu_{\mathcal{A}}-\mu_{\mathcal{B}}|$ are expected to be larger in the case of inter-class comparisons than in the case of intra-class comparisons. The ordered pair $[|\mu_{\mathcal{A}}-\mu_{\mathcal{B}}|, d_{KL}(\mathcal{A},\mathcal{B})]$ is used to distinguish between soft and loud utterances of a speaker, as described below.

Let $\mathcal{S}$ denote the set of values of η of a given speaker, derived from the 20 utterances collected in soft voice. Let $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ denote three distinct subsets of $\mathcal{S}$, such that the values of η in each subset are derived from six utterances in the soft voice. For the same speaker, let $\mathcal{L}$, $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ denote the corresponding sets derived from the loud utterances. For each speaker, the following ordered pairs are computed: (a) $(|\mu_{\mathcal{S}i}-\mu_{\mathcal{L}j}|, d_{KL}(\mathcal{S}_i,\mathcal{L}_j))$, for $i=1,2,3$, and $j=1,2,3$; (b) $(|\mu_{\mathcal{S}}-\mu_{\mathcal{L}}|, d_{KL}(\mathcal{S},\mathcal{L}))$; (c) $(|\mu_{\mathcal{S}i}-\mu_{\mathcal{S}j}|, d_{KL}(\mathcal{S}_i,\mathcal{S}_j))$ for $i=1,2,3$, $j=1,2,3$, and $i\neq j$; and (d) $(|\mu_{\mathcal{L}i}-\mu_{\mathcal{L}j}|, d_{KL}(\mathcal{L}_i,\mathcal{L}_j))$, for $i=1,2,3$, $j=1,2,3$, and $i\neq j$. It was observed that $\mu_{\mathcal{L}}>\mu_{\mathcal{S}}$, and $\mu_{\mathcal{L}i}>\mu_{\mathcal{S}j}$, for $i=1,2,3$, and $j=1,2,3$, for all the speakers. The ordered pairs in (a) and (b) denote inter-class comparisons within a speaker, while those in (c) and (d) denote intra-class comparisons. Each ordered pair can be plotted as a point in a two-dimensional plane. For each speaker, there are ten points due to inter-class comparisons and six points due to intra-class comparisons [since $d_{KL}(\mathcal{A},\mathcal{B})=d_{KL}(\mathcal{B},\mathcal{A})$]. Figures 9(a) and 9(c) show the intra-class points for the recorded 13 female and 31 male speakers respectively, while Figs. 9(b) and 9(d) show the inter-class points. For both female and male speakers, the intra-class points are clustered closer to the origin compared to the inter-class points which are farther from the origin and have a greater spread. Thus, the distribution of strength of excitation does help in distinguish-

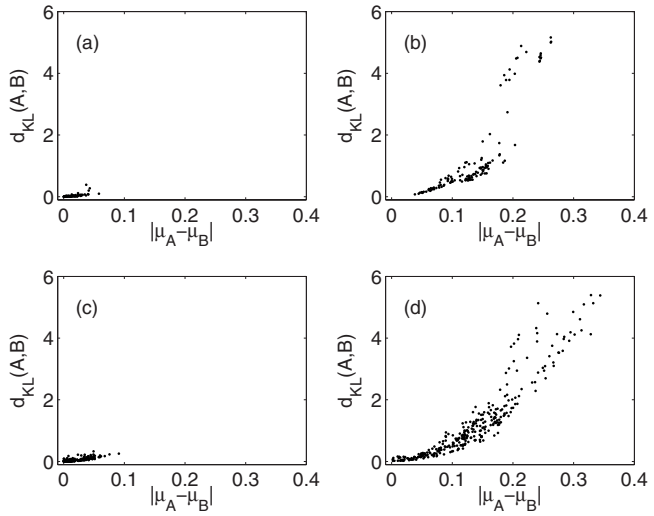

FIG. 9. Illustration of the variation of speaker-specific loudness. (a) and (b) show the results of intra-class comparisons and inter-class comparisons, respectively, for 13 female speakers. (c) and (d) show the results of intra-class comparisons and inter-class comparisons, respectively, for 31 male speakers.

ing between the loudness levels of a speaker. The distribution of η may also be useful in identifying those segments of speech signal of a speaker which are relatively soft. Such segments can then be processed, if necessary, to improve the loudness of the signal.

B. Comparison of loudness across speakers

Given two speakers, human listeners can judge in most cases, if the speech of one speaker is louder relative to the speech of the other speaker. This is particularly so if both the speakers belong to the same gender. This section describes an experiment to compare the subjective judgment of loudness and an objective measure of loudness.

The effect of prosodic factors such as changes in duration and pitch, on the perception of loudness in spontaneous speech, can be significant. For instance, speakers tend to raise their pitch while producing loud speech. Hence, when the utterances used for subjective listening consist of read speech corresponding to the same sentence, the prosodic variations across different speakers may be comparable. The subjective judgment of loudness is likely to be influenced more by the excitation characteristics of the voices than by the prosodic factors.

1. Speech material and subjective listening

Twenty-five subjects participated in the listening test to judge the relative loudness of several pairs of utterances. These 25 subjects belonged to the set of 44 subjects who participated in the experiment described in Sec. V A 1. For listening, speech signals corresponding to the utterances of the same text were selected from 20 speakers (6 female and 14 male speakers) of TIMIT database (Garofalo *et al.*, 1993). The average duration of the utterances is about 3 s. The signals were sampled at 16 kHz. The data were organized into pairs of utterances where each pair belonged to one of the following three types: (a) *XY*, when the pair consists of two different speakers; (b) *YX*, when the order of the speakers in the pair is reversed; and (c) *XX*, when the pair is a repetition of an utterance of the same speaker. For listening tests, 40 pairs of utterances were used, with 30 pairs of type *XY*, 5 pairs of type *YX*, and 5 pairs of type *XX*. The pairs of type *YX* and *XX* were used to check the consistency in the judgment of the subjects. Of the 30 pairs of type *XY*, 10 pairs consist of female speakers, and 20 pairs consist of male speakers. The utterances in each pair were normalized so that the energy of the signal was same in both the utterances.

The 40 pairs of utterances were presented in a random order. The subjects did not know the identity of the speakers. The subjects were asked to mark *A* or *B*, depending on whether the first or the second utterance in the pair was judged louder. They were asked to mark *C* if they observed no perceptible difference in loudness between the two utterances in a pair. For all the five pairs of type *XX*, all the subjects marked *C*. If a subject's decision on louder voice in the pairs of type *XY* and *YX* was not consistent, then the subject was regarded as inconsistent. If a subject was found inconsistent in two or more of the five such cases, then the decisions made by that subject were ignored. Out of the 25 subjects, four subjects were found to be inconsistent. Hence the decisions by the remaining 21 subjects were considered for evaluation. Out of these 21 subjects, the decision by the majority of the subjects on a pair of utterances of type *XY* was taken as the correct one. Here, the term majority denotes that at least 11 subjects out of the 21 subjects have voted in favor of one particular speaker, in a pair of utterances of type *XY*.

The subjective tests gave a clear decision on louder voice consistently, only for 21 out of the 30 pairs of the type *XY*. For the remaining nine pairs, there was no clear decision on the louder voice. This observation correlated with the objective measure described in Sec. V B 2.

2. Objective measurement of loudness

The loudness of two speakers in a pair is compared using the distributions of the strength (η) of excitation. For each speaker, ten utterances (including the utterance used in the listening test) were used to derive the distribution of η. The durations of the utterances varied from 2 to 4 s. All speech signals were downsampled to 8 kHz for processing. Let $\mathcal{A}$ and $\mathcal{B}$ denote the sets of values of η for a pair of speakers. The ordered pair $(|\mu_{\mathcal{A}}-\mu_{\mathcal{B}}|, d_{KL}(\mathcal{A},\mathcal{B}))$ is computed, as described in Sec. V A 2. There are 30 such ordered pairs corresponding to the 30 pairs of utterances of type *XY*, and these ordered pairs are plotted as points in a two-dimensional plane, as shown in Fig. 10. In Fig. 10(a), the points marked by "○" correspond to the nine pairs of speakers, for whom a clear decision could not be made by the listeners in the subjective test. Eight of these nine points lie close to the origin, indicating lack of discrimination between the distributions of η for the speakers in these pairs. Note that the subjective tests were conducted using only one utterance per speaker, whereas the distribution of η is obtained using ten utterances per speaker. The points marked by "+" in Fig. 10(a) denote the 21 pairs of speakers where one speaker was rated as louder in the subjective test. Here, the

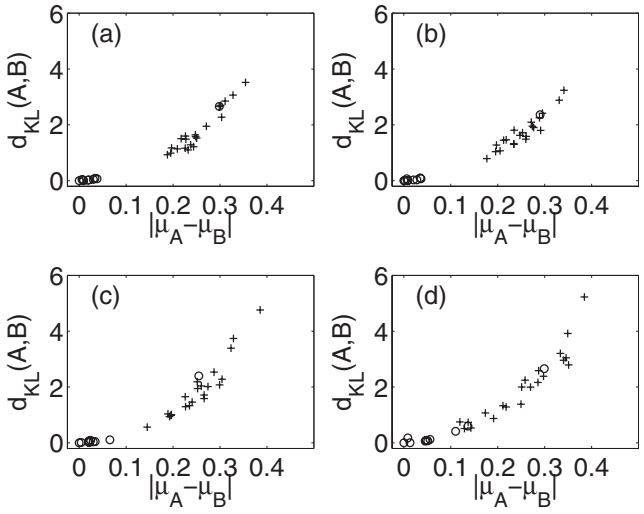

FIG. 10. In each plot, the points marked by + denote those pairs for which one speaker in each pair was judged as louder, based on the subjective listening. The points marked by o denote those pairs for which neither speaker in a pair was decisively voted as louder. For each speaker, the distribution of strength of excitation was obtained by processing (a) ten utterances, (b) six utterances, (c) three utterances, and (d) one utterance.

points are located farther away from the origin compared to the points marked by ○. Thus, discrimination between the loudness of two speakers based on the distributions of the strength of excitation is in agreement with the subjective judgment of loudness.

Figures 10(a)–10(d) correspond to the cases where the distribution of η for each speaker is obtained using ten, six, three, and one utterances, respectively. It is evident that the discrimination between the clusters of + and ○ points is reduced when the amount of speech data is reduced. Thus, 10–15 s of speech material may be required per speaker to obtain a reliable estimate of the distribution of the strength of excitation. The reliability of subjective judgment of loudness may also improve with increase in the duration of speech material used for listening.

VI. CONCLUSION

This paper presents a measure of perceived loudness in the form of distribution of a feature called strength of excitation. The strength of excitation represents the impulse-like nature of excitation in speech production. Observation of the electroglottograph signals indicates that the abruptness of glottal closure during the production of voiced speech plays an important role in the perception of loudness. The abruptness of the glottal closure lends the impulse-like characteristic to the excitation. Two features of the impulse-like excitation are investigated, namely, the amplitude and the strength of excitation. These features are derived from the Hilbert envelope of the LP residual of speech signal. The method proposed for estimating the amplitude of excitation is based on filtering the Hilbert envelope of LP residual through a zero frequency resonator. The strength of excitation is derived from a short segment of the Hilbert envelope of the LP residual of speech signal, around the instant of impulse-like excitation. The feature of the strength of excitation is that it is independent of the period of glottal vibration, and does not require parametrization of the glottal flow derivative. Experiments show that the distribution of the strength of excitation is strongly related to perceived loudness. The ability of the distribution of the strength of excitation to distinguish between soft and loud utterances of individual speakers is demonstrated using speech signals collected from a set of 44 speakers. Also, discrimination between the loudness of two speakers obtained based on the subjective judgment is in agreement with the discrimination between the distributions of the strength of excitation of the two speakers. This is illustrated on a set of 30 pairs of utterances, spoken by 20 speakers. Thus the distribution of the strength of excitation is useful for comparison of loudness of speakers. The significance of the amount of speech material required for reliable estimation of the distribution is also discussed. Since loudness varies over different segments of speech signal, it is more appropriately described by the *distribution* of the strength of excitation, than by the *strength* itself.

The proposed feature highlights the significance of the nature of excitation in the perception of loudness. The feature of the strength of excitation can help in measuring the loudness level of a speaker's voice on a quantitative basis. The proposed feature can be used to automatically identify the segments or regions of speech signal with relatively less loudness. Such segments may then be processed to enhance their loudness, if necessary. The reliability of the distribution of the strength of excitation improves with the amount of speech data. This is mainly due to variation in loudness over different segments of speech signal. Perceived loudness could be different for different sounds, due to loading of the vocal tract system on the vocal source during the production of speech sounds. The perception of loudness in human listeners is also influenced by prosodic factors, such as variations in pitch and duration, which are manifested over longer durations of speech. Hence, the assessment of loudness by human listeners is likely to improve with the duration of speech data. Prosodic factors, such as stress on particular syllables of words and accentuation of stressed syllables, also affect the perception of loudness. The influence of such factors on the perception of loudness needs to be studied and quantified, in order to obtain a more comprehensive measure of loudness of speech.

Alku, P., Airas, M., Björkner, E., and Sundberg, J. (**2006**). "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," J. Acoust. Soc. Am. **120**, 1052–1062.
Alku, P., Vintturi, J., and Vilkman, E. (**2002**). "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," Speech Commun. **38**, 321–334.
Allen, G. D. (**1971**). "Acoustic level and vocal effort as cues for the loudness of speech," J. Acoust. Soc. Am. **49**, 1831–1841.
Ananthapadmanabha, T. V., and Yegnanarayana, B. (**1979**). "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust., Speech, Signal Process. **27**, 309–319.
Bäckström, T., Alku, P., and Vilkman, E. (**2002**). "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range," IEEE Trans. Speech Audio Process. **10**, 186–192.
Beranek, L. L., Marshall, J. L., Cudworth, A. L., and Peterson, A. P. G. (**1951**). "Calculation and measurement of the loudness of sounds," J. Acoust. Soc. Am. **23**, 261–269.
Cairns, D. A., and Hansen, J. H. L. (**1994**). "Nonlinear analysis and classification of speech under stressed conditions," J. Acoust. Soc. Am. **96**, 3392–3400.

Childers, D. G., and Lee, C. K. (**1991**). "Voice quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**, 2394–2410.
Cover, T. M., and Thomas, J. A. (**1991**). *Elements of Information Theory* (Wiley, New York).
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.
d'Alessandro, C., and Scherer, K. R. (**2003**). "Voice quality: Functions, analysis and synthesis (VOQUAL'03)," ISCA Tutorial and Research Workshop, Geneva, Switzerland, http://archives.limsi.fr/VOQUAL/voicematerial.html (Last viewed 6/7/2008).
Doval, B., d'Alessadro, C., and Henrich, N. (**2006**). "The spectrum of glottal flow models," Acta. Acust. Acust. **92**, 1026–1046.
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.
Eskenazi, L., Childers, D. G., and Hicks, D. M. (**1990**). "Acoustic correlates of vocal quality," J. Speech Hear. Res. **33**, 298–306.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).
Fletcher, H., and Munson, W. A. (**1933**). "Loudness, its definition, measurement and calculation," J. Acoust. Soc. Am. **5**, 82–108.
Fletcher, H., and Munson, W. A. (**1937**). "Relation between loudness and masking," J. Acoust. Soc. Am. **9**, 1–10.
Garofalo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom," Linguistic Data Consortium, Philadelphia, PA.
Gauffin, J., and Sundberg, J. (**1989**). "Spectral correlates of glottal voice source waveform characteristics," J. Speech Hear. Res. **32**, 556–565.
Glave, R. D., and Rietveld, A. C. M. (**1975**). "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?," J. Acoust. Soc. Am. **58**, 875–879.
Gramming, P., and Sundberg, J. (**1988**). "Spectrum factors relevant to phonetogram measurement," J. Acoust. Soc. Am. **83**, 2352–2360.
Harris, C. M., and Weiss, M. R. (**1964**). "Effects of speaking condition on pitch," J. Acoust. Soc. Am. **36**, 933–936.
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.
Johnstone, T., and Scherer, K. R. (**1999**). "The effects of emotions on voice quality," in Proceedings of the 14th International Conference on Phonetic Sciences, San Francisco, pp. 2029–2032.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Kullback, S. (**1968**). *Information Theory and Statistics* (Dover, Mineola, NY).
Ladd, D. R., Verhoeven, J., and Jacobs, K. (**1994**). "Influence of adjacent pitch accents on each other's perceived prominence: Two contradictory effects," J. Phonetics **22**, 87–99.
Ladefoged, P., and McKinney, N. P. (**1963**). "Loudness, sound pressure, and subglottal pressure in speech," J. Acoust. Soc. Am. **35**, 454–460.
Lane, H. L., Catania, A. C., and Stevens, S. S. (**1961**). "Voice level: Autophonic scale, perceived loudness, and effects of sidetone," J. Acoust. Soc. Am. **33**, 160–167.
Laver, J. (**1994**). *Principles of Phonetics*, Cambridge Textbooks in Linguistics (Cambridge University Press, Cambridge).
Lieberman, P., Knudson, R., and Mead, J. (**1969**). "Determination of the rate of change of fundamental frequency with respect to subglottal air pressure during sustained phonation," J. Acoust. Soc. Am. **45**, 1537–1543.
Liénard, J.-S., and Benedetto, M.-G. D. (**1999**). "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422.
Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.
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.
Murty, K. S. R., and Yegnanarayana, B. (**2008**). "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **16**, 1602–1613.
Oppenheim, A. V., and Schafer, R. W. (**1975**). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
Orlikoff, R. F. (**1991**). "Assessment of the dynamics of vocal fold contact from the electroglottogram: Data from normal male subjects," J. Speech Hear. Res. **34**, 1066–1072.
Pickett, J. M. (**1956**). "Effects of vocal force on the intelligibility of speech sounds," J. Acoust. Soc. Am. **28**, 902–905.
Rothenberg, M. (**1983**). "The effect of flow dependence on source-tract acoustic interaction," J. Acoust. Soc. Am. **73**, S72–S72.
Schulman, R. (**1989**). "Articulatory dynamics of loud and normal speech," J. Acoust. Soc. Am. **85**, 295–312.
Smits, R., and Yegnanarayana, B. (**1995**). "Determination of instants of significant excitation in speech using group delay functions," IEEE Trans. Speech Audio Process. **3**, 325–333.
Stevens, S. S. (**1956**). "Calculation of the loudness of complex noise," J. Acoust. Soc. Am. **28**, 807–832.
Stevens, S. S. (**1961**). "Procedure for calculating loudness: Mark VI," J. Acoust. Soc. Am. **33**, 1577–1585.
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., and Nordenberg, M. (**2006**). "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," J. Acoust. Soc. Am. **120**, 453–457.
Sundberg, J., Fahlstedt, E., and Morell, A. (**2005**). "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," J. Acoust. Soc. Am. **117**, 879–885.
Ternström, S., Bohman, M., and Södersten, M. (**2006**). "Loud speech over noise: Some spectral attributes, with gender differences," J. Acoust. Soc. Am. **119**, 1648–1665.
Traunmüller, H., and Eriksson, A. (**2000**). "Acoustic effects of variation in vocal effort by men, women, and children," J. Acoust. Soc. Am. **107**, 3438–3451.
Warren, R. M. (**1973**). "Anomalous loudness function for speech," J. Acoust. Soc. Am. **54**, 390–396.
Zwicker, E. (**1977**). "Procedure for calculating loudness of temporally variable sounds," J. Acoust. Soc. Am. **62**, 675–682.
Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics: Facts and Models*, Springer Series in Information Sciences Vol. **22**, 2nd ed. (Springer, Berlin).