Journal of Phonetics (2001) **29,** 451–480
doi:10.006/jpho.2001.0146
Available online at http://www.idealibrary.com on IDEAL®

Towards models of phonation

Helen M. Hanson*

Sensimetrics Corporation, 48 Grove St., Suite 305, Somerville, MA 02144, U.S.A. and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Kenneth N. Stevens

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. and Sensimetrics Corporation, 48 Grove St., Suite 305, Somerville, MA 02144, U.S.A.

Hong-Kwang Jeff Kuo

Bell Laboratories, Murray Hill, NJ 07974, U.S.A.

Marilyn Y. Chen

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Janet Slifka

Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

The earliest models of phonation were based on the assumption that the glottis is closed during a part of the vibration cycle, that is, the phonation is modal. Nonmodal phonation, however, commonly occurs not only for disordered voice but also for normal voices, which often exhibit a breathy quality or irregular vibration. In this paper, we review recent work that examines acoustic data and models of nonmodal phonation in both normal and disordered voice. We first describe acoustic models that predict how the glottal source varies from modal phonation to phonation resulting from glottal configurations that are partially abducted, including a posterior glottal opening. These models are applied first to vowels of nondisordered adults, and, later in the paper, to vowels produced by adults with dysarthria. We also present results from a study in which a modified version of the two-mass model is used to resolve a seeming conflict among aerodynamic and acoustic data collected from adult female subjects with vocal-fold nodules. Some discussion of nonmodal phenomena that occur due to prosodic and emotional influences is included. Overall, it appears that current models of modal phonation can be extended to include a range of nonmodal phonation types.

*E-mail: hanson@sens.com

0095-4470/01/040451 + 30 $35.00/0

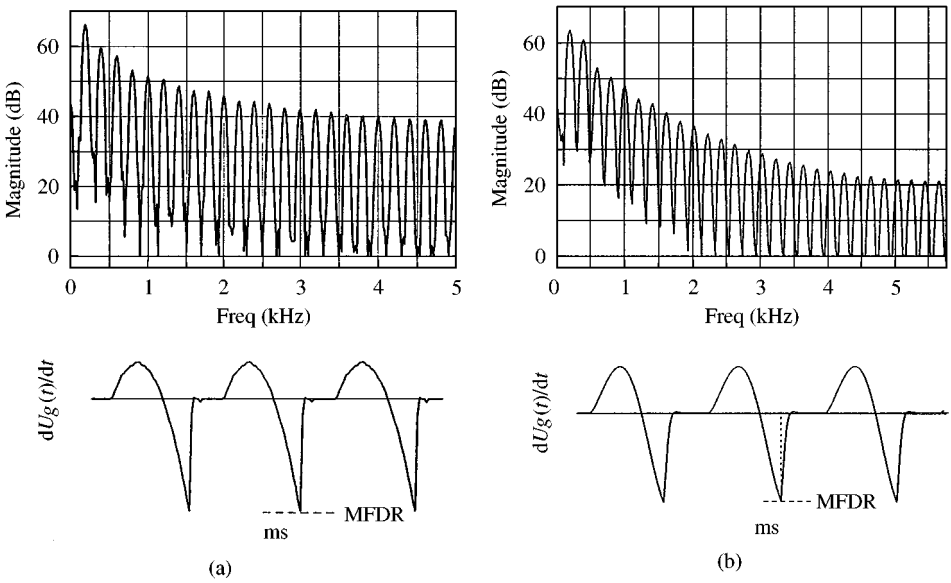

Figure 1. Examples of glottal-waveform derivatives and corresponding spectra, generated using the KLGLOTT88 model (Klatt & Klatt, 1990). (a) Derivative of a modal volume-velocity waveform. At high frequencies, the spectrum falls off at 6 dB/octave. (b) Derivative of a volume-velocity waveform with a nonzero return phase (the dotted line indicates a return phase of 0). At high frequencies, the spectrum falls off at 12 dB/octave. The dashed line indicates the maximum flow declination rate (MFDR).

1. Introduction

Models of phonation generally fall into two categories: models of the volume-velocity waveform (or its derivative) generated by the oscillating vocal folds, or models of the vocal-fold motion (or area function). Examples of volume-velocity waveform models include the Rosenberg (1971) model, the LF model (see, for example, Fant, Liljencrants & Lin, 1985), the KLGLOTT88 model (Klatt & Klatt, 1990), and the R++ model (Veldhuis, 1998). Models of vocal-fold motion include the two-mass model (Ishizaka & Matsudaira, 1968; Ishizaka & Flanagan, 1972) and the "body-cover" model (Story & Titze, 1995).

The earliest models only included modal phonation. Modal phonation has been defined as phonation in which full contact occurs between the vocal folds during the closed phase of a phonatory cycle (Titze, 1995). The volume-velocity waveform or area function produced by a modal model is zero during the closed phase, and its first derivative has a discontinuity at the moment that closure occurs. In Fig. 1(a), we show the derivative of a modal volume-velocity waveform produced using the KLGLOTT88 model, and its power spectrum. At mid and high frequencies, the spectrum falls off at 6 dB/octave, as expected when there is a discontinuity in the derivative of the waveform (see, for example, Siebert, 1986).

For some time now, researchers have realized that models that include nonmodal phonation are necessary if further advances are to be made in several areas of speech research and applications. For example, natural-sounding synthesized speech can only be achieved by including nonmodal phenomena. Likewise, in the study of voice

disorders, modal phonation models may not be useful. Moreover, speech recognition by computers can also benefit from the ability to model nonmodal phonation.

In this paper, we discuss how nonmodal phonation has been included in several phonation models, and how some of these models have been applied to study nonmodal phonation in both normal and disordered speech. We begin by reviewing some of the phenomena that nonmodal models should account for. After a brief description of some nonmodal models proposed by other researchers, we review some of our own work relating the acoustic speech signal to certain glottal configurations. This review includes experimental data from several studies which demonstrate how voice quality varies across individuals, within individuals, and across populations. A speech synthesizer which allows one to incorporate such variations is also briefly described. Finally, we discuss the application of phonation models to disordered speech. In particular, we discuss the use of a two-mass model to study and interpret data on the phonation of nodules patients, and the application of the acoustic models presented earlier to the assessment of the phonation of individuals with neuromotor disorders.

2. Nonmodal phenomena

In this section, we discuss some characteristics that a model of phonation should include if it is to account for observations of normal and disordered phonation. While not every aspect of nonmodal phonation is included, we cover those that we consider most relevant.

2.1. *Breathy voice*

Breathy voice quality is one of the most commonly studied types of nonmodal phonation. In some languages it is used linguistically, as a contrastive feature that distinguishes between words (see, for example, Gordon & Ladefoged, 2001). However, breathiness as an overall voice quality can also vary from one individual to another. Many believe that women tend to have breathier voice quality than males, although it is not known if this difference is due to anatomy and physiology, or to social construct. Excessively breathy voice is usually a symptom of a voice disorder. When discussing breathy voice, care must be taken to differentiate between breathiness as a voice disorder, as a habitual voice quality, and as a linguistic feature. The physical source may be the same (see below), but the underlying causes are quite different.

Breathiness is the result of excessive air leakage at the glottis when the vocal folds do not fully approximate during phonation. For normal speakers, this leakage usually occurs when there is a glottal opening at the arytenoid cartilages which is maintained throughout a phonatory cycle. For disordered speakers, on the other hand, air may escape at other places along the length of the folds (Titze, 1995). This glottal opening can have two effects on the acoustic source at the glottis: it can modify the spectrum of the periodic component by reducing the amplitude at mid and high frequencies, and it can introduce a noise component to the spectrum at mid and high frequencies (Section 3.1). Fig. 2 shows schematically how the amplitude of the third-formant peak might vary as a function of average glottal opening, for both periodic and noise components of the glottal excitation. For relatively small average glottal openings, as in modal phonation, the noise component is much lower in amplitude than the periodic component, while for relatively large glottal openings, voicing does not occur and only noise is generated

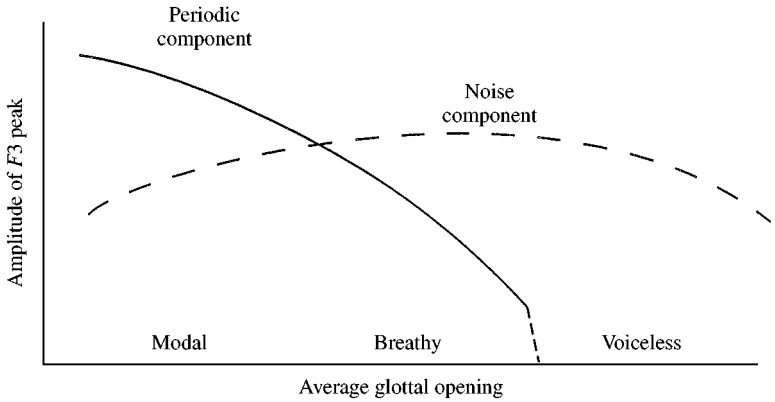

Figure 2. The amplitude of the F_3 peak as a function of the average glottal opening, for both periodic and noise components of the glottal source. For small glottal openings, the noise component is insignificant and phonation is modal. For intermediate glottal openings, the noise component may be greater than the periodic component and phonation is breathy. For large openings, the vocal folds will not vibrate and only the noise component contributes to the source.

(aspiration). For moderate glottal openings, voicing may continue, but the noise component in the F_3 region may be comparable to or stronger than the periodic component. It is this significant noise component in the mid to high-frequency range that is believed to create the impression of breathy voice quality (Klatt & Klatt, 1990).

2.2. *Failure to vibrate*

The vocal folds will only vibrate under certain conditions of transglottal pressure, configuration of the glottis, and state of the vocal folds. First, the folds must be adducted, without being too close together; if they are either too tightly pressed or too widely adducted, they will not vibrate. Second, the folds must have an appropriate amount of tension; too much tension will limit the range of conditions under which the vocal folds can vibrate. Finally, a minimum pressure drop must exist across the folds. The latter requirement is referred to as the phonation threshold pressure; it is the minimum transglottal pressure necessary to initiate or maintain vocal-fold vibration (Baer, 1975; Titze, 1988, 1992; Lucero, 1995). These three variables are not independent. For example, the phonation threshold pressure is higher if the folds are relatively tense or if they are somewhat adducted. Models of phonation that include nonmodal effects should include these three factors and their interdependence.

2.3. *Irregular vibration*

Although the folds will oscillate when the variables transglottal pressure, vocal-fold tension, and vocal-fold adduction are in particular ranges, the vibrations may become irregular for certain combinations of the variables in these ranges. As with breathy voice, humans have learned to sometimes take advantage of these regions of instability to enhance certain contrasts in language. Consequently, irregular vibration of the vocal folds can function on several levels in speech production. Some languages use it as a contrastive feature at the segmental level, while others may use it at a prosodic level or as a social marker (Redi & Shattuck-Hufnagel, 2001). In all languages, it could occur

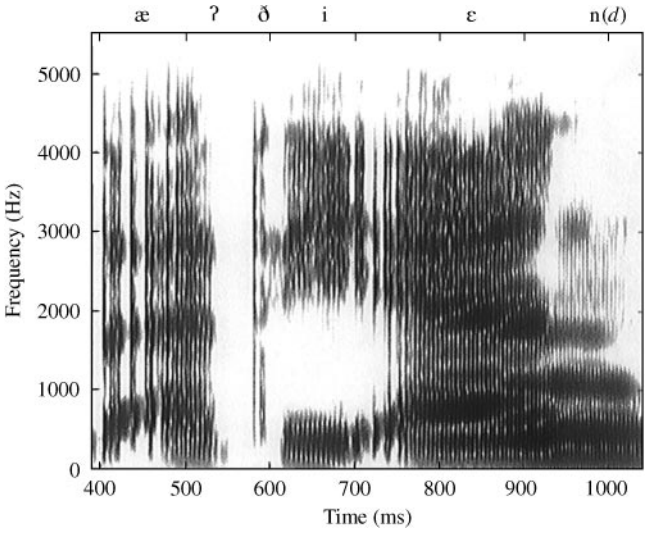

Figure 3. A spectrogram illustrating allophonic uses of glottalization in American English. The utterance is "at the end", excised from the rainbow passage; it occurred phrase initially. The word "at" is glottalized at vowel onset. Glottalization also occurs at the vowel-to-vowel junction in "the end". The /t/ at the end of 'at' is probably produced as a glottal stop. The speaker is a female with no known voice disorder.

simply as a byproduct of variations in the glottal configuration; that is, as the glottal configuration moves from one setting to another, it could move through regions of instability (Slifka, 2000). There may be individual differences as to whether it occurs and in its details (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Slifka, 2000). Irregular vibration may also be the symptom of a voice disorder.

Although the term *glottalization* is often used interchangeably with *irregular vibration*, glottalization is one among several types of irregular vibration. Titze (1995) has defined glottalization as "transient sounds resulting from the relatively forceful adduction or abduction" of the vocal folds. In American English, glottalization is not contrastive, but it is used allophonically; vowel-initial words may be glottalized at onset (e.g., "elephant") (Dilley & Shattuck-Hufnagel, 1995), and in syllable-final environments, voiceless stop consonants, particularly /t/, may be realized by a glottal stop (e.g., in "hat rack") (Pierrehumbert, 1995). Fig. 3 illustrates the glottalization of vowel-initial words at onset. The spectrogram is of the utterance "at the end", extracted from a recording of the rainbow passage. The speaker is a female, with no known voice disorder. The word "at" occurs phrase initially, and the /æ/ is seen to be glottalized. The transition from vowel /i/ to vowel /ɛ/ is also glottalized, marking the word-initial pitch-accented vowel. In addition, the /t/ at the end of "at" was probably realized by terminating the vowel with a glottal closure, although it is not evident in the spectrogram.

Irregular vibration may occur in phrase-final position. Although irregular vibration in this environment has been referred to as glottalization, recent evidence (Slifka, 2000, in prep.) suggests otherwise. Fig. 4 shows data extracted from the last two words of a recording of the phrase "Ali will have the best seat in the show". The top panel is the acoustic signal, in which it can be seen that irregular vibration occurs during the final

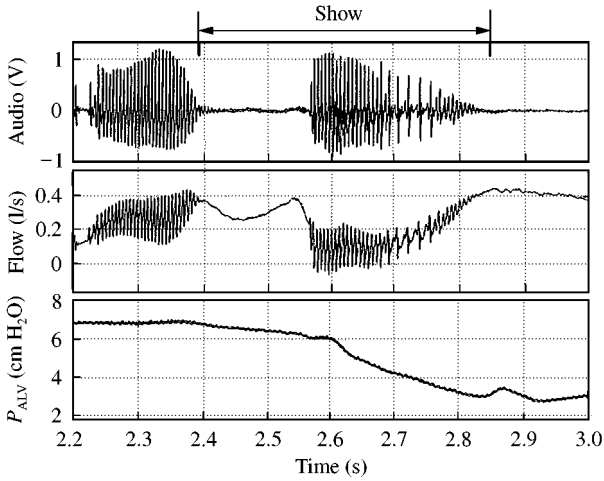

Figure 4. Voicing termination example for a female speaker. Panels show the audio signal (V), oral flow (l/s), and estimated alveolar pressure (cm H_2O). The utterance ends with the word "show" and has an irregular glottal waveform during the second half of the word. During the irregular phonation, subglottal pressure falls while oral flow increases, implying vocal-fold abduction. Therefore, the irregular vibration cannot be glottalization, but is rather due to an unstable combination of glottal area, subglottal pressure, and vocal-fold tension. (After Slifka, 2000.)

half of the word "show". The middle panel shows oral airflow measured using a Rothenberg mask and the bottom panel shows the alveolar pressure, estimated via an esophageal balloon. The alveolar pressure is falling during the period of irregular vibration; if the glottal area was being maintained or if the folds were being adducted during this time, the flow would be expected to decrease as well. However, it is evident that the flow is increasing smoothly as the pressure falls. Therefore, contrary to what one might expect during glottalization, it is likely that the vocal folds are being gradually abducted. Slifka (2000, in prep.) observed many examples of this pattern, and concluded that this particular instance of irregular vibration is not glottalization, but is the result of unstable vibration due to a combination of vocal-fold abduction, vocal-fold slackening, and reduced transglottal pressure.

Another environment in which irregular vibration may occur is the low pitch accent. Again, this type of irregular vibration has also been referred to as glottalization. As the above discussion suggests, this assumption should perhaps be reconsidered. It could be that in an attempt to lower pitch, speakers reduce subglottal pressure or slacken the vocal folds, leading to a combination of the vocal-fold parameters that results in unstable vibration.

In summary, irregular vibration should be included in models of phonation, and these models must take into account the observation that different constellations of the parameters describing glottal configuration or state can result in irregular vibration.

2.4. *Disordered voice*

Certain organic or neurological disorders can have a negative impact on the production of voice, resulting in nonmodal phonation. The degree to which communication is

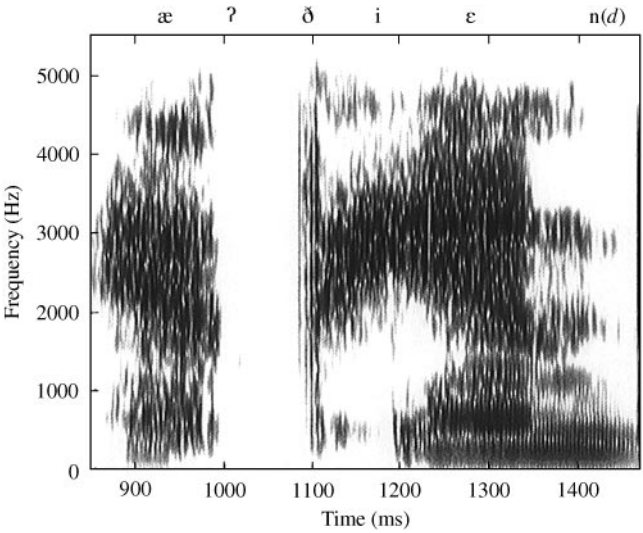

Figure 5. Spectrogram of the phrase "at the end" extracted from a recording of the rainbow passage produced by a female with dysphonia. Formants above about 1000 Hz are primarily excited by aspiration noise. Vocal-fold vibration ceases during the word "at" and for most of the word "the".

impaired ranges from little to severe. Models of phonation for these disorders certainly should include nonmodal effects. Fig. 5 is a spectrogram of the phrase "at the end" extracted from a recording of the Rainbow Passage, produced by a female with severe dysphonia. Formants above about 1000 Hz are seen to be primarily excited by noise. Vocal-fold vibration seems to stop completely halfway through the word "at" and for most of the word "the". A phonation model that produces breathy phonation and that would fail to result in vibration under the appropriate circumstances would clearly be useful in the study of such a voice.

3. Models, acoustic measurements, and synthesis

3.1. *Models*

One of the first phonation models intended to include a nonmodal effect was the LF model (Fant *et al*., 1985). The LF model is a four-parameter model of the derivative of the volume-velocity waveform. Where it differs from previous models is in the transition from the open phase to the closed phase: rather than always being abrupt, the LF model includes the capability of having a nonabrupt transition. In the derivative of the glottal waveform, this nonabrupt transition translates to the lack of a discontinuity at the moment of closure, thereby reducing the high-frequency content of the periodic source. The degree to which it is attenuated depends on the "return phase", or how quickly the volume velocity is cut off at closure. Fig. 1 compares the waveform and spectrum for a volume-velocity waveform with an abrupt return phase (part a) and with a nonzero return phase (part b). At high frequencies, the spectrum in (b) falls off at a faster rate than 6 dB/octave. The reduction in high-frequency content is sometimes referred to as an

increase in spectral tilt, or the balance between the high- and low-frequency energy in the signal.

Another model of the volume-velocity waveform, KLGLOTT88, was proposed by Klatt & Klatt (1990). This model includes a tilt, or return-phase, parameter, which allows one to control the high-frequency content of the periodic source. In addition, aspiration noise can be added to the periodic source. Another parameter allows one to include a diplophonia-like effect, in which every other glottal waveform is delayed and attenuated by a certain percentage.

More recently, Wilhelms-Tricarico (unpublished) has extended the two-mass model to include a posterior glottal opening and nonsimultaneous closure of the folds. Cranen & Schroeter (1995) have modeled leaks occurring at the posterior end of the horizontal folds and along the length of the membranous folds. Story & Titze (1995) have developed the "body-cover" model, an extension of the two-mass model which allows separate vibration of masses representing the vocal-fold body and cover. This model is intended to provide a more realistic vehicle for research into both normal and disordered vocal-fold oscillations.

In the remainder of this section, we describe two studies in which a model was developed to interpret experimental data on nonmodal phonation by normal speakers (Hanson, 1997; Hanson & Chuang, 1999). Although the two studies were done separately, we will discuss them as though they formed a single study. Rather than developing a model of the volume-velocity waveform or the glottal area function, we developed models of how the glottal waveform and the vocal-tract transfer function are affected for certain glottal configurations believed to be common in normal voice production. The configurations of interest are (1) modal (full contact occurs between the vocal folds during the closed phase of a phonatory cycle (Titze, 1995)); (2) the membranous portions of the folds close completely and simultaneously along their length, while a posterior glottal opening (PGO) is maintained at the arytenoid cartilages during the closed phase of the glottal cycle; (3) the closure is nonsimultaneous, beginning at the anterior ends of the membranous folds and progressing towards the vocal processes, and there is no PGO; (4) there is a PGO and the membranous folds close nonsimultaneously. We were primarily interested in how the spectrum of the voice source would be affected by these various configurations, and how those spectral characteristics of the source could be estimated from the acoustic speech signal.

When vocal-fold vibration is modal, glottal characteristics can still show some variation. The open quotient, or ratio of the time that the folds are open to the duration of a complete cycle of vibration, can vary. In addition, there can be variations in the speed quotient, or ratio of the duration of the opening gesture to the duration of the closing gesture. If a speaker has a posterior glottal opening, significant changes can occur. In particular, our models show that the source spectral tilt, the aspiration noise generated near the folds, and the bandwidth of the first formant will all be increased. The degree by which they increase depends on the size of the PGO, with larger openings resulting in greater increases. Therefore, posterior glottal openings are a source of great variability in the characteristics of the volume-velocity waveform.

When the vocal folds close completely but nonsimultaneously along their length, the source spectral tilt will increase because the cutoff of flow is gradual rather than abrupt, resulting in a loss of high-frequency energy. The amount by which the tilt increases depends on the speed at which the folds close from the anterior to posterior ends. Finally, if there is a PGO and the folds close nonsimultaneously, spectral tilt, aspiration noise,

and first-formant bandwidth increase, but the spectral tilt increases even more than it would have if there was only a PGO. The additional amount by which tilt increases is difficult to predict because the time constant of the nonsimultaneous closure is independent of the size of the posterior glottal opening.

Based on the model, we propose a group of measures that reflect the characteristics predicted to be affected by the glottal configurations of interest: spectral tilt, open quotient, first-formant bandwidth, and aspiration noise. It is expected that as open quotient (OQ) increases, the glottal waveform more closely approximates a sinusoid of frequency F_0, and therefore in the frequency domain the amplitude of the first harmonic increases relative to the amplitudes of the higher harmonics. The measure $H1$–$H2$ (relative amplitude of the first two harmonics, in dB), has been used by other researchers to reflect OQ and has been shown to be correlated with OQ by Holmberg, Hillman, Perkell, Guiod & Goldman (1995). The amplitudes of $H1$ and $H2$ for a typical vowel spectrum are shown in Fig. 6, together with several other relevant measures. If the first-formant bandwidth ($B1$) increases, the amplitude $A1$ of the first-formant peak in the spectrum is expected to decrease. Therefore, the relative amplitude of the first harmonic and the first-formant peak ($H1$–$A1$, in dB) is selected as an indicator of $B1$. It may also be affected if spectral tilt extends down to the first-formant frequency. A more direct estimate of $B1$ is also proposed: this measurement involves first bandpass filtering the speech waveform at the F_1 frequency, and then measuring the decay of the F_1 oscillations, as illustrated by the dashed line at the bottom of Fig. 6. An increase in spectral tilt is most evident at higher frequencies, and therefore the measure $H1$–$A3$ (the relative amplitude of the first harmonic and the third-formant peak, in dB) is selected as a reflection of tilt.

Because we make the spectral measures on the speech spectrum, the measures $H1$ and $H2$ are corrected for the boosting effects of the first formant and the measure $A3$ is corrected for the boosting effects of the first and second formants. These corrections are based on the acoustic theory of speech production (Fant, 1960). They are necessary to allow comparison of the measures across vowels and speakers, and are denoted with an asterisk. For example, $H1^*$–$H2^*$ denotes the measure $H1$–$H2$ is corrected for the effects of F_1. The equations for these corrections are given in an earlier paper (Hanson, 1997).

Finally, aspiration noise is measured by using an extended version of a subjective rating system proposed by Klatt & Klatt (1990). In this system, noise in the third-formant region is observed by bandpass filtering the speech waveform at the F_3 frequency, and then rating the periodicity of the resulting signal on a scale of 1–4, where 1 corresponds to "no evidence of noise excitation" and 4 corresponds to "little evidence of periodic excitation". This system is also extended by observing for the vowel spectrum in the frequency region of the third formant, using a time window of 25–35 ms, depending on the gender of the subject. The noise rating based on the waveform is referred to as N_W (illustrated in Fig. 6), while the rating based on the spectrum is referred to as N_S.

Next, experimental data (House & Stevens, 1958; Fant, 1972; Holmberg, Hillman & Perkell, 1988; Holmberg, Hillman, Perkell & Gress, 1994; Perkell, Hillman & Holmberg, 1994) and analysis of the KLGLOTT88 model of the glottal waveform (Klatt & Klatt, 1990) are combined with our models to predict mean, minimum, and maximum expected values for the measures $H1^*$–$A1$, $H1^*$–$A3^*$, and $B1$ for male and female adult speakers without voice disorders. The maximum values for $H1^*$–$A3^*$ can be estimated only for the configuration in which a PGO exists and closure is simultaneous along the length of the folds. The reason for the latter is that experimental data on nonsimultaneous closure are not available. Therefore, the estimated maximum for $H1^*$–$A3^*$ is

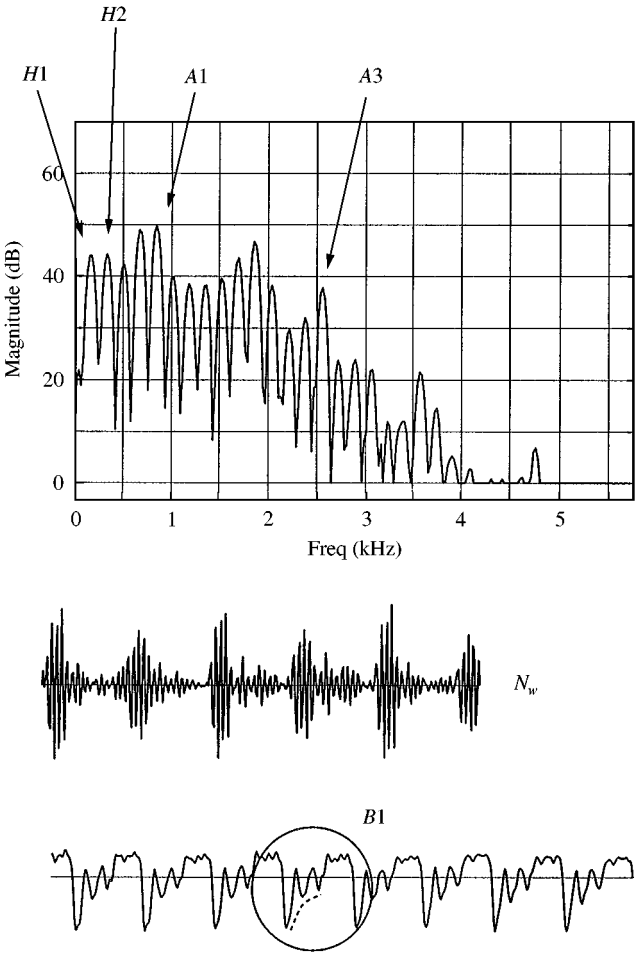

Figure 6. Speech waveforms and a vowel spectrum produced by female speakers. The acoustic parameters labeled in the spectrum are the amplitudes of the first harmonic ($H1$), second harmonic ($H2$), first formant ($A1$), and third formant ($A3$). The top waveform was obtained by bandpass filtering the sound-pressure waveform at the F_3 frequency, and is an example of those used to make the noise rating N_w described in the text. The bottom waveform illustrates the decay of the first-formant oscillation, used to calculate the estimated bandwidth $B1$. (Note that each of these examples is from a different speaker.) (From Hanson & Chuang, 1999.)

a "soft" upper threshold in that $H1^*–A3^*$ is expected to range at least that high, but possibly higher when a PGO is combined with nonsimultaneous closure.

With these estimates in hand, several predictions can then be made about how these measures might vary among adult speakers. First, a wide variation for all of the measures is expected within both gender groups. The females are expected to show slightly greater variation, because they are believed to be more likely to have posterior glottal openings, which can vary in size. Male speakers are expected to have lower mean values than females on all measures. It is also predicted that most of the measures should be positively correlated. The measures $H1^*–A1$, $H1^*–A3^*$, $B1$, N_W, and N_S should be

correlated because they all should increase if the size of a PGO increases. Because breathy voice quality has been found to be correlated with both PGO size and $H1^*–H2^*$, we also expect $H1^*–H2^*$ to be correlated with the other measures.

3.2. *Experimental data for citation words*

Data from 43 adults (21 male, 22 female) were collected. Five tokens each of the vowels /æ, ʌ, ɛ/ in a carrier phrase were recorded and analyzed. An average value for each acoustic measure was obtained for each vowel for each speaker. As expected, fairly wide variations among both male and female speakers were observed, and the females had a tendency to show slightly greater variations for each measure. The estimates in Section 3.1 for the mean, minimum, and maximum values were good predictions of what was observed.

As an example, Fig. 7 is a graph of the measure $H1^*–A3^*$ *vs.* $H1^*–A1$ for the female speakers. Each point on this graph represents one vowel produced by a speaker; thus, there are 66 points (3 vowels × 22 female speakers). Overlaid on this graph are lines representing the minimum and maximum expected values for the two measures. For the measure $H1^*–A1$, nearly all of the data fall between these lines. For the measure $H1^*–A3^*$, all but one of the points fall above the line representing the expected minimum. Nearly half of the points fall above the expected maximum. Recall, however, that this maximum is based on the models in which there is simultaneous closure along

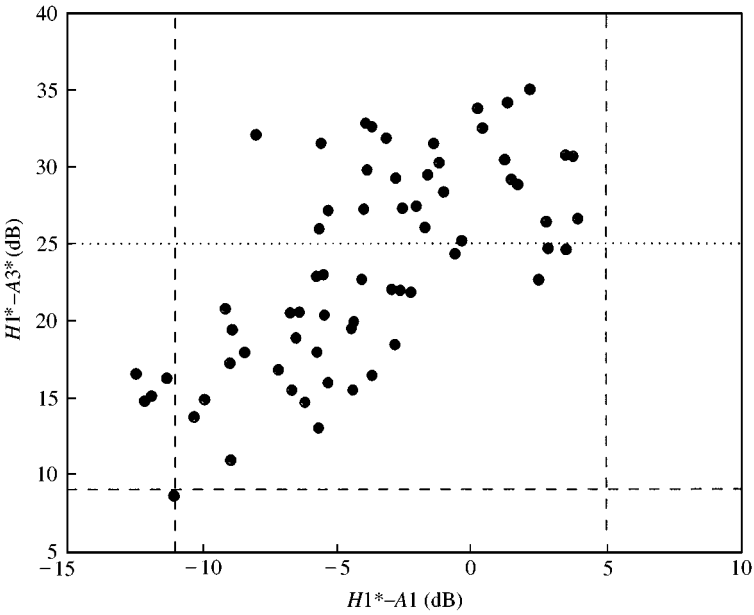

Figure 7. Relation between the measures $H1^*–A3^*$ and $H1^*–A1$ for 22 female speakers. Each point represents data for one of the three vowels per speaker. The vertical dashed lines indicate minimum and maximum values expected for $H1^*–A1$. Likewise, the horizontal dashed line indicates the minimum value expected for $H1^*–A3^*$. The horizontal dotted line indicates the maximum value expected for $H1^*–A3^*$ for glottal configurations in which vocal-fold closure is simultaneous along the length of the membranous folds. See text for details.

the length of the folds. Therefore, the points that fall above the maximum line do not cast doubt on the models, but rather suggest that nonsimultaneous closure of the vocal folds is fairly prevalent among this group of female speakers. Because the points falling above the line also have relatively large values of $H1^*–A1$, it is also assumed that these vowels were produced with a glottal configuration that included a posterior glottal opening.

The data points are divided into two groups, those with $H1^*–A3^*$ less than or equal to 23 dB and those with $H1^*–A3^*$ greater than 23 dB (see Hanson (1997) for more details). Closer analysis of the data reveals that only three of the speakers have data points falling into both groups. Therefore, the female speakers are classified as belonging to one of two groups: Group 1 speakers (with $H1^*–A3^*$ below 23 dB) are hypothesized to have glottal configurations for which vocal-fold closure is simultaneous along the length of their vocal folds, while Group 2 speakers are hypothesized to have nonsimultaneous closure. Because of the range of $H1^*–A1$ observed for both groups, it is also hypothesized that

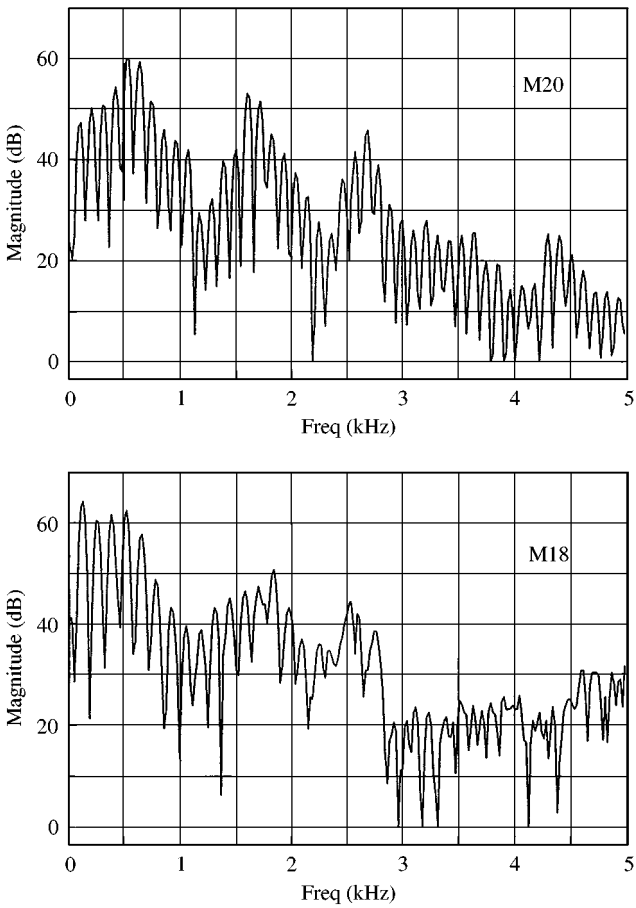

Figure 8. Typical vowel spectra for two male subjects. The vowel is /ɛ/. Subject M20 falls at the low end of the range for most measures, while Subject M18 falls at the high end. The spectrum for M20 has a relatively low value of $H1^*–H2^*$ and the first three formant peaks are more sharply defined than those in the spectrum for M18. In addition, the spectrum for M18 is noisy above 2000 Hz, while that for M20 shows little evidence of noise. (From Hanson & Chuang, 1999.)

most of the speakers in Group 1 and all of the speakers in Group 2 have posterior glottal openings.

For the male speakers, means, maxima, and minima are predicted equally well. To give some idea of the variation observed for our subjects, Fig. 8 shows typical vowel spectra for two male speakers, one (M20) who falls at the low end of the male range for the acoustic measures and one (M18) at the high end. Subject M20 has a relatively low value of $H1$–$H2$ (about -3 dB, compared with about 4 dB for subject M18). In addition, the first three formant peaks, but especially that of F_1, are more sharply defined than those of subject M18. Another difference is that the spectrum for M18 becomes noisy above about 2000 Hz, while that of M20 does not show any evidence of noise.

Fig. 9 is similar to Fig. 7 expect that both male and female data are included; different symbols are used to distinguish the Group 1 and Group 2 females. This graph shows that these parameters vary widely for both males and females. In addition, the mean values for both measures are lower for the male speakers than for the females. This difference is especially large for $H1^*$–$A3^*$, suggesting that spectral tilt may be important for distinguishing male and female voice quality. Finally, there is very little overlap among the Group 2 females and the male subjects. If our hypothesis that Group 2 females have relatively large posterior glottal openings and nonsimultaneous vocal-fold closure is correct, this result suggests that this glottal configuration, while common among female speakers, is not characteristic of males.

These results provide support for our acoustic models, and show that there is rather wide variation in glottal configuration among both males and females. Not only can the models make predictions about a normal population, but it is also possible to work

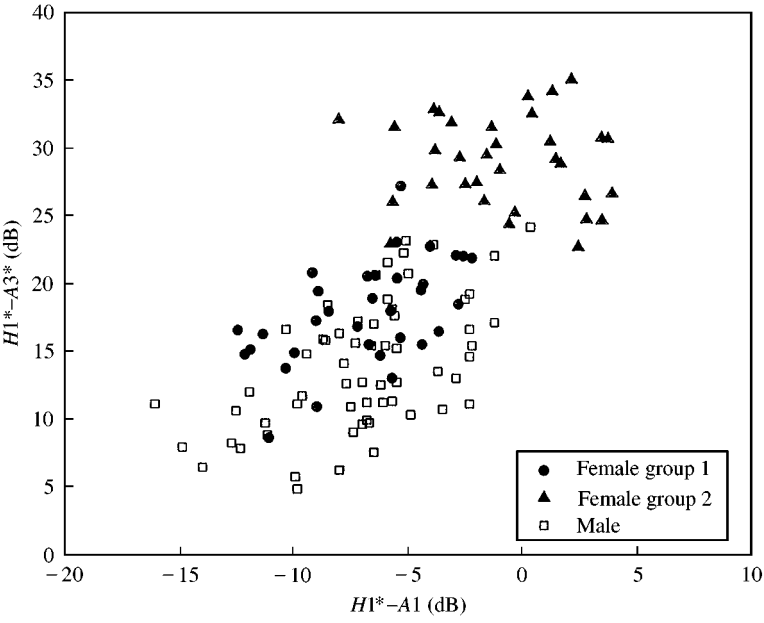

Figure 9. Relation between the measures $H1^*$–$A3^*$ and $H1^*$–$A1$ for 22 female speakers and 21 male speakers. Each point represents data for one of three vowels per speaker. The data for the female speakers have been divided into two groups, described in the text.

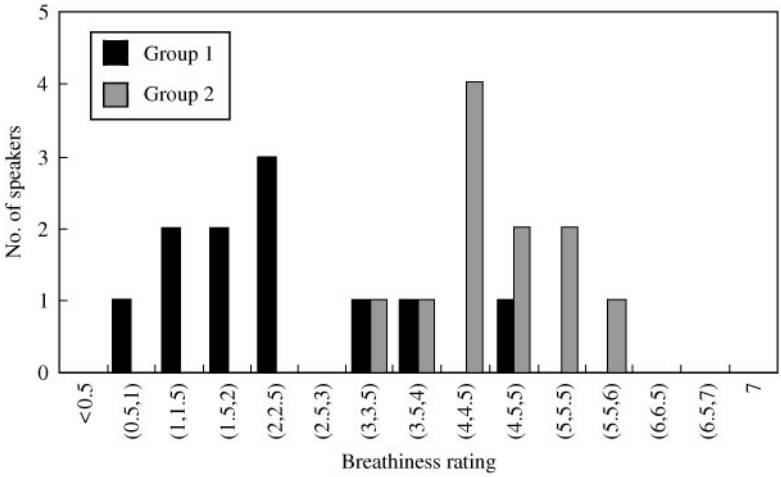

Figure 10. Results of a listening test in which the 22 female speakers were rated for perceived breathiness. A rating of 1 corresponds to "full voice" and a rating of 7 corresponds to "very breathy". Group 2 speakers tended to be perceived as being much breathier than Group 1 speakers. (From Hanson, 1995.)

backwards from the experimental data to make hypotheses about the glottal configuration. Some further tests with the female data were carried out to determine how strong these hypotheses might be (Hanson, 1995). The results are summarized here. First, the vocal folds during sustained vowel production were observed via nasal endoscopy for four of the subjects. Two of these subjects were from Group 1 and two were from Group 2. The Group 2 speakers were observed to have posterior glottal openings that were large relative to those of the Group 1 speakers. Second, four listeners rated the vowels of all of the female speakers for breathiness on a scale of 1–7, where 1 corresponds to full (that is, modal) voice and 7 corresponds to extremely breathy. The hypothesis was that the Group 2 speakers would be perceived to be breathier than the Group 1 speakers. A histogram of the results, shown in Fig. 10, verifies that Group 2 speakers were perceived to be breathy more often than Group 1 speakers. Because breathiness has been found to be correlated with the size of a posterior glottal opening (Södersten & Lindestad, 1990), these results provide further support for our hypotheses about the glottal configurations of our female speakers.

3.3. *Variations within an individual*

In the experimental data described in the previous section, variations were observed among individuals for syllables carrying prominence. In this section, we briefly review evidence that glottal characteristics can also vary widely for a given individual, depending on prosody or emotional state.

Stevens (1994) presented preliminary data on the effects of prosody on glottal characteristics. Fig. 11 compares spectra and waveforms sampled from the sentence "The bat you gave him is lazy", produced by a female speaker. The first spectrum is sampled from the full vowel in "bat" and the second is sampled from the reduced vowel in "is". The measure $H1–H2$ is much larger for the reduced vowel than for the full vowel. It can also

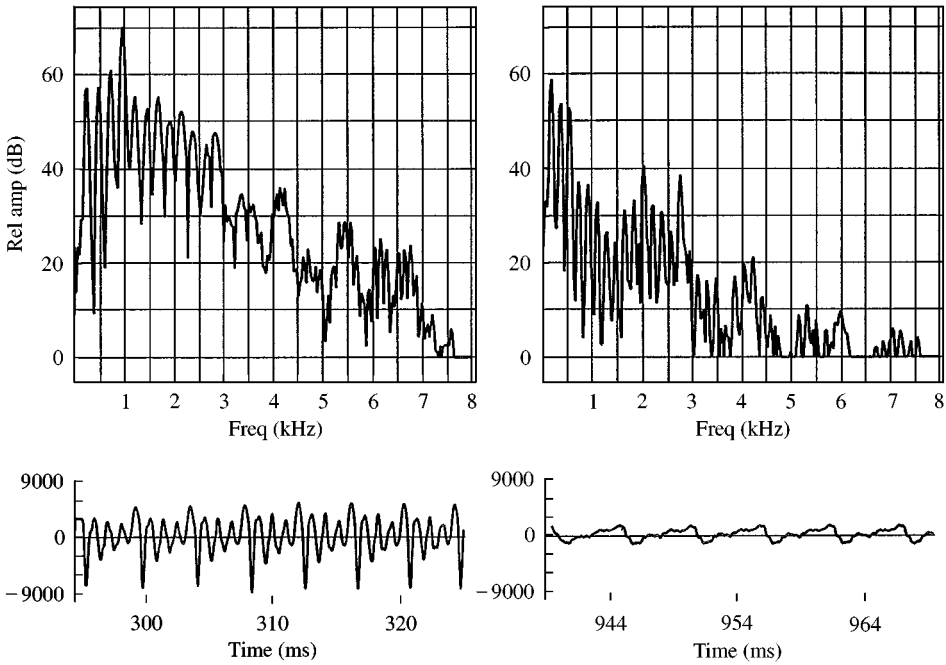

Figure 11. Spectra sampled from the two italicized vowels in the sentence "The b*a*t you gave him *i*s lazy", produced by a female speaker. The spectra were obtained with a 30-ms window. (After Stevens, 1994.)

be seen, from both the spectra and the waveforms, that the first-formant bandwidth is much greater for the reduced vowel. In addition, the spectral tilt is greater for the reduced vowel. These differences suggest that the reduced vowel is produced with a greater open quotient and a larger glottal opening. Stevens (1994) made measures on utterances collected from four subjects to compare glottal characteristics of (1) reduced vowels *vs.* nonreduced; (2) full vowels occurring before and after a nuclear pitch accent; and (3) full vowels occurring early and late in an utterance. All of these conditions resulted in considerable changes in the glottal characteristics of the vowels.

An example of the results of this study is given in Fig. 12. For each of the speakers, spectra were sampled in 17 reduced vowels in six sentences. Similar spectra were also obtained for a pitch-accented vowel that had an amplitude equal to or greater than that of any other vowel in the sentence. The figure shows the average amplitude of the F_1 peaks of the prominent vowels relative to the amplitudes for the reduced vowels. For the reduced vowels, the amplitude is weaker by 7–13 dB, on average, but with considerable variability within each speaker. This difference in amplitude is presumably caused by a larger glottal opening for the reduced vowels. As illustrated in Fig. 11, the reduced vowels also showed greater spectral tilt than the full vowels.

Emotional or attitudinal state may also cause the voice source to vary within a given individual. In an early work, Williams & Stevens (1972) studied the effect of emotions on the acoustic speech signal. Their subjects were professional actors who were asked to speak as if they were conveying various emotions. They reported that for three speakers the energy in various frequency bands varied with emotional state. For example,

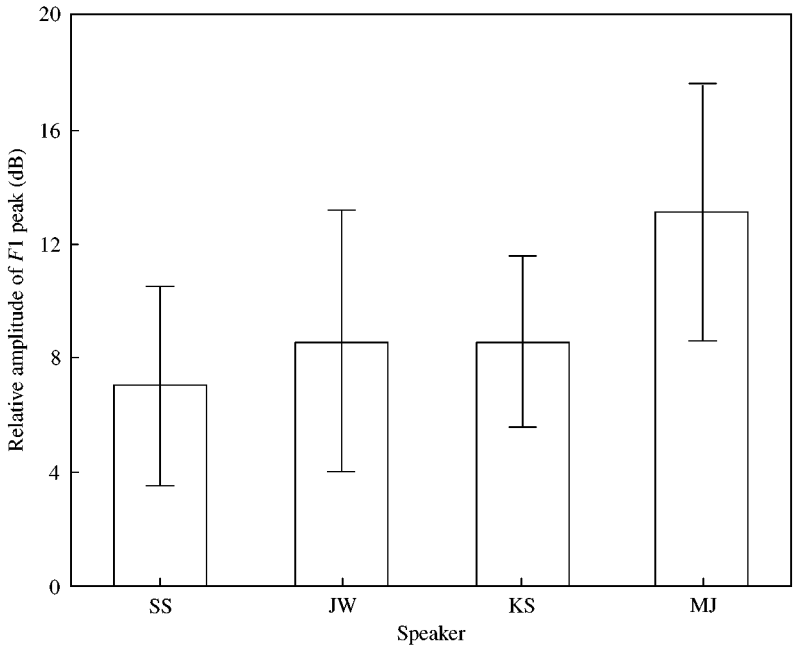

Figure 12. In several sentences, the amplitude of the F_1 peak in the spectra of reduced vowels was measured relative to the strongest $F1$ peak in the pitch-accented vowels. Average differences for 17 reduced vowels are given for each of two female (SS and JW) and two male (KS and MJ) speakers. Standard deviations are shown by vertical lines. (After Stevens, 1994.)

expressing anger seemed to increase the energy at high frequencies, while expressing sorrow decreased it. Liénard & Di Benedetto (1999) found that changes in the voice source occur simply with changes in vocal effort, for example when speakers are located at varying distances from listeners and must adjust their speaking level accordingly.

In summary, we have described a body of work suggesting that glottal characteristics can vary widely across individuals and between male and female speakers. In addition, depending on emotion and prosodic events, they can also vary widely for a given individual. In most of these studies we have used acoustic models that include nonmodal effects to predict and analyze data. The results show that inclusion of nonmodal effects is necessary to account for the observed variations.

3.4. *Speech synthesis with nonmodal effects*

The traditional approach to the synthesis of speech is to implement a model that has separate components for the sources of sound in the vocal tract and for the filtering of these sources by the resonances of the airways that form the vocal tract. Because of the wide variation of the properties of the glottal source both across speakers and within a speaker, it is necessary to include parameters that can control properties that cover a range of nonmodal types of phonation. Early attempts to synthesize speech with a fixed waveform of the glottal pulse led to an unnatural sounding speech output. We describe here a speech synthesizer, HLsyn, in which the glottal characteristics are manipulated

with quasi-articulatory parameters. The settings can also be varied to produce different individual voice qualities. HLsyn has been described in some detail in previous papers (Stevens & Bickley, 1991) and in an upcoming paper (Hanson & Stevens, submitted). This synthesizer provides "higher-level" control of a Klatt formant synthesizer (Klatt & Klatt, 1990), hence the name HLsyn. It has only 13 parameters, which are quasi-articulatory. HLsyn contains a set of mapping relations which transform its 13 parameters to a set of about 50 Klatt acoustic parameters. A Klatt synthesizer is then used to produce output speech.

The Klatt parameters related to voice quality are the amplitude of voicing (AV), amplitude of aspiration noise (AH), spectral tilt (TL), formant bandwidths ($B1$, $B2$, etc.), especially $B1$, open quotient (OQ), fundamental frequency ($F0$), and degree of diplophonia (DI). The HLsyn parameters that affect these Klatt parameters include subglottal pressure (**ps**), average area of the glottis between the membranous portion of the folds (**ag**), area of a posterior glottal opening (**ap**), and fundamental frequency (**f0**). There is also a parameter than can be used to change the compliance of the vocal folds (**dc**).

The following examples illustrate how the HLsyn mapping relations produce Klatt parameters that result in variations in voice quality corresponding to nonmodal phonation:

- The parameters AV and AH are controlled by the glottal areas and the subglottal pressure. For example, if the parameter **ag** is increased at the end of an utterance to emulate abduction of the vocal folds, AV decreases while AH increases. An increase in subglottal pressure (**ps**), assuming all else remains the same, results in increases in both AV and AH.
- An increase in the parameter **ap** is mapped to increases in TL, AH, and the formant bandwidths, producing a breathier percept.
- OQ is increased if **ag** increases. Parameter **ag** is usually increased when voiceless consonants are synthesized. By starting the abduction gesture during a preceding vowel, and extending this gesture somewhat into a following vowel, one can emulate the breathy phonation often reported to be observed in vowels adjacent to voiceless obstruents (see, for example, Ní Chasaide & Gobl, 1993). Prior to forming the constriction for the obstruent, vowel amplitude decreases, aspiration noise increases, and open quotient increases. Likewise, at the onset of voicing, there might be a slightly reduced amplitude of voicing, increased aspiration noise, and increased open quotient, depending on the time taken to complete the adduction gesture for the vowel.
- Voicing is cutoff (that is, AV is set to 0) if the glottal area is either too small or too large, or if transglottal pressure is too low. The transglottal pressure threshold can be changed by adjusting the tension of the vocal folds using parameter **dc**.
- When the parameter **ag** is relatively small (but not small enough to cut off voicing), the Klatt parameter DI is available to emulate a form of irregular vocal-fold vibration to simulate glottalization.

The top panel of Fig. 13 is a spectrogram of the phrase "played basketball" excised from the sentence "Five women played basketball", produced by a female speaker; the bottom panel is our best attempt at copy synthesis of this utterance using HLsyn. The natural utterance has several instances of irregular vocal-fold vibration, indicated by the label "I"; two of these are brief, but the third extends throughout the phrase-final syllable

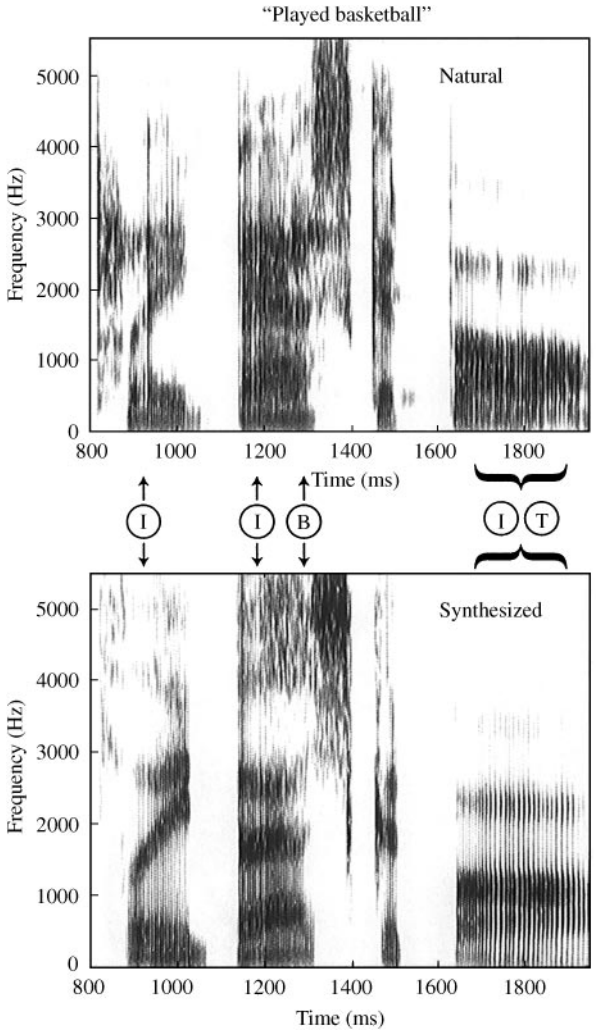

Figure 13. An example of copy synthesis using HLsyn. The utterance is "played basketball", excised from the sentence "Five women played basketball", produced by a female speaker. The top panel is the natural utterance and the bottom is the copy synthesis. Points labelled "I" indicate irregular vibration of the vocal folds; "B" indicates breathy phonation; and "T" indicates increased spectral tilt.

"ball". Also note that in the /æ/ of "basket", the glottal source appears to become breathy for about 50 ms before the /s/ (labeled "B"). The spectrogram becomes noisier above about 1500 Hz, as might be expected during a vowel preceding a voiceless obstruent. During the word "ball" there is a sharp reduction in high-frequency energy, relative to vowels occurring earlier in the utterance. In the synthesized version, the irregularities during "played" and "basket" are difficult to see, but they are a bit more clear in the word "ball". Note that the spectral tilt also increases during "ball", due to spreading of the glottis (**ag**) and a steady reduction in subglottal pressure. The transition from /æ/ to /s/ in "basket" becomes progressively breathier. Listeners judge the synthetic version to sound

like the speaker who produced the original utterance (McGowan, Hanson, Stevens & Gow, 1999). This example, then, illustrates how the incorporation of acoustic models of nonmodal phonation into the synthesizer allows us to emulate nonmodal phenomena that occur in normal speakers.

4. Applications to disordered speech

In the previous section, we described the application of nonmodal phonation models to normal speakers. Acoustic measures believed to reflect glottal configuration were shown to vary widely across a group of 43 adult speakers. In addition, we discussed how these measures can vary throughout an utterance or in different emotional situations for a given individual. As we suggested in the introduction, we might also expect to find that models of nonmodal phonation can be useful to understand differences among populations, specifically between speakers with and without speech disorders. In this section, we explore this possibility.

Although one might assume that such models would be relevant to disorders that are specific to voice, for example nodules, they are also applicable to disorders such as the dysarthrias and deaf speech. Dysarthria is a motor speech disorder, and voice production can be affected as much as articulation. Likewise, the lack or reduction of auditory feedback in the hearing impaired may lead to deviations in voice quality that are nonmodal, as well as imprecise articulation.

In this section, we describe the application of nonmodal phonation models to two disorders. We begin by describing a study by Kuo (1998) on phonation models and speakers with vocal-fold nodules. We then turn to work by Chen (in prep.), in which the acoustic measures described in the previous section were applied to the speech of individuals with neuromotor disorders.

4.1. *Vocal nodules*

Vocal-fold pathologies are known to affect the vibratory characteristics of the folds, although the details for particular pathologies are not well understood (see, for example, Hillman, Holmberg, Perkell, Walsh & Vaughan, 1989). In recent work, Kuo (1998; Kuo, Holmberg & Hillman, 1999, in prep.) focused on the effects of vocal-fold nodules on both aerodynamic and acoustic characteristics of speech. When vocal nodules are present, the vocal folds often do not close completely during phonation (Colton, Woo, Brewer, Griffin & Casper, 1995). One might expect, then, that the acoustic effects of this incomplete closure would be similar to those described in Section 3 for normal speakers with posterior glottal openings; that is, the larger glottal opening would lead to increases in the first-formant bandwidth, source spectral tilt, and aspiration noise. As we will describe, however, aerodynamic measures for a group of female speakers with nodules were found to be significantly different from those for a group of normal speakers, while acoustic measures were not. Kuo (1998) was able to explain this surprising result by using a modified version of the two-mass model (Ishizaka & Flanagan, 1972).

Kuo analyzed data from 26 women: 12 normal subjects and 14 with bilateral vocal nodules. The women with nodules were judged by a clinician to be mildly dysphonic. Subjects were seated in a sound-isolated booth and fitted with a Rothenberg mask, which was used to measure oral airflow during speech production. A transducer was fitted to

TABLE I. Average data collected from normal subjects and nodules patients speaking in both comfortable and loud voices

Condition	Group	P_s (cm H_2O)	$H1$–$A1$ (dB)	SPL (dB)
Comfortable	Normals	6	− 1	77.5
	Nodules	11	1	80.1
Loud	Normals	9	− 9	86.6
	Nodules	15	− 6	88.8

the mask and inserted between the lips to measure intraoral pressure. Subjects were asked to repeat the syllable /pæ/ at both comfortable and loud speaking levels. The acoustic speech signal was recorded at a distance of 15 cm. Following the recordings, the oral airflow signal was lowpass filtered at 1100 Hz and inverse-filtered to derive the glottal airflow, from which aerodynamic features were extracted. The acoustic signal was digitized at a rate of 10 kHz. The SPL was calculated from the root-mean-square amplitude of the acoustic signal and a calibration signal. The subglottal pressure was inferred from interpolation of the intraoral pressure during the closures of adjacent /p/'s in the /pæ pæ … / sequences. (For more details, see Perkell, Holmberg & Hillman (1991), which closely approximates the procedures that were followed in collecting the data used by Kuo.)

Several aerodynamic and acoustic measures were extracted from the middle of each vowel. Aerodynamic features included subglottal pressure, average flow, open quotient, and maximum flow declination rate (MFDR). Acoustic measures included SPL, and the measures $H1$–$A1$ and $H1$–$A3$ described in Section 3. In this paper, we briefly describe the results; greater detail is given in Kuo (1998) and Kuo *et al.* (1999, in prep.). As we stated earlier, one might expect to find significant differences between the normal controls and the nodules patients in both the aerodynamic and acoustic measures. The nodules patients were found to have, on average, higher values for all the aerodynamic measures. They also had SPLs that were 2–3 dB higher, on average. However, the remainder of the acoustic measures differed by less than 3 dB, on average.

As an example, Table I summarizes the average data across speakers for one aerodynamic measure, subglottal pressure P_s, and two acoustic measures, SPL and $H1$–$A1$. For both the "comfortable" and "loud" conditions, the average subglottal pressures for the nodules subjects are far outside the ranges typically reported for normal females (cf. Holmberg *et al.*, 1988). However, for the measure $H1$–$A1$ there is only a 2–3 dB difference between the averages for the two groups of subjects. Referring to the range of $H1$–$A1$ found for normal females by Hanson (1997) (Fig. 7), we see that this difference is negligible. Note also that the differences in SPL between the two groups are only about 2–3 dB, far less than what might be expected based on the subglottal-pressure differences. Looking only at the acoustic measures, we might surmise that the glottal configurations for the two groups are similar, despite the fact that one group has nodules, which should increase the average glottal area. The aerodynamic data, however, are clues to the real story. Differences in subglottal pressure such as those exhibited by the two groups studied should lead to large differences in the acoustic measures, all else being the same. That is, if the average glottal area were the same for the nodules and normal subjects, the differences in SPL should be larger than those observed. In addition, the glottal losses

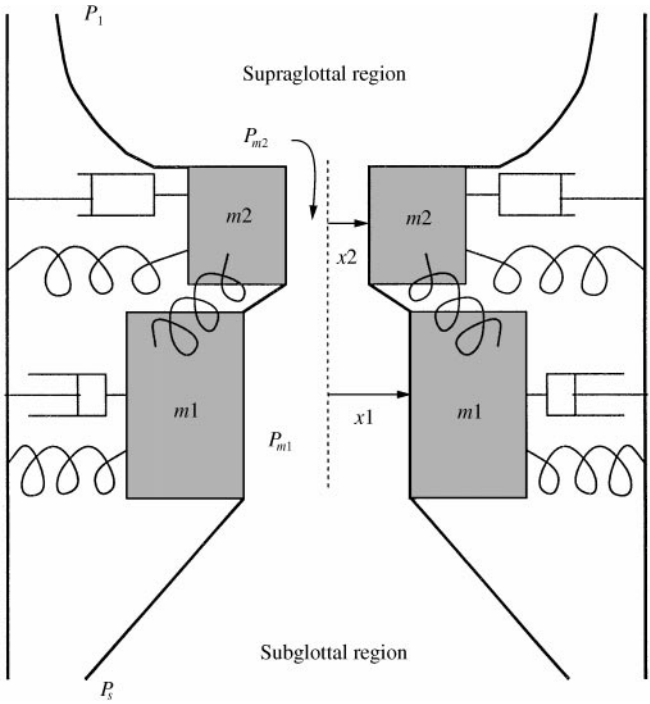

Figure 14. A two-mass model of the vocal folds. The variables m1 and m2 are the lower and upper masses, respectively. The subglottal pressure is P_s and the pressure above the folds is P_1. The pressure on the lower and upper masses are P_{m1} and P_{m2}, respectively. (From Kuo, 1998.)

would decrease (Hanson, 1997), and the measure $H1–A1$ should decrease significantly. Taken together, then, the aerodynamic and acoustic data suggest that nodules patients do have larger glottal areas, on average, but they compensate for the acoustic changes that would occur by speaking with increased subglottal pressure.

To further understand the effects of vocal-fold nodules on the aerodynamics and acoustics of speech production, Kuo (1998) implemented a version of the two-mass model of vocal-fold vibration (Ishizaka & Flanagan, 1972). Fig. 14 shows a schematic of the mechanical portion of a two-mass model. In his model, Kuo replaced the aerodynamic component proposed by Ishizaka & Flanagan (1972) with one proposed by Story & Titze (1995), and adjusted certain parameters to be more appropriate for female speakers. In one version of his model, Kuo also adjusted certain parameters to model the effects of nodules, schematicized in Fig. 15. Namely, (1) the masses were increased, (2) the stiffness of the coupling between the upper and lower masses was increased, and (3) collision forces were brought into play when the nodules came into contact. As can be seen in Fig. 15, the third change means that air can continue to flow through the glottis after collision occurs.

Fig. 16(a) shows the output U_g of Kuo's "normal" model, that is, for the case which does not include nodule effects. Its derivative $\dot{U}_g$ is shown in part (b). Subglottal pressure was set to 8 cm H_2O. The output is fairly typical for a modal, two-mass model. Fig. 17(a) and (b) shows the output for the "nodules" model, with subglottal pressure at 8 cm H_2O.

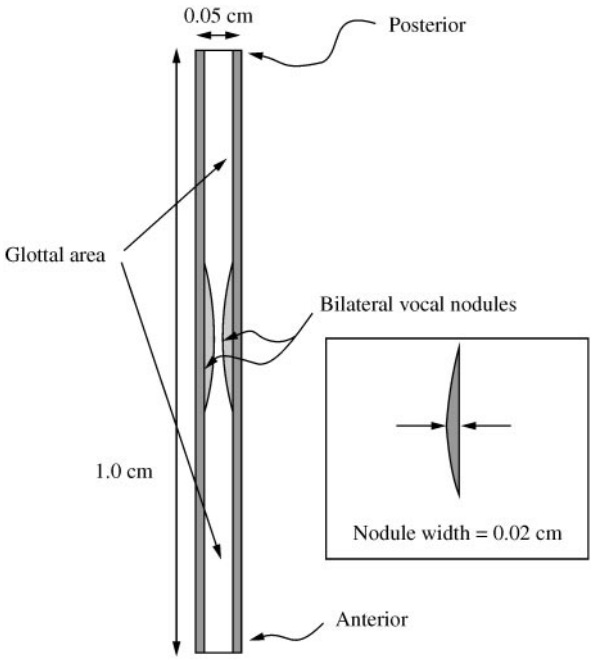

Figure 15. A schematic of bilateral vocal-fold nodules. Note that when the nodules are included in the Kuo (1998) two-mass model, collision forces come into play when the nodules meet and air can continue to flow through the folds after collision occurs. (From Kuo, 1998.)

The changes introduced by the nodules are what one might expect when the folds do not close completely during a vibratory cycle: the flow is not completely cut off during the maximally closed portion of the cycle, the slope of U_g is less abrupt during the closing phase, and following the point at which the folds collide at the nodules (that is, following the large negative peak in $\dot{U}_g$) there is a "return phase". In addition, the open quotient is increased. Finally, Fig. 17(c) and (d) shows the results of a simulation in which the subglottal pressure in the "nodules" model was set to 11 cm H_2O, the average value found for nodules subjects in comfortable voice. By increasing the subglottal pressure, the output of the "nodules" model more closely approximates the output of the "normal" model (Fig. 16): the minimum flow component disappears, the slope of U_g prior to closure increases, and the return phase is more abrupt. However, the waveform varies from the normal case in that open quotient and AC flow are relatively large.

Recall that although the nodules subjects employed relatively high subglottal pressures while speaking, the resulting SPLs were not much higher than those of the normal subjects. Kuo used both versions of his model to generate glottal waveforms for ranges of subglottal pressures typical of normal and nodules subjects. These glottal waveforms were then used as sources in the synthesis of vowels, and the corresponding SPLs were computed. The results are plotted in Fig. 18(a). The straight line in the plot has a slope of 3/2 and illustrates the relationship between SPL and P_s predicted by the 3/2 power law reported in several earlier studies (e.g., Ladefoged, 1962). Much higher subglottal pressures are required by the "nodules" model than by the "normal" model to get similar SPLs. Fig. 18(b) plots SPL as a function of P_s for two of Kuo's subjects, one normal and

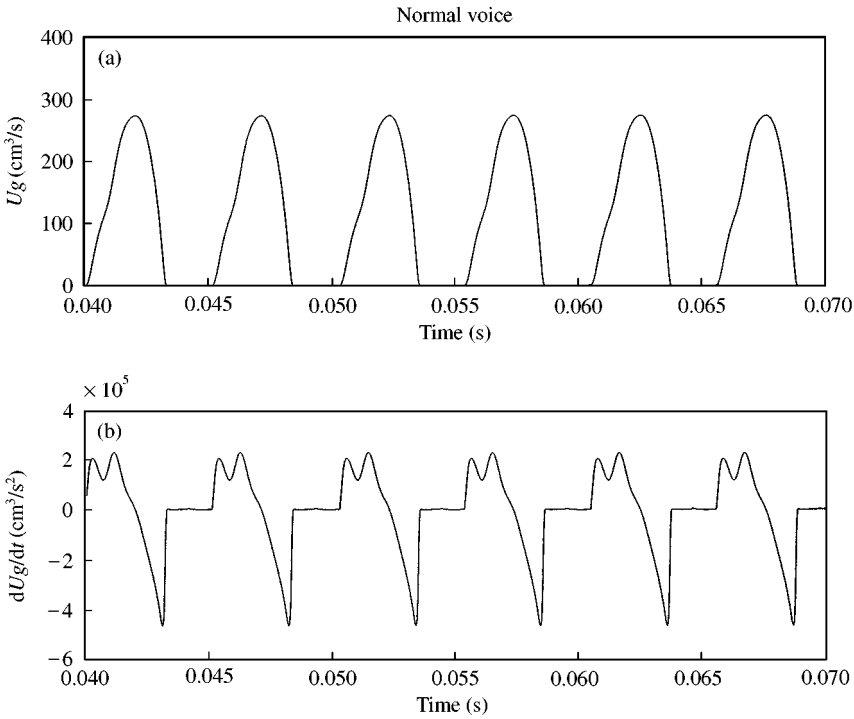

Figure 16. A simulation of normal female voice, with subglottal pressure set at 8 cm H_2O. (a) The volume-velocity waveform $U_g(t)$. (b) The derivative of the volume-velocity waveform $\dot{U}_g(t)$. (From Kuo, 1998.)

one with nodules. These data exhibit a relationship quite similar to those generated by the models, suggesting that Kuo's model captures fairly accurately the effects of nodules on vocal-fold vibration.

4.2. *Dysarthria*

The dysarthrias are motor speech disorders that typically result from neurological disorders such as Parkinson's disease or multiple sclerosis (Kent, Kent, Duffy & Weismer, 1998). Respiration, phonation, and articulation may all be affected. The degree to which they are affected varies from one individual to another, depending in part on the type of dysarthria. The perceived voice quality of dysarthric speakers covers a large range, including strained, harsh, hoarse, or breathy. Fig. 19 shows spectrograms of the word "see" produced by a normal subject (a) and a dysarthric subject with spastic cerebral palsy (b). Comparison of the voiced portions of the two spectrograms suggests that the voice source of the dysarthric subject is noisier and less stable than that of the normal subject. Voice onset for the normal subject is rather abrupt, but for the dysarthric subject there seem to be two phases. For the first 50 ms or so, the periodic source appears to have mainly low-frequency components, while the excitation above about 1500 Hz is primarily aspiration noise. At about 625 ms, the high-frequency components of the periodic source suddenly gain in amplitude. During the last 100 ms or so of the vowel, the

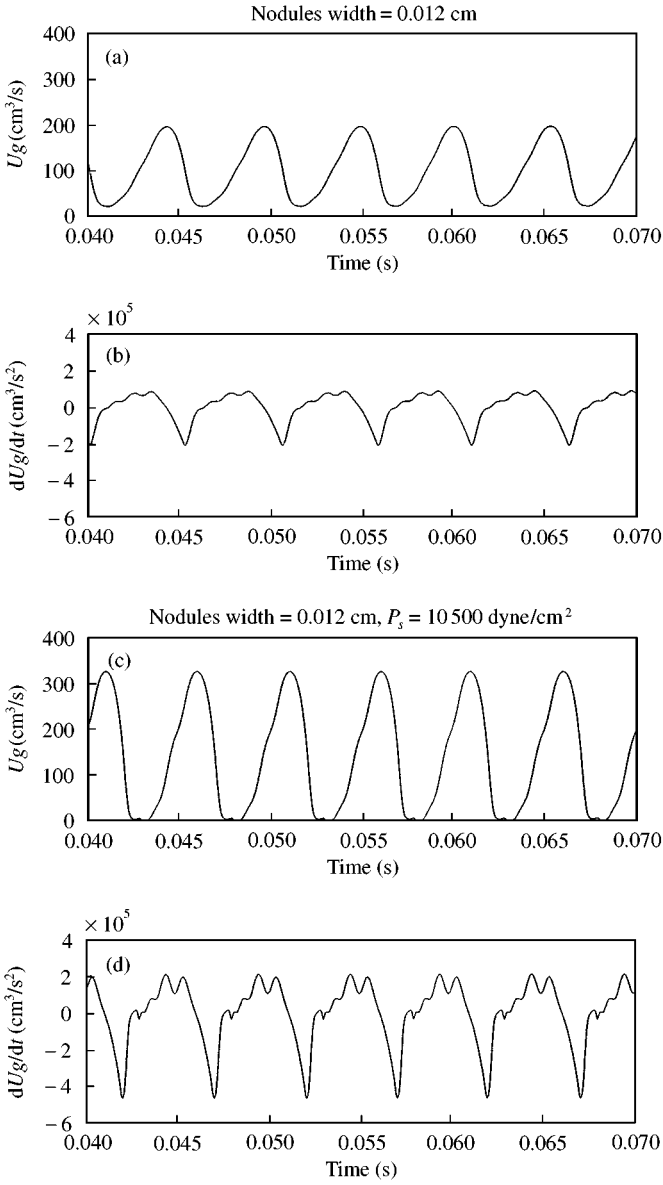

Figure 17. Simulations of female voice with nodules. (a) The volume-velocity waveform $U_g(t)$ for the case in which subglottal pressure was set at 8 cm H_2O. (b) The derivative of the volume-velocity waveform $\dot{U}_g(t)$, corresponding to (a). (c) The volume-velocity waveform $U_g(t)$ for the case in which subglottal pressure was set at 11 cm H_2O. (d) The derivative of the volume-velocity waveform $\dot{U}_g(t)$, corresponding to (c). (After Kuo, 1998.)

source appears to become breathier. The latter observation is based on the increase in spectral tilt beginning at about 825 ms.

In recent work, Chen (in prep.) applied the models proposed by Hanson (1995, 1997) and described in Section 3.1 to vowels produced by dysarthric speakers. Her subjects

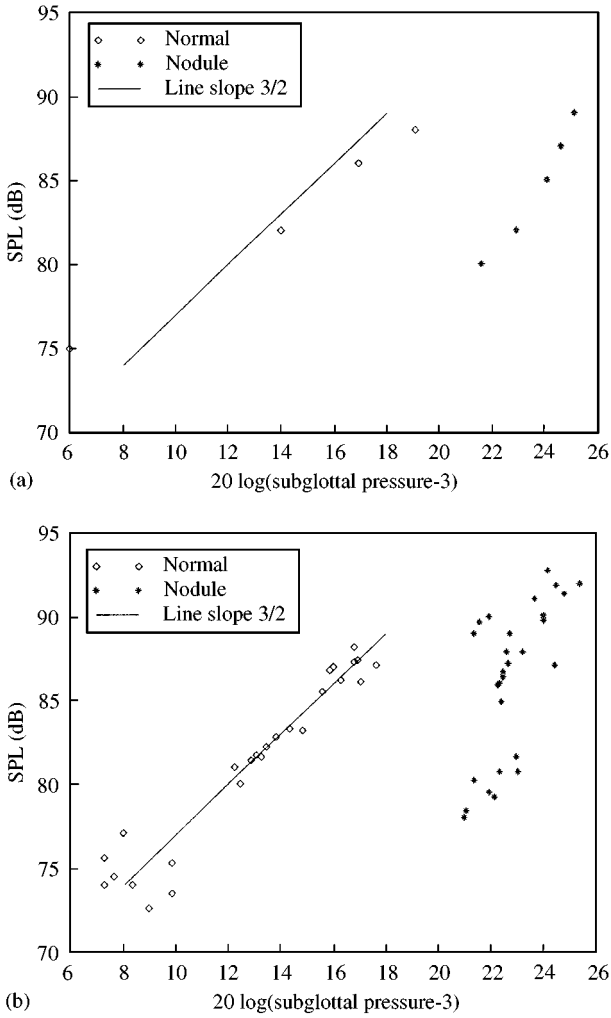

Figure 18. SPL *vs.* subglottal pressure. (a) Simulation of normal and nodules subjects. (b) Data collected from one normal subject and one nodules patient. (From Kuo, 1998.)

included six speakers with spastic, ataxic, or athetoid dysarthria (three female and three male) and four normal controls (two female and two male). Recordings of 29 words were analyzed. The words were monosyllabic, beginning with either /s/, /ʃ/, or a stop consonant, and containing one of 10 vowels. The corpus included five tokens of each word. The acoustic measures $H1^*–H2^*$, $H1^*–A1$, and $H1^*–A3^*$ were made near the center of the vowel, and the noise ratings N_w and N_s were made based on observations at the beginning, 100 ms from the beginning, and at the center of the vowel.

Fig. 20(a) is a plot of $H1^*–A3^*$ *vs.* $H1^*–A1$ for the female speakers. The solid points on this graph represent words produced by two normal speakers and the open points represent words produced by dysarthric speakers. The solid lines indicate the ranges for these measures found for normal females by Hanson (1995, 1997) and shown in Fig. 7;

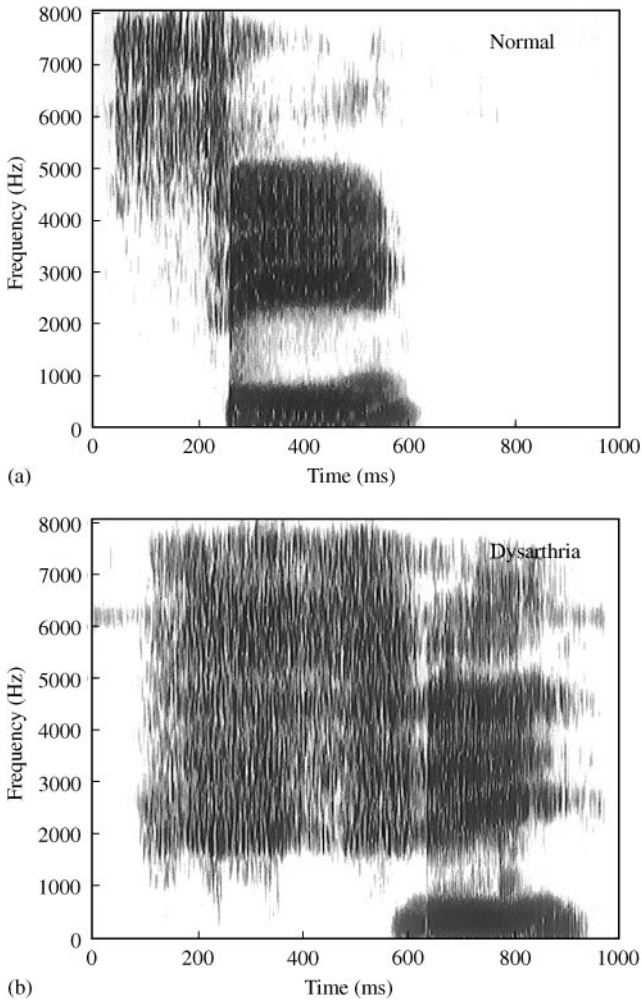

Figure 19. Spectrograms of the word "see" by (a) a normal and (b) a dysarthric speaker with spastic cerebral palsy. The voice source for the dysarthric subject appears to be noisier and less stable.

as can be seen, very few of the points fall outside of these ranges. The data for the two groups of speakers almost entirely overlap. Fig. 20(b) is the same plot for the male speakers. Here we see a different picture. One of the dysarthric speakers, DM2, falls within the normal ranges for both measures, and most of his data seem to overlap with those of the normal subjects. The other two dysarthric speakers are very different. Along the $H1^*–A1$ dimension, most of their data fall within the normal range, but they both display much greater variability. Along the $H1^*–A3^*$ dimension, most of their data fall outside of the normal range, and they also display more variability than the normal speakers. Speaker DM1 has $H1^*–A3^*$ values which are greater than normal, while DM3 falls below the normal range.

Fig. 21 is a bar graph of the noise ratings N_s and N_w. The data for the normal subjects have been averaged for this graph. We see that the ratings for the normal subjects fall

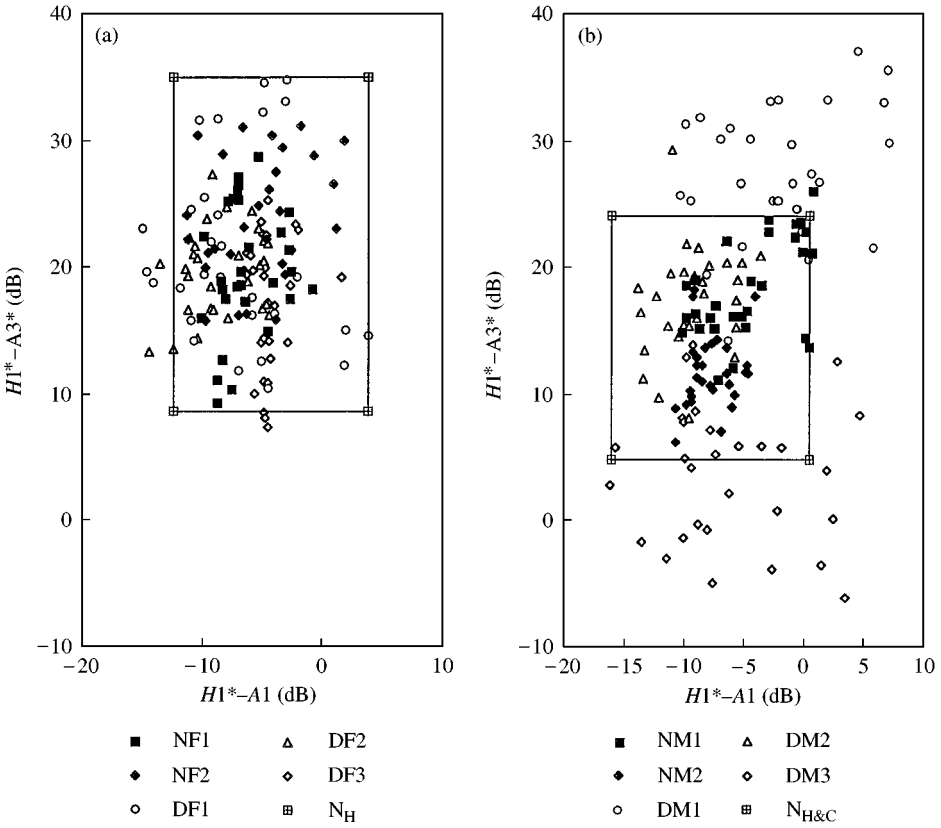

Figure 20. Relation between $H1^*–A3^*$ and $H1^*–A1$ for normal and dysarthric subjects studied by Chen (in prep.). (a) Female subjects. Solid lines indicate ranges reported by Hanson (1995, 1997) for normal females. (b) Male subjects. Solid lines indicate ranges reported by Hanson & Chuang (1999) for normal males. For the female subjects, the type of dysarthria is spastic. For the male subjects, DM1 is spastic, DM2 is ataxic, and DM3 is athetoid.

below a rating of 2; that is, there is little evidence of high-frequency noise in their vowel spectra or in the waveform bandpass-filtered in the F_3 region. The dysarthric subjects, however, have average ratings close to or above 3, meaning that there is little evidence of periodicity in their spectra and waveforms at high frequencies.

How is one to interpret these acoustic data? Given the plots in Fig. 20, it would seem that the glottal characteristics of the dysarthric subjects often overlap those of normal subjects, but for some individuals they can also range well above and below those of normal subjects. Based on Fig. 20, one might expect that only subject DM3 would have noise ratings higher than the normal subjects. However, as we have just seen, the noise ratings for the dysarthric subjects were uniformly higher than those for the normal subjects. Therefore, it seems unlikely that the glottal configurations for the dysarthric subjects are normal.

Although we cannot make any conclusions based on the small amount of data presented here, it is possible that, much like the nodules patients studied by Kuo (1998), the dysarthric subjects may be able to produce speech that is within the normal range

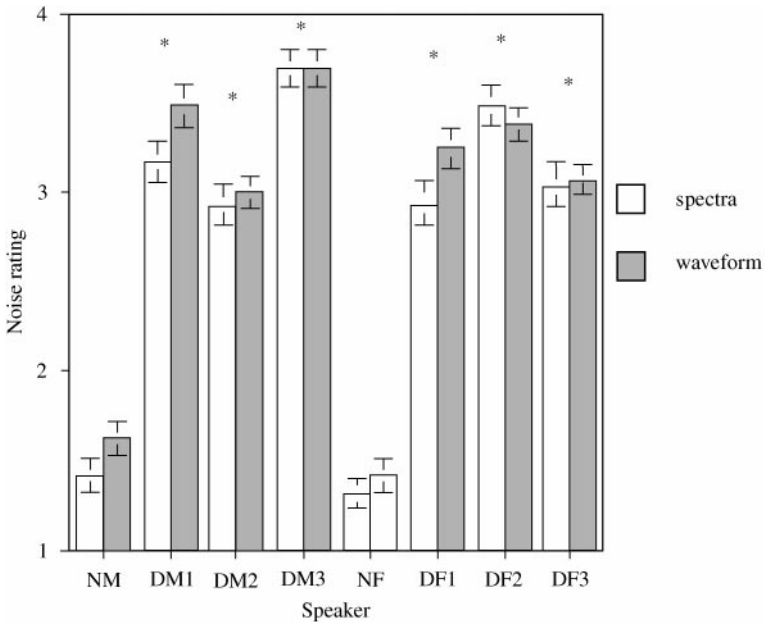

Figure 21. Noise ratings for normal and dysarthric subjects. Asterisks above bars indicate significant differences from normal subjects. (From Chen, in prep.)

along certain acoustic dimensions, despite having abnormal glottal configurations. They may use a high subglottal pressure to create a periodic component of the source that has spectral characteristics within normal limits (at least for some speakers) but with a stronger turbulence noise component because of the increased average flow. Other acoustic parameters or aerodynamic measures may be necessary to more completely describe the disordered phonation.

5. Summary

In this paper, we have shown that models of modal vocal-fold phonation can be extended to include a range of nonmodal phonation types. These models predict some aspects of the sound source and filtering when the glottal configuration and vocal-fold state are known. They can also be used to infer the glottal configuration and state from acoustic measurements on the radiated sound. We have reviewed acoustic data which suggest a large range of glottal configurations across female and male speakers for normal phonation. For a given speaker, there also appears to be a range of glottal configurations, depending on prosodic characteristics within an utterance and on the emotional state of the speaker. The models can also be used to examine attributes of phonation associated with neuromotor and laryngeal disorders. Study of the characteristics of the glottal source for speakers with these disorders suggests that these speakers may modify their subglottal pressure (and possibly other physiological parameters) in relation to those used by normal speakers in order to achieve acoustic patterns that are closer to the normal range. Some aspects of the phonation of speakers with vocal-fold nodules can be simulated with a modified phonation model.

The work described in this paper was supported in part by NIH grants DC00075, DC00266, and F32 DC00205 to MIT; NIH grant MH52358 to Sensimetrics Corp.; and a US Department of Education fellowship award to the Harvard University Division of Applied Sciences. The preparation of the manuscript was supported in part by NIH grant DC04331. We gratefully acknowledge Stefanie Shattuck-Hufnagel for her comments on an earlier version of the manuscript.

References

Baer, T. (1975) *Investigation of phonation using excised larynges*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Chen, M. Y. (in prep.) Voice characteristics of vowels spoken by dysarthric speakers.

Colton, R. H., Woo, P., Brewer, D. W., Griffin, B. & Casper, J. (1995) Stroboscopic signs associated with benign lesions of the vocal folds, *Journal of Voice*, **9,** 312–325.

Cranen, B. & Schroeter, J. (1995) Modeling a leaky glottis, *Journal of Phonetics*, **23,** 165–177.

Dilley, L. C. & Shattuck-Hufnagel, S. (1995) Variability in glottalization of word onset vowels in American English. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, *ICPhS '95*, Stockholm, Vol. 4, pp. 586–589.

Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M. (1996) Glottalization of word-initial vowels as a function of prosodic structure, *Journal of Phonetics*, **24,** 423–444.

Fant, G. (1960) *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1972) Vocal tract wall effects, losses, and resonance bandwidths, *Speech Transmission Laboratory Quarterly Progress and Status Report*, Royal Institute of Technology, Stockholm. Vol. 2–3, pp. 28–52.

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

Gordon, M. & Ladefoged, P. (2001) Phonation types: a cross-linguistic overview, *Journal of Phonetics*, **29,** 383–406. doi:10.1006/jpho.2001.0147.

Hanson, H. M. (1995) *Glottal characteristics of female speakers*. PhD thesis, Harvard University, Cambridge, MA.

Hanson, H. M. (1997) Glottal characteristics of female speakers: acoustic correlates, *Journal of the Acoustical Society of America*, **101,** 466–481.

Hanson, H. M. & Chuang, E. S. (1999) Glottal characteristics of male speakers: acoustic correlates and comparison with female data, *Journal of the Acoustical Society of America*, **106,** 1064–1077.

Hanson, H. M. & Stevens, K. N. (submitted) Control of acoustic source parameters in speech synthesis: a quasi-articulatory approach.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M. & Vaughan, C. (1989) Objective assessment of vocal hyperfunction: an experimental framework and initial results, *Journal of Speech and Hearing Research*, **32,** 373–392.

Holmberg, E. B., Hillman, R. E. & Perkell, J. S. (1988) Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice, *Journal of the Acoustical Society of America*, **84,** 511–529. Plus Erratum, *ibid*, **85,** 1787.

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

Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. & Goldman, S. L. (1995) Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice, *Journal of Speech and Hearing Research*, **38,** 1212–1223.

House, A. S. & Stevens, K. N. (1958) Estimation of formant bandwidths from measurements of transient response of the vocal tract, *Journal of Speech and Hearing Research*, **1,** 309–315.

Ishizaka, K. & Flanagan, J. L. (1972) Synthesis of voiced sounds from a two-mass model of the vocal cords, *Bell System Technical Journal*, **51,** 1233–1268.

Ishizaka, K. & Matsudaira, M. (1968) What makes the vocal cords vibrate. In *Proceedings of Sixth International Congress of Acoustics*, Tokyo, pp. B-9–B-12.

Kent, R. D., Kent, J. F., Duffy, J. & Weismer, G. (1998) The dysarthrias: speech-voice profiles, related dysfunctions, and neuropathology, *Journal of Medical Speech-Language Pathology*, **6,** 165–211.

Klatt, D. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, **87,** 820–857.

Kuo, H.-K. J. (1998) *Voice source modeling and analysis of speakers with vocal-fold nodules*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Kuo, J., Holmberg, E. B. & Hillman, R. E. (1999) Discriminating speakers with vocal nodules using aerodynamic and acoustic features. In *Proceedings of the IEEE ICASSP '99*, Vol. 1, pp. 77–80.

Kuo, J., Holmberg, E. B. & Hillman, R. E. (in prep.) Aerodynamic and acoustic characteristics of female speakers with vocal-fold nodules.
Ladefoged, P. (1962) Subglottal activity during speech. In *Proceedings of the Fourth International Congress of Phonetic Sciences*, pp. 73–91. The Hague: Mouton.
Liénard, J.-S. & Di Benedetto, M.-G. (1999) Effect of vocal on spectral properties of vowels, *Journal of the Acoustical Society of America*, **106,** 411–422.
Lucero, J. C. (1995) The minimum lung pressure to sustain vocal fold oscillation, *Journal of the Acoustical Society of America*, **98,** 779–784.
McGowan, R. S., Hanson, H. M., Stevens, K. N. & Gow, D. W. (1999) Talker transformation through synthesis. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 137–140.
Ní Chasaide, A. & Gobl, C. (1993) Contextual variation of the vowel voice source as a function of adjacent consonants, *Language and Speech*, **36,** 303–330.
Perkell, J. S., Hillman, R. E. & Holmberg, E. B. (1994) Group differences in measures of voice production and revised values of maximum airflow declination rate, *Journal of the Acoustical Society of America*, **96,** 695–698.
Perkell, J. S., Holmberg, E. B. & Hillman, R. E. (1991) A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production, *Journal of the Acoustical Society of America*, **89,** 1777–1781.
Pierrehumbert, J. (1995) Prosodic effects on glottal allophones. In *Vocal fold physiology*: *voice quality control* (O. Fujimura & M. Hirano, editors), pp. 39–60. San Diego: Singular.
Redi, L. & Shattuck-Hufnagel, S. (2001) Variation in the realization of glottal events in normal speakers, *Journal of Phonetics*, **29,** 407–429. doi:10.1006/jpho.2001.0145.
Rosenberg, A. E. (1971) Effect of glottal pulse shape on the quality of natural vowels, *Journal of the Acoustical Society of America*, **49,** 583–590.
Siebert, W. M. (1986) *Circuits, signals, and systems.* Cambridge: MIT Press.
Slifka, J. (2000) *Respiratory constraints at prosodic boundaries in speech.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.
Slifka, J. (in prep.) Respiratory system correlates at utterance termination.
Södersten, M. & Lindestad, P.-Å. (1990) Glottal closure and perceived breathiness during phonation in normally speaking subjects, *Journal of Speech and Hearing Research*, **33,** 601–611.
Stevens, K. N. (1994) Prosodic influences on glottal waveform: preliminary data. In *Proceedings of the International Symposium on Prosody*, September 1994, pp. 53–64.
Stevens, K. N. & Bickley, C. A. (1991) Constraints among parameters simplify control of Klatt formant synthesizer, *Journal of Phonetics*, **19,** 161–174.
Story, B. H. & Titze, I. R. (1995) Voice simulation with a body-cover model of the vocal folds, *Journal of the Acoustical Society of America*, **97,** 1249–1260.
Titze, I. R. (1988) The physics of small-amplitude oscillation of the vocal folds, *Journal of the Acoustical Society of America*, **83,** 1536–1552.
Titze, I. R. (1992) Phonation threshold pressure: a missing link in glottal aerodynamics, *Journal of the Acoustical Society of America*, **91,** 2926–2935.
Titze, I. R. (1995) Definitions and nomenclature related to voice quality. In *Vocal fold physiology*: *voice quality control* (O. Fujimura & M. Hirano, editors), pp. 335–342. San Diego: Singular.
Veldhuis, R. (1998) A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation, *Journal of the Acoustical Society of America*, **103,** 566–571.
Wilhelms-Tricarico, R. (unpublished) A modified two-mass model of the vocal folds with a chink and gradual closure.
Williams, C. E. & Stevens, K. N. (1972) Emotions and speech: some acoustical correlates, *Journal of the Acoustical Society of America*, **52,** 1238–1250.