Glottal Characteristics of Female Speakers

A thesis presented

by

Helen M. Hanson

to

The Division of Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Engineering Sciences

Harvard University

Cambridge, Massachusetts

Adviser: Kenneth N. Stevens

May 1995

Abstract

The aim of the research reported in this thesis is to develop a set of descriptors of the voicing source that reflect individual differences in the voice qualities of female speakers. The descriptors are derived from measurements of the spectra of vowels produced by the speakers.

The configuration of the membranous part of the vocal folds and the arytenoid cartilages shapes waveform of the airflow that passes through the glottis during phonation and affects the amount of turbulence noise that is generated at the glottis. Theoretical analysis and observations of experimental data suggest that a more open glottal configuration results in a glottal waveform with relatively greater low-frequency and weaker high-frequency components, compared to a waveform produced with a more adducted glottal configuration. This more open glottal configuration should also result in a greater source of aspiration noise and larger bandwidths of the natural frequencies of the vocal tract (formants), particularly the first formant. These effects of glottal configuration are theorized to be measurable directly from the speech spectrum or waveform. In our work we have developed such acoustic measurements. By applying these measurements to the speech of a group of female speakers, we have shown that they can be used to classify speakers according to glottal configuration. Physiological measures derived from airflow waveforms and from fiberscopic observations for a subset of the subjects are in accord with the classifications based on acoustic measurements. These classifications have also been found to be correlated with perceived voice quality. Through a speech synthesis experiment, we found that the variations in glottal characteristics observed in our group of female subjects are perceptible and contribute to improved synthesis of a given female's speech.

This research contributes to the literature that seeks to describe the normal variation of voicing characteristics across speakers, and therefore has implications for speech and speaker recognition. In addition, it contributes to a continuing effort to improve the analysis and synthesis of female speech, which have often been termed 'difficult'. Furthermore, this research may have diagnostic and therapeutic applications in clinical settings.

Acknowledgments

I have so many people to thank—and those who know me well will not be surprised to hear that I haven't left myself much time in which to do it.

Most of all, I want to thank my adviser extraordinaire, Ken Stevens, who has been a great teacher and a mentor to me since I stopped by MIT almost six years ago to 'shop' one of his classes (as we say at Harvard). Ken has been encouraging and supportive over the years, and his excitement about speech research (especially when he gets his hands on some data) has truly been inspirational, if not downright infectious. I am especially grateful that he was willing to add another graduate student (from Harvard, no less) to his always busy schedule when I was looking for an adviser. On top of all that, he's pretty good about bike repairs, too.

I would also like to thank my committee members: Eva Holmberg has made an invaluable contribution to this thesis, not only by helping with the aerodynamic measures, but by her *very* thorough reading of the thesis and many many valuable comments, suggestions, and corrections. I also thank her for her enthusiasm about and interest in my work—it has been very much appreciated. Barbara Grosz has been very encouraging and I thank her for all her help. And Alan Yuille was there when I needed him, a week before the defense—I can't thank him enough for stepping in at the last minute.

Thanks are also due to my first adviser, Petros Maragos. I have always felt that he gave me my start in graduate school because he was so patient and believing when I was struggling through that first semester and feeling like I would never make it. I don't think that I would be where I am today without his support, and I want to thank him for that.

Next I would like to thank the 22 women who were my subjects, especially the three who let me have a fiberscope threaded through their nasal cavities. Their patience and willingness is much appreciated. In addition, I thank the members of the speech lab who took part in several listening tests and more than several pilot listening tests—they are real troopers.

The members of the Speech Communication Group at MIT deserve special thanks for welcoming me (I think?) to the group and making me feel a part of it. It really is a special place to work. I especially thank Arlene Wint for all her help, for her patience with my incessant procrastination, and, yes, for laughing at me when I run around trying to get everything done at the last minute. Thanks to Stefanie Shattuck-Hufnagel for always brimming with enthusiasm and seeing the best in people, and just generally making us feel good about ourselves and our work. Melanie Matthies has been very patient with my many questions about statistics, multivariate analysis of variance, Bonferroni-Dunn corrections, etc. Plus, she's pretty good with computer-related stuff, and I thank her for all her help. Thanks to Seth Hall for solving many computer problems, especially on Sunday mornings when only God knows why he was reading his email. Sharon Manuel has also been very helpful, especially with improving my listening tests and, more importantly, participating in my listening tests. Thank you also for getting that Router going—it's been invaluable.

Joyce Manzella has been a wonderful friend and made coming to work fun. Thanks for letting me be the crabby person that I really am! Alice Turk helped out alot with designing listening tests and doing pilot tests and reading chapters and making comments and suggestions and all kinds of other things. She is also just a great person to be around and I'm really glad she's my ole buddy ole pal. Jane Wozniak has also been a reliable source of support—she's always willing to be a subject, even for *long, boring* listening tests, she gives us candy and bakes yummy cookies and cakes, and most importantly, she's a true friend.

Marilyn Chen has also been very helpful over the years, especially the night before my defense! Corine Bickley was also very helpful that night, on top of being a good subject, and I thank her for that. Thanks to Xu Yi for his help with understanding glottal anatomy, computers, statistics, and listening tests. And thank you to Kelly Poort, Mario Svirsky, Reiner Wilhelms-Tricarico, and Joe Perkell for their help and support.

Thank you, Keith North, wherever you are, for keeping the lab and equipment running smoothly, and for just being Keith. I really missed seeing you shake your head at me during that last leg of the thesis when I was driving myself crazy doing the procrastination thing and working late at night.

I would also like to thank members of the Robotics Lab at Harvard for their help and friendship, especially the Tea Club: Navin Saxena, Dan Friedman, Peter Bellehumeur, and Mike Wiseman. Jon Christensen, who is only in the Computer Science Dept., but I still let him be my friend, has also been a great pal, and importantly, he's been making me laugh since we met.

Finally, I would like to thank my parents, Sue and Joe Hanson, and my grandmother, Margaret Bottari, for their love and support through it all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis has its roots in a combined interest in speaker characteristics in general and speech characteristics of females in particular. Speaker characteristics give a voice its quality and individuality, and are the characteristics that listeners use to identify or distinguish speakers. Most people often have the experience of recognizing an unseen speaker based on just a few words out of the speaker's mouth. This type of recognition occurs on such a routine basis that it might seem unremarkable. In the same way, most people understand immediately what is meant by a 'female voice' or a 'male voice'. As with many aspects of human behavior, this ability to recognize and distinguish voices is effortless, yet the ability to explain and emulate this behavior is elusive.

Speaker characteristics are complex, having contributions from many levels of the speech production process. These contributions range from the level of the speech production mechanism, that is, differences between individual sound sources and the natural frequencies of the vocal tract, to higher levels, such as prosody and dialect. Part of the challenge in the study of speaker characteristics is to separate these many influences on the speech signal, and at the same time to understand their interaction. Applications of speaker characteristics are several, including speaker verification and identification, speech recognition, and speech synthesis by computer. Speaker characteristics also have applications for speech disorders and therapy, and in the field of forensics.

The distinction between male and female speech has long been believed to be due simply to anatomical differences that lead to females having higher fundamental frequencies of the voicing source and higher natural frequencies of the vocal tract. Yet most efforts to apply the same analysis tools and speech-related applications to both male and female speech have seen better results for male speech. As a consequence, female speech has been considered 'difficult' to analyze and a hard problem. A particular problem has been synthesized female speech, which continues to sound more unnatural than synthesized male speech. Only recently have researchers begun to concentrate on studies that either compare male and female speech, or focus entirely on female speech.

At the level of speech production, fundamental frequency has often been considered to be the primary descriptor of both speaker and gender characteristics. Second to fundamental frequency are the natural frequencies of the vocal tract, or formants. Yet the performance of many speech-related applications based on these descriptors is far from satisfactory. For example, speaker recognition systems may not perform well unless speech is collected under strict conditions and the system is limited to a relatively small number of speakers. And as mentioned above, the synthesis of natural-sounding female speech is difficult to achieve, and thus the synthesis of a particular female's speech must also be considered unrealistic.

However, there are other sources of individuality in speech production. Our goal in the current research was to explore some of these other sources and to determine how they might contribute to speaker characteristics. The basic dimensions of speech production that can lead to individuality include the following:

Sound sources These sources include the voicing source at the vocal folds and turbulence sources that can occur along the length of the vocal tract. Many aspects of these sound sources, such as the intensity and spectrum, can vary from speaker to speaker.

Supraglottal filtering Supraglottal filters include both the oral and nasal cavities. The natural frequencies of these cavities can vary from one individual to another.

Subglottal coupling When coupling between the sub- and supraglottal cavities occurs, vowel spectra can be affected. This type of coupling is expected particularly for speakers who do not close the glottis completely during phonation, and the degree of coupling may vary from speaker to speaker.

Source-filter interaction Contrary to the simple source-filter theory of speech production, interaction may occur between the voicing source and the vocal tract filter. The amount of this interaction can vary from one individual to another.

Kinematics The movements of the articulators, including the vocal folds, can vary from person to person.

Aside from the fundamental frequency and the natural frequencies of the oral cavity, the above sources of individuality provide what might be considered fine details; that is, they may be suspected of having little to add compared to the major influences of pitch, formants, and prosody. However, we will show that for at least one item on this list, these details are not overwhelmed by stronger influences. In particular, we will study the voicing source of female speakers. The aim of the research is to develop a set of descriptors of the voicing source that reflect individual differences in the voice qualities of female speakers. These descriptors go beyond the fundamental frequency to include other details of the voicing source. The descriptors are derived from measurements of the spectra of vowels produced by the speakers. We will show that these details of the voicing source are perceptible and affect the perceived quality of the female voice.

The outline of the thesis is as follows. In Chapter 2 we describe the voicing source in more detail and define the parameters of the glottal waveform, or the glottal characteristics. We also discuss current methods of measuring these characteristics and give a literature review of other efforts to describe the characteristics of the voicing source as it varies across speakers. It will be seen that several researchers have used a method called inverse filtering to obtain glottal waveforms from which to measure glottal characteristics. However inverse filtering is not without its difficulties and requires special equipment. A few researchers have begun to develop methods for measuring glottal parameters from the speech waveform or spectrum, but there has been little work in this area.

We begin Chapter 3 by developing the theoretical background necessary for measuring glottal characteristics directly from the speech waveform and spectrum. Several measures are suggested and the theoretical background is used to predict ranges of values of these of coupling may vary from speaker to speaker.

Source-filter interaction Contrary to the simple source-filter theory of speech production, interaction may occur between the voicing source and the vocal tract filter. The amount of this interaction can vary from one individual to another.

Kinematics The movements of the articulators, including the vocal folds, can vary from person to person.

Aside from the fundamental frequency and the natural frequencies of the oral cavity, the above sources of individuality provide what might be considered fine details; that is, they may be suspected of having little to add compared to the major influences of pitch, formants, and prosody. However, we will show that for at least one item on this list, these details are not overwhelmed by stronger influences. In particular, we will study the voicing source of female speakers. The aim of the research is to develop a set of descriptors of the voicing source that reflect individual differences in the voice qualities of female speakers. These descriptors go beyond the fundamental frequency to include other details of the voicing source. The descriptors are derived from measurements of the spectra of vowels produced by the speakers. We will show that these details of the voicing source are perceptible and affect the perceived quality of the female voice.

The outline of the thesis is as follows. In Chapter 2 we describe the voicing source in more detail and define the parameters of the glottal waveform, or the glottal characteristics. We also discuss current methods of measuring these characteristics and give a literature review of other efforts to describe the characteristics of the voicing source as it varies across speakers. It will be seen that several researchers have used a method called inverse filtering to obtain glottal waveforms from which to measure glottal characteristics. However inverse filtering is not without its difficulties and requires special equipment. A few researchers have begun to develop methods for measuring glottal parameters from the speech waveform or spectrum, but there has been little work in this area.

We begin Chapter 3 by developing the theoretical background necessary for measuring glottal characteristics directly from the speech waveform and spectrum. Several measures are suggested and the theoretical background is used to predict ranges of values of these measures for several glottal configurations that might be expected for female speakers in normal voice. Following the theoretical background, we describe an experiment in which the speech of 22 female speakers was analyzed using the proposed measures. The results are found to fall into the predicted ranges and are used to classify the speakers according to hypothesized glottal configurations.

In Chapter 4 we attempt to further explore and perhaps validate these results through physiological measurements on four subjects. These include the measurement of the glottal waveform through inverse filtering and the extraction of glottal characteristics from that waveform. We also describe visual observations of the vocal folds during phonation. The results of these two experiments support the hypotheses made in Chapter 3.

Chapter 5 describes listening tests designed to determine the effect of the variation of glottal characteristics on voice quality perception. We find a strong correlation between several of our spectrum-based measures and the perception of a breathy voice quality in our group of female speakers. In addition, a test using synthesized stimuli for which only the glottal parameters are varied shows that the variation of these parameters is perceptible to listeners. We also find that in order to generate synthesized vowels that are similar to naturally-produced vowels by different speakers, it is necessary to select glottal parameters that differ from one speaker to another. The acoustic measures developed in Chapter 3 are found to have potential for guiding the improved synthesis of voice quality.

Finally, Chapter 6 summarizes our work and discusses future work.

The contributions of this work are several. First, it adds to research efforts aimed at finding quantitative measures that describe dimensions along which normal voices vary across speakers. It also improves the understanding, analysis, and synthesis of female voice and speech. In addition, the work may also have medical applications, including the diagnosis and classification of voice and speech disorders, and speech or voice therapy.

Chapter 2

Background and literature review

2.1 Introduction

This chapter provides the background necessary for the remainder of this thesis. We begin with a brief description of the voicing source at the vocal folds, and of how airflow is modulated to produce the glottal waveform that excites the vocal tract during the production of voiced sounds. Next, we define attributes of the glottal waveform that are of interest for (1) what they reflect about glottal configurations and (2) how a given configuration affects voice quality. This section also discusses ways in which the glottal configurations of female speakers differ from those of males, indicating the importance of a separate consideration of male and female glottal characteristics. In the next section, we turn to techniques for measuring the glottal waveform and viewing glottal activity during phonation. The last section is a review of research related to the work to be presented in Chapters 3–5 of this thesis.

2.2 Glottal characteristics

Speech is commonly thought of as the product of sound sources that are filtered by the vocal tract. For the purposes of this thesis, we are interested in the voicing source at the vocal folds. Figure 2.1 is a schematic of the vocal folds and surrounding structures as viewed from above. The upper part of each panel of the figure represents the posterior

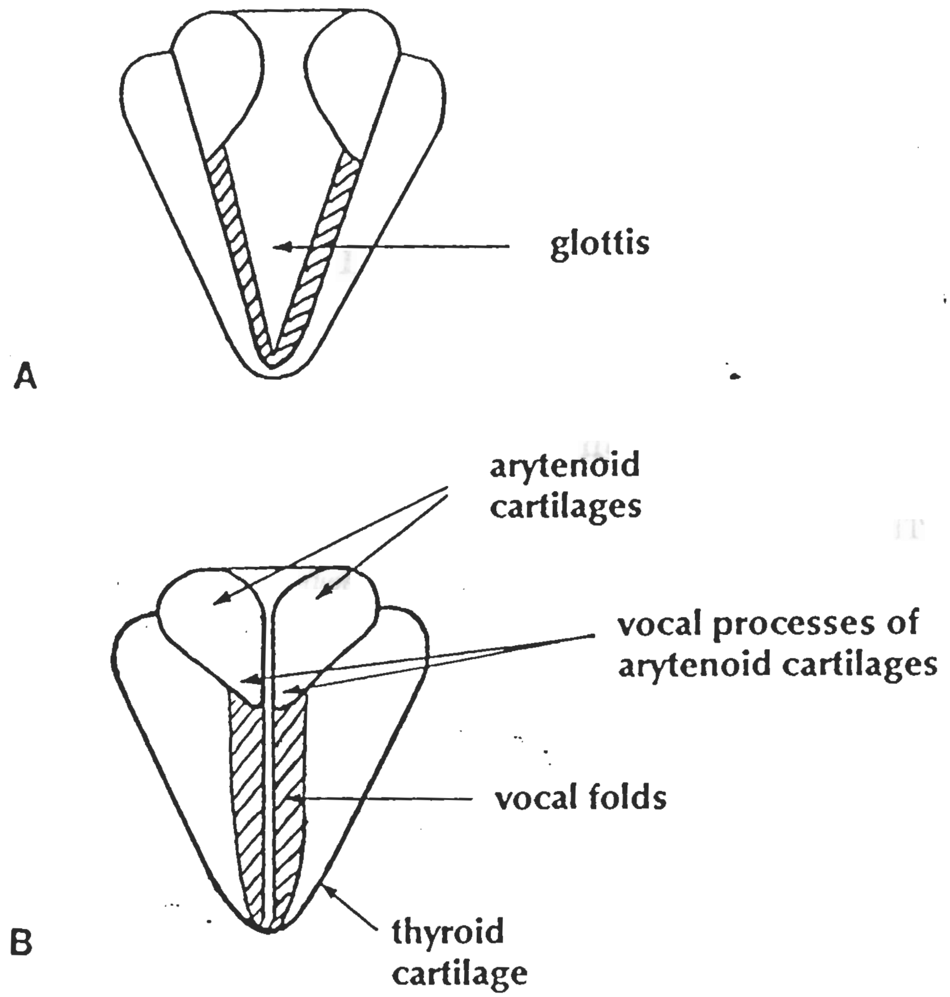


Figure 2.1: *The vocal folds, as viewed from above. (a) Configuration during quiet breathing. (b) Configuration during vocal fold vibration (from Stevens, 1994).*

end of the vocal folds. The membranous part of the folds runs from the anterior end of the vocal folds to the vocal processes, which connect them to the arytenoid cartilages. The space between the folds is referred to as the glottis. Through movements of the vocal folds, the cross-sectional area of the glottis is periodically modulated, shaping the airflow that passes through. Figure 2.2(a) shows lateral sections of the vocal folds at various points in time during one cycle of vibration. The lungs act as a constant pressure source below the folds. In the first panel the folds are closed along their entire vertical length. The transglottal pressure builds up until the folds begin to separate at their lower edge, as in the second panel. The separation propagates along the vertical length until finally the folds are completely separated as in the third panel. The fourth panel shows the folds when they are most widely separated; at this point the pressure within the glottis is small. In the fifth panel the lower edges of the folds come back together, which will be followed by the upper edges. The cycle begins anew back at the first panel. Figure 2.2(b) shows a schematic drawing of a glottal waveform. It is labelled to show which points in the waveform correspond to the panels of Figure 2.2(a).

As with any mechanical system, the voicing source has a natural frequency, usually referred to as the fundamental frequency, or F0. This frequency is relatively low, with a range of about 100–300 Hz for adult speakers. However, for speech it is necessary to produce sound energy over a broad frequency range, up to about 5000 Hz. The key to producing such sound energy is that the folds close rapidly, resulting in an abrupt cessation of the airflow. In this way acoustic energy is produced over a wide range of frequencies (Stevens, in preparation). This abrupt cessation in the airflow can be seen in Fig. 2.2(b) at about 4.5 ms.

The waveform of Fig. 2.2(b) is associated with what is called modal phonation. In this type of phonation, complete closure occurs simultaneously along the length of the folds. Note that for modal phonation the flow is zero during the time that the glottis is closed along its vertical length, and the cutoff of airflow at closure is abrupt.

However, it is often the case, especially for female speakers, that there is incomplete closure of the vocal folds during phonation; that is, for many speakers there is always some opening at the glottis during the phonatory cycle (see, for example, Hertegård et al.,

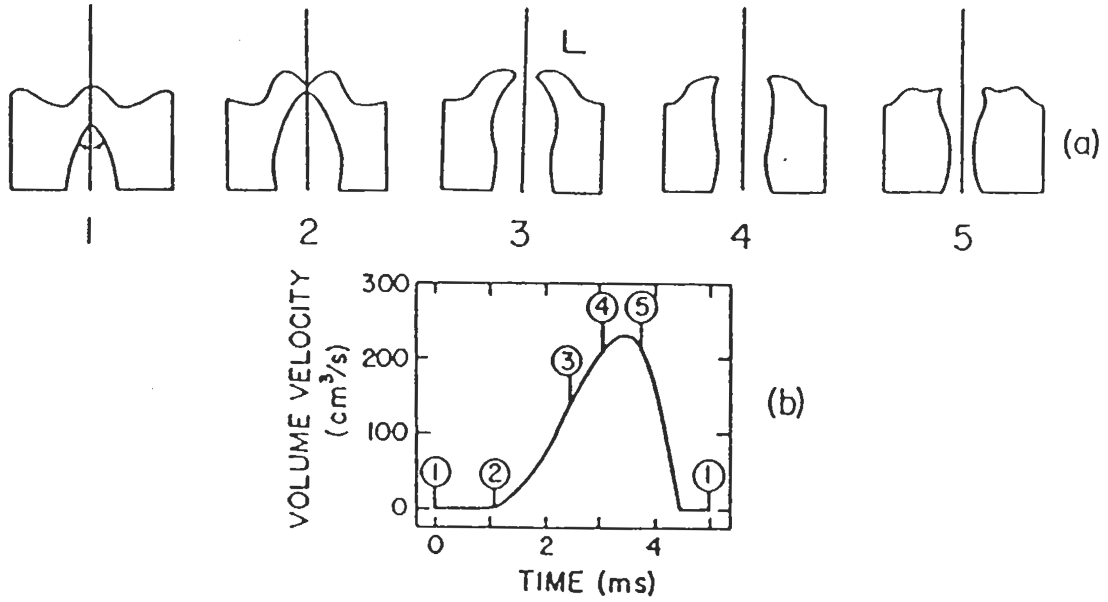


Figure 2.2: *(a) Schematized lateral sections of the vocal folds at various times during a vibratory cycle. The folds are closed in panels 1 and 2, and open in panels 3–5. See text for additional explanation. (b) Schematized glottal waveform labelled to indicate which points in the waveform correspond to the panels in (a) (from Stevens, 1994).*

1992; Linville, 1992; Peppard et al., 1988; Södersten and Lindestad, 1990; Södersten et al., 1991). This opening is often at the arytenoid cartilages at the posterior end of the folds, and in this case is often referred to as a *glottal* or *posterior chink*. Other types of openings can occur, for example at the anterior end of the folds or at the center of the folds. More generally, an opening at the glottis is referred to as a *glottal gap*, or sometimes as a *fixed opening*. Figure 2.3 shows schematic drawings of possible posterior openings (chinks) that may occur during the closed phase of a glottal cycle. We see that the size of these posterior openings can vary, and they may extend beyond the vocal processes to the membranous part of the folds.

The preponderance of such openings among female speakers has lead several researchers to suggest that they should be considered normal for women (Biever and Bless, 1989; Södersten and Lindestad, 1990; Rammage et al., 1992). Hirano et al. (1988) offered explanations based on anatomical differences between males and females. Södersten and Hammarberg (1993) and Hertegård et al. (1992) found that these openings occur even in women with trained voices. Rammage et al. (1992) found that the size of posterior chinks was not significantly reduced for female patients following voice therapy. Such findings support the hypothesis that these kinds of gaps are due to anatomy (Södersten and Hammarberg, 1993).

The implications of incomplete closure for the glottal waveform $U_g(t)$ are that there is an airflow bypass even during the so-called closed phase of the glottal vibratory cycle, and that abrupt cutoff of the airflow is not possible due to the mass of air in this pathway. A synthesized glottal waveform illustrating what might result when there is incomplete glottal closure is shown in Fig. 2.4(a). This waveform has a DC offset, due to the airflow bypass at the vocal folds. This DC offset is often referred to as *DC flow*, *minimum flow*, or *residual flow*. In clinical literature it is also referred to as "unmodulated flow". The waveform at the time of closure is smoothed compared to that of Fig. 2.2(b). The latter is more obvious in the derivative of the glottal waveform $dU_g(t)/dt$, illustrated in Fig. 2.4(b): if closure had occurred abruptly, the derivative would change abruptly from a negative value to zero, as shown by the dotted line. But as we can see, for the case of nonabrupt closure this change occurs more gradually, thereby reducing the high-frequency

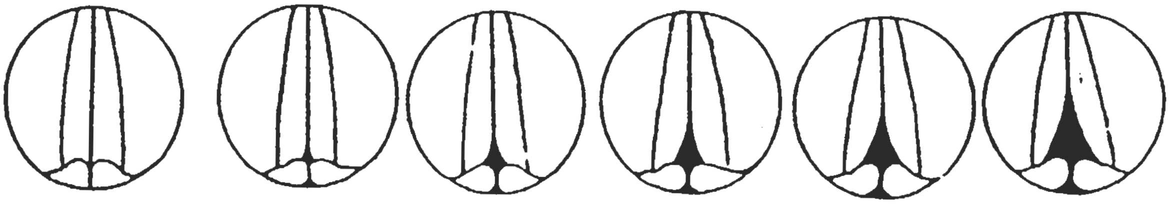


Figure 2.3: *Schematics indicating various glottal configurations that might occur during the closed phase of a glottal cycle. The posterior ends of the folds are at the bottom of the figures. Posterior openings can range in size from being nonexistent, as in the first panel, to extending beyond the vocal processes into the membranous part of the folds, as in the last panel (from Rammage et al., 1992).*

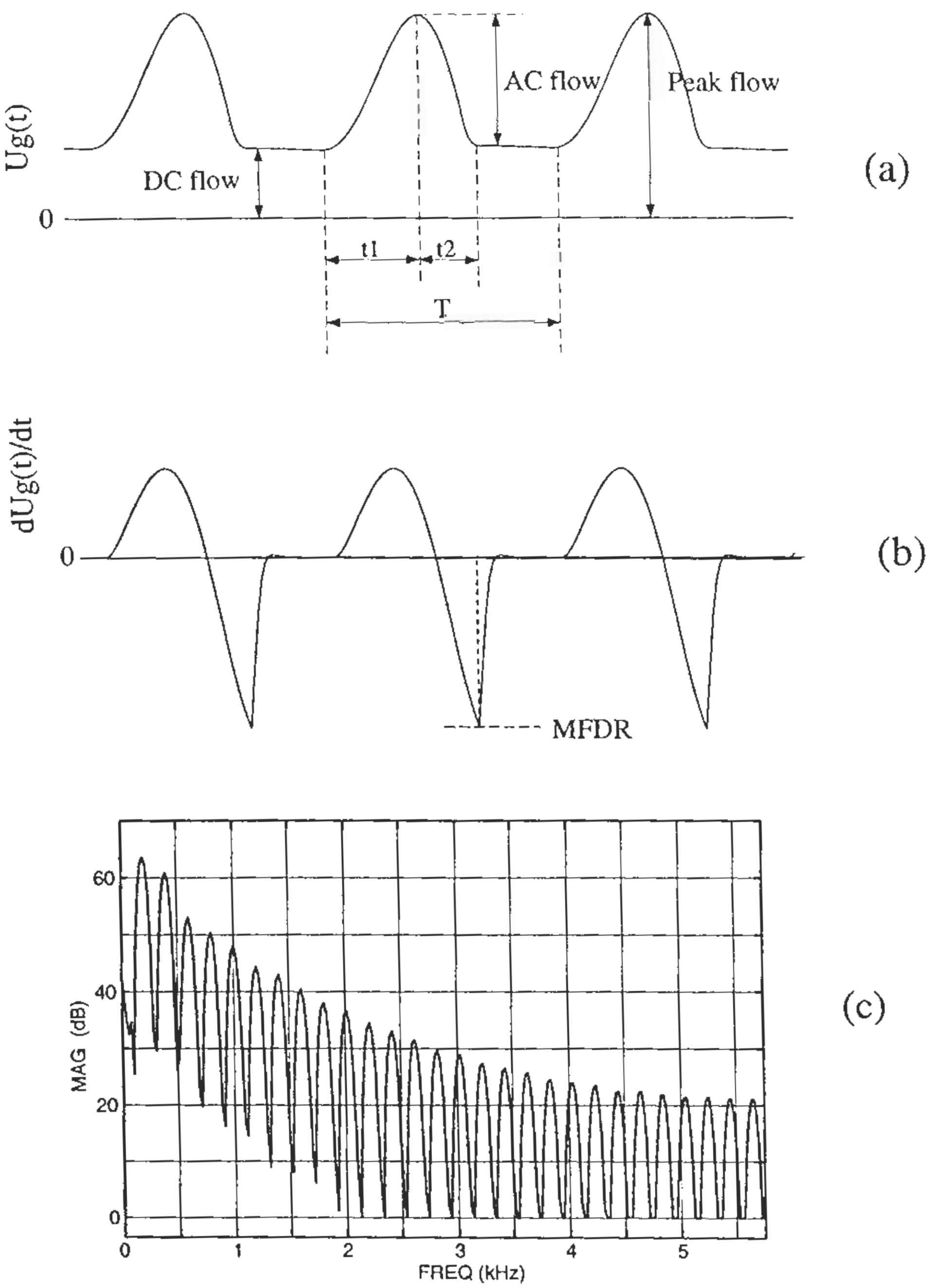


Figure 2.4: *Schematic of a glottal waveform $U_g(t)$, and its derivative dU_g/dt, synthesized using the KLSYN88 formant synthesizer (Klatt and Klatt, 1990). (a) The glottal waveform $U_g(t)$. The glottal parameters AC flow, DC flow, peak flow, and the pitch period T are indicated. Speed quotient is defined as $t1/t2$ (ratio of rise time to fall time), and open quotient is defined as $(t1 + t2)/T$ (ratio of open time to pitch period). (b) The derivative of the glottal waveform dU_g/dt. MFDR is indicated. The vertical dotted line indicates how the derivative would appear if abrupt closure had occurred. (c) Spectrum of the waveform in (b).*

content of the glottal waveform. Due to the radiation characteristic at the mouth, which can be approximated as a derivative, the derivative of the glottal waveform is the effective excitation (Fant, 1982), and thus can be of more interest than the glottal airflow itself. The spectrum of the derivative is also of interest because according to the simple source-filter theory, the output speech spectrum is the product of the source spectrum, the frequency response of the vocal tract, and the radiation characteristic. Figure 2.4(c) shows the spectrum of the derivative in (b).

There are other parameters of interest besides the DC flow and the abruptness of closure, and these are illustrated in Fig. 2.4 as well. These parameters have been found to be important in terms of the intensity (SPL) and quality of voice. The *pitch period*, or 1/F0, is indicated by T, and the *rise time* and *fall time* of the glottal waveform are indicated by t1 and t2, respectively. One of the most important parameters for SPL is the *maximum flow declination rate*, or, more compactly, the MFDR. This parameter is the greatest negative slope that occurs during the fall time, and it reflects how rapidly the folds are closing. More importantly, as seen in the derivative $dU_g(t)/dt$, MFDR represents the point of greatest excitation of the vocal tract. The *AC flow* represents the component of the flow that is modulated by the vibrating vocal folds, and it becomes larger when the amplitude of vibration increases. The *peak flow* is simply the sum of the DC and AC flows, and represents the maximum airflow through the glottis. Another parameter of interest for perceived voice quality is the *open quotient*, the ratio of time that the vocal folds are open to the pitch period, T, or $(t1 + t2)/T$. Finally, the *speed quotient*, or *skewness* of the waveform, is measured as the ratio of the rise time to the fall time, or $t1/t2$.

The values for the parameters just defined can vary depending on the glottal configuration, and it is expected that these variations may lead to different voice qualities and intensities. Some voice qualities are usually associated with disordered voice, such as harshness, but our main concern for this thesis are those that occur for voices that are not considered to be disordered. One mode of vibration that can occur in normal speech is *pressed voice* or *laryngealization*. This phonation is characterized by a reduced open quotient and a reduction in airflow through the glottis, compared to what might be expected for modal phonation.

Breathy voice occurs when there is a greater amount of air passing through the glottis than might be expected for modal phonation. This large amount of air is thought to be due to an incomplete glottal closure or a gradual closing movement, which also result in a larger open quotient. Breathy voice can result from voice disorders, but also occurs in voices that are not disordered. In fact, some languages use breathy voice phonemically (see, for example, Ladefoged and Antoñanzas-Barroso, 1985; Huffman, 1987). Hammarberg et al. (1984) found that there were two types of phonation that lead to breathy voice. A breathy/hypofunctional voice is produced with low laryngeal effort, and the waveform is expected to be almost sinusoidal, with increased DC flow. A breathy/hyperfunctional voice is produced with a great aerodynamic and laryngeal effort, and the glottal waveform is also expected to have increased DC flow, but rather than being sinusoidal, the waveform will be skewed to the right and have increased MFDR and AC flow. Holmberg et al. (1994b) have suggested that this latter type of phonation may be the result of speakers trying to compensate for the increased glottal losses associated with breathy voice.

We will now turn to a description of techniques that are used to measure glottal characteristics.

2.3 Measurement of glottal characteristics

There are several methods that are used to measure or observe characteristics of the glottal configuration or waveform. One such method is *inverse filtering*. This method is based on the simple source-filter theory of speech production, which says that a source of sound is filtered linearly by the vocal tract to produce speech. Therefore, if the effects of the formants are removed from a speech waveform with a filter that is the inverse of the vocal-tract filter, the resulting signal will be the glottal waveform. Glottal parameters can then be extracted from this waveform and its derivative. Inverse filtering can be done on the acoustic sound pressure or the oral airflow. For the first method, speech is recorded using a microphone. For the second method oral airflow is usually measured during speech production via a Rothenberg mask (Rothenberg, 1973), pictured in Fig. 2.5, which is a high-time resolution pneumotachograph.

Both methods of inverse filtering have advantages and disadvantages. The Rothenberg

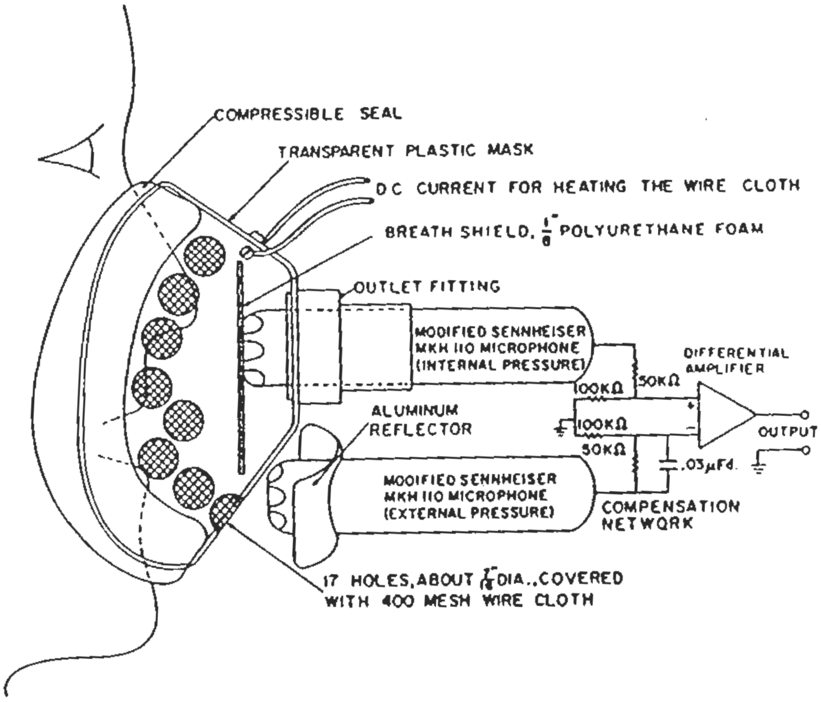

COMPRESSIBLE SEAL
TRANSPARENT PLASTIC MASK
DC CURRENT FOR HEATING THE WIRE CLOTH
BREATH SHIELD, 1/8" POLYURETHANE FOAM
OUTLET FITTING
MODIFIED SENNHEISER MKH 110 MICROPHONE (INTERNAL PRESSURE)
DIFFERENTIAL AMPLIFIER
50KΩ
100KΩ
100KΩ
50KΩ
+
−
OUTPUT
.03μFd.
ALUMINUM REFLECTOR
MODIFIED SENNHEISER MKH 110 MICROPHONE (EXTERNAL PRESSURE)
COMPENSATION NETWORK
17 HOLES, ABOUT 1/16" DIA., COVERED WITH 400 MESH WIRE CLOTH

FIG. 2. Circumferentially vented pneumotachograph mask.

Figure 2.5: *Rothenberg mask (from Rothenberg, 1973).*

mask preserves the zero flow level, but has a limited bandwidth, the frequency response being flat only up to about 1200 Hz. Consequently the airflow must be lowpass filtered at about this frequency, and the higher part of the source spectrum cannot be studied. Most importantly, information about the abruptness of glottal closure is lost. Another disadvantage of the mask is that unless a tight fit is maintained between the mask and the subject's face, mask leak will occur and the measures of glottal flow amplitude will be erroneously reduced. Mask leak is difficult to detect unless the glottal waveform falls below the zero baseline. Microphone recordings are much easier to collect and mid- to high frequencies of the glottal waveform are preserved. However, these recordings must be obtained under very strict conditions, using phase-true microphones (Karlsson, 1988, 1992b). In addition, the inverse-filtered microphone recordings are very sensitive to low-frequency noise and do not preserve the zero flow level. Consequently, the absolute transglottal airflow cannot be measured from these recordings (Hertegård et al., 1992).

There have been other attempts to measure oral airflow that do not involve the Rothenberg mask. Cranen and Boves (1985, 1988) have used pressure transducers to estimate the pressure gradient across the glottis, which can then be used to estimate the oral airflow. This method measures the flow close to the glottis, whereas the Rothenberg mask method measures the flow at the mouth. However, the DC component is not preserved. In other work, Teager (1980) has used hot-wire anemometers to measure oral airflow (see also, Teager and Teager, 1983a, 1983b).

Most researchers have extracted parameters directly from the glottal waveform that results from inverse filtering, or from a glottal waveform model that is fit to the natural glottal waveform. However, there have also been attempts to relate these parameters to the spectrum of the glottal waveform, or to the speech waveform and spectrum. In particular, Fant (Fant, 1979, 1993; Fant et al., 1985, 1994; Fant and Lin, 1988) and Ananthapadmanabha (1984) have pioneered in developing these techniques.

A second method for studying the glottis during phonation is visual observation of the vocal folds, using either an endoscopic system or a fiberscopic system. These two systems are illustrated in Fig. 2.6. In the first method a rigid endoscope is inserted into the oral cavity and positioned so that the vocal folds can be observed during phonation, as

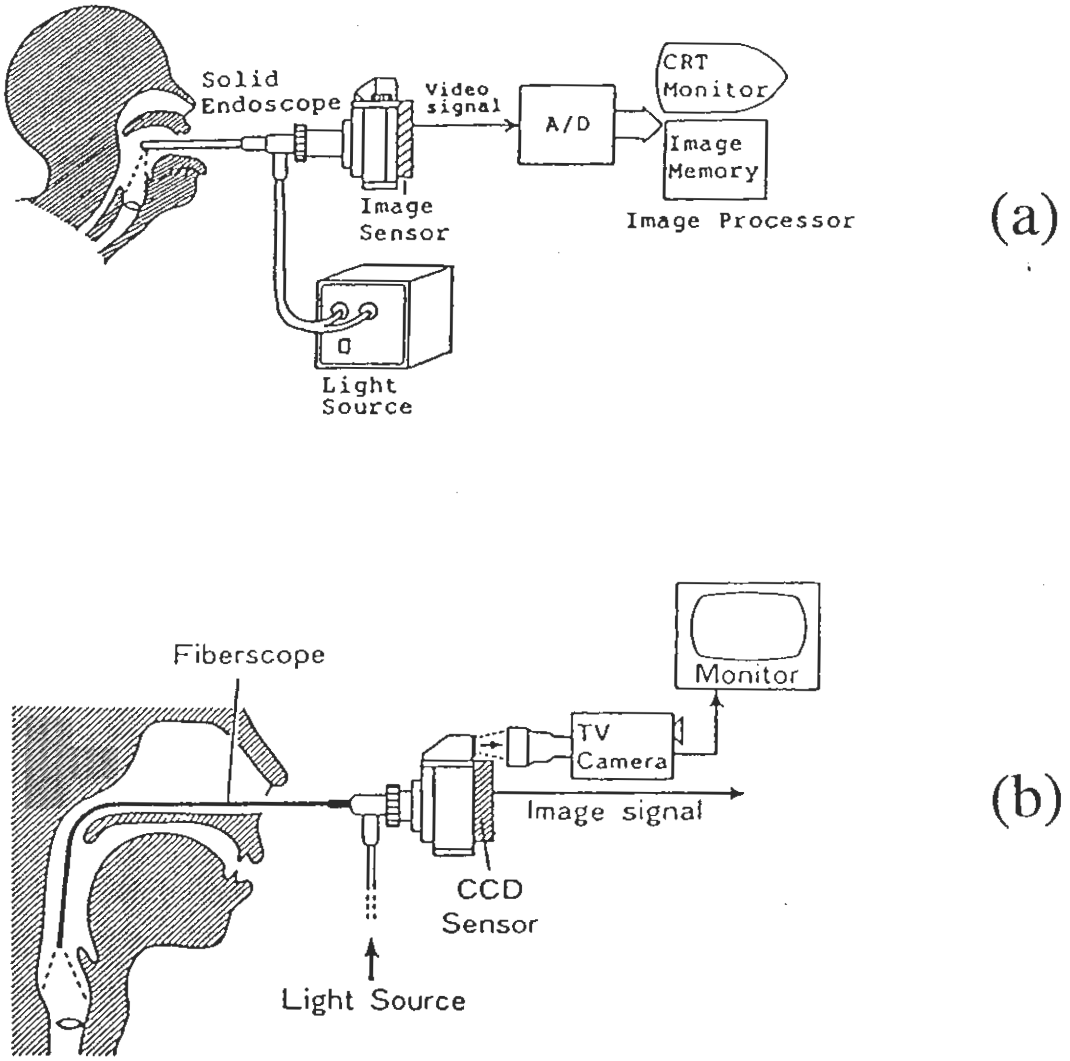


Figure 2.6: *(a) Schematic of an endoscope system. (b) Schematic of a fiberscope system (both figures from Kiritani et al., 1990).*

illustrated in Fig. 2.6(a). Speech materials are necessarily limited to open vowels. In the second method, a flexible fiberscope is inserted through the nasal cavity and positioned above the vocal folds so that the folds can be observed during phonation, as illustrated in Fig. 2.6(b). With this system, there are more choices of speech materials available. However, the image quality is not as good as with the rigid endoscope system.

2.4 Related work

Previous work that has studied individual variations in glottal characteristics tends to fall into the following categories:

- Glottal characteristics during phonation are measured from glottal waveforms obtained by inverse filtering, or are based on visual inspections of the vocal folds.
- Acoustic measures of glottal characteristics are made directly on the speech spectrum or waveform. Examples of such measures are measures of perturbation, such as jitter and shimmer. Another measure that is often made is the level of the first harmonic in the speech spectrum relative to the level of the second harmonic, because several studies have found this measure to be correlated with degree of perceived breathiness and open quotient (for a review, see Klatt and Klatt, 1990).
- Perceptions of voice quality, particularly breathiness, are gathered from listeners, and related to the glottal characteristics as measured from the glottal waveform or from the speech spectrum.

Holmberg and her colleagues have done extensive studies, making aerodynamic measures of both male and female voice (Holmberg et al., 1988, 1989, 1994a, 1994b, in press; Perkell, et al., 1994), and extracting glottal parameters directly from the glottal waveform. Their goals in this work were to gain an improved understanding of normal voice production and the determination of normal ranges of glottal characteristics (Holmberg et al., in press) to be used in studies of voice disorders. These studies included relatively large groups of subjects (from 20 to 45) phonating in different speech conditions, including soft, normal, and loud speech, and low, medium, and high pitch. The main findings in their studies are the existence of parameter differences between females and males, and across speech conditions. Several parameters were found to be significantly related to SPL. For normal loudness and pitch, they found that female speakers had lower peak flow, AC flow, and maximum flow declination rate (MFDR) relative to male speakers, and that the females had more gradual opening or closing times, leading to less well-defined closed portions of the glottal waveform (Holmberg et al., 1988). Several of the aerodynamic measures were found to be well correlated, indicating that there are complex relationships between these measures. Their studies also included acoustic measures made on the speech spectrum, and they related these spectral measures to the aerodynamic results (Holmberg et al., in press). They found that open quotient, as measured from the glottal waveform, has a strong relationship with the difference in amplitude of the first two harmonics. In addition, MFDR is in some cases reflected by the difference in amplitude of the first and third formants.

Karlsson (1986, 1988, 1990, 1991a,b, 1992a,b) also studied the glottal waveform obtained by inverse filtering, but she focused on a small group of female speakers. Her aim was to achieve descriptors of voice differences that can be used to synthesize speech for individual voices (Karlsson, 1992a). Her methods included inverse filtering of both oral airflow and speech. Glottal parameters were obtained by fitting a theoretical model to the resulting glottal waveform. She found that most female speakers have a DC offset in their glottal airflow, suggesting incomplete glottal closure. She also attempted to correlate the glottal parameters with the voice qualities of her subjects, as judged by speech therapists. In these tests, she found that different voice qualities were separated by the degree of spectral tilt of the voice source, and the presence of noise excitation at mid- to high frequencies (Karlsson, 1988, 1989). Speakers perceived to be breathy were found to have higher minimum flows, steeper tilts, and more aspiration noise (Karlsson, 1988, 1992). Voices with a tight, strained quality had weaker lower harmonics (Karlsson, 1992b).

Södersten and her colleagues studied glottal closure via fiberscopy and related the degree of closure to perceptions of breathiness. They also used an acoustic measure, the level of the fundamental relative to the level of the first formant, that they refer to as L0-L1. Their results show that female speakers have a higher degree of incomplete closure and perceived breathiness than male speakers (Södersten and Lindestad, 1990), and significant correlations between breathiness and the acoustic measure L0-L1 (Södersten, Lindestad, and Hammarberg, 1991). They also found that even after voice training, incomplete glottal closure, while reduced, still existed in female subjects (Södersten and Hammarberg, 1993), supporting a hypothesis that incomplete closure in females is due to anatomy rather than behavior.

Gobl (1989) studied voice quality correlates by extracting glottal parameters from the inverse-filtered waveform and making measurements on the glottal spectrum. These acoustic measures were the average of the harmonic amplitudes in four frequency bands. Breathy voice was found to have a steeper spectral tilt than modal voice. Gobl and Ní Chasaide (1988) and Ní Chasaide and Gobl (1993) extracted glottal parameters both from the glottal waveform and from vowel spectra. They found that as glottal abduction increased, so did the downward spectral slope of the vowel spectrum, and that formant amplitudes, especially F1, decreased as well.

Klatt and Klatt (1990) studied variations in voice quality, from pressed to breathy, for both male and female speakers, using reiterant and synthesized speech. Through perception tests they too found that female speakers were perceived to be breathier than male speakers. Unlike the researchers discussed above, they relied entirely on acoustic measures made directly on the speech spectrum and waveform. They found that relevant cues to perception of breathy voice are increases in the amplitude of the fundamental component, aspiration noise, and lower formant bandwidths. Aspiration noise was found to be the most important cue to breathy voice.

Kasuya and Ando (1991), in a study limited to two female subjects, found that an increased amplitude of the fundamental and the amount of glottal turbulence noise were important cues for perceived breathiness.

2.5 Summary

We have seen in this chapter that the glottal waveform has several parameters that can vary among speakers, and that some these variations may affect perceived voice quality. Several researchers have studied these variations using inverse filtering to obtain the glottal waveform, or fiberscopic and endoscopic examination of the vocal folds. Some of these variations have been related to perceptions of voice quality. However, inverse filtering has several problems associated with it, including the difficulty of gathering airflow or speech recordings, and errors that can be introduced due to improper choice of the inverse filter. Examination of the vocal folds is necessarily invasive. Thus, there is a need to develop methods of measuring glottal characteristics directly from the acoustic sound pressure which do not require special equipment to record. In the next chapter we will discuss the theoretical basis for such measurements, and then apply these measurements to speech recordings gathered from a group of 22 female speakers.

Chapter 3

Acoustic measures of glottal characteristics

3.1 Introduction

In this chapter we discuss methods of measuring glottal characteristics from the sound pressure, without inverse filtering and without direct physiological measurements, and we apply these methods to acoustic data from female speakers. We begin by reviewing theoretical background that can be used as a basis for predicting how differences in glottal configuration are manifested in the sound. This theoretical background draws on previous work, particularly Fant et al. (1985) and Klatt and Klatt (1990). As a result of this theoretical development, several measures of glottal characteristics will be suggested. Following the theoretical development, acoustic data for 22 female speakers will be given, and we will attempt to interpret these data in terms of the theoretical models and to classify individual differences based entirely on the inferences derived from the measurements of the sound pressure.

3.2 Theoretical background

3.2.1 Complete glottal closure during a vibratory cycle

We begin by reviewing the parameters that can influence the glottal waveform for the case in which the glottis closes completely during a part of the glottal cycle. The waveform can show several kinds of differences from one individual to another. Certain of these differences are manifested in the spectrum at low frequencies, in the vicinity of the lowest two or three harmonics, whereas other differences modify the spectrum of the volume-velocity waveform at middle and high frequencies.

For example, if a speaker modifies her production such that it results in a glottal waveform with a larger open quotient but the same rate of decrease of volume velocity at closure, the spectrum of the source undergoes a change only at low frequencies, with essentially no change in the spectrum amplitude at high frequencies (Klatt and Klatt, 1990). At a given fundamental frequency, differences in open quotient could arise from differences in the degree of lateral compression of the vocal folds and the extent to which there is a tapering of the glottis from the inferior to the superior surface. With greater tapering, one might expect that the lower edges remain closed during a shorter time interval, possibly leading to a longer open interval.

Figures 3.1(a) and 3.1(c) show the derivatives of the volume-velocity waveform and the spectra of these derivatives for two synthesized waveforms having different values of open quotient (OQ) (30 percent and 70 percent, respectively). When OQ varies from 30 to 70 percent, the difference between the amplitudes of the first two harmonics ($H1-H2$) changes by about 10 dB. Figs. 3.1(b) and 3.1(d) show spectra for the vowel /æ/ synthesized using these glottal waveforms. The difference between the values of $H1 - H2$ that were observed in the glottal spectra are also evident in the spectra of the synthesized vowels.

The spectrum of the derivative of the glottal waveform at middle and high frequencies, when the derivative has a discontinuity at the time of closing, has a downward slope of 6 dB/octave. This spectrum is influenced by the abruptness with which the flow is cut off when the membranous part of the vocal folds closes during the vibration cycle. As has been shown by Fant et al. (1985) and by Klatt and Klatt (1990), this abruptness can be affected in two ways, for a given open quotient, when there is complete closure of the glottis during some part of the vibratory cycle. One mechanism that leads to a change in abruptness is a glottal closing that does not occur simultaneously at all points along the anterior-posterior length of the vocal folds. Closing is a type of "zipper" action, with initial closure at the anterior end of the glottis and the closure sliding back along the length of the glottis (cf. Ananthapadmanabha, 1993). This type of closure leads to a more gradual cutoff of the flow, as illustrated in Fig. 3.1(e). The glottal volume velocity

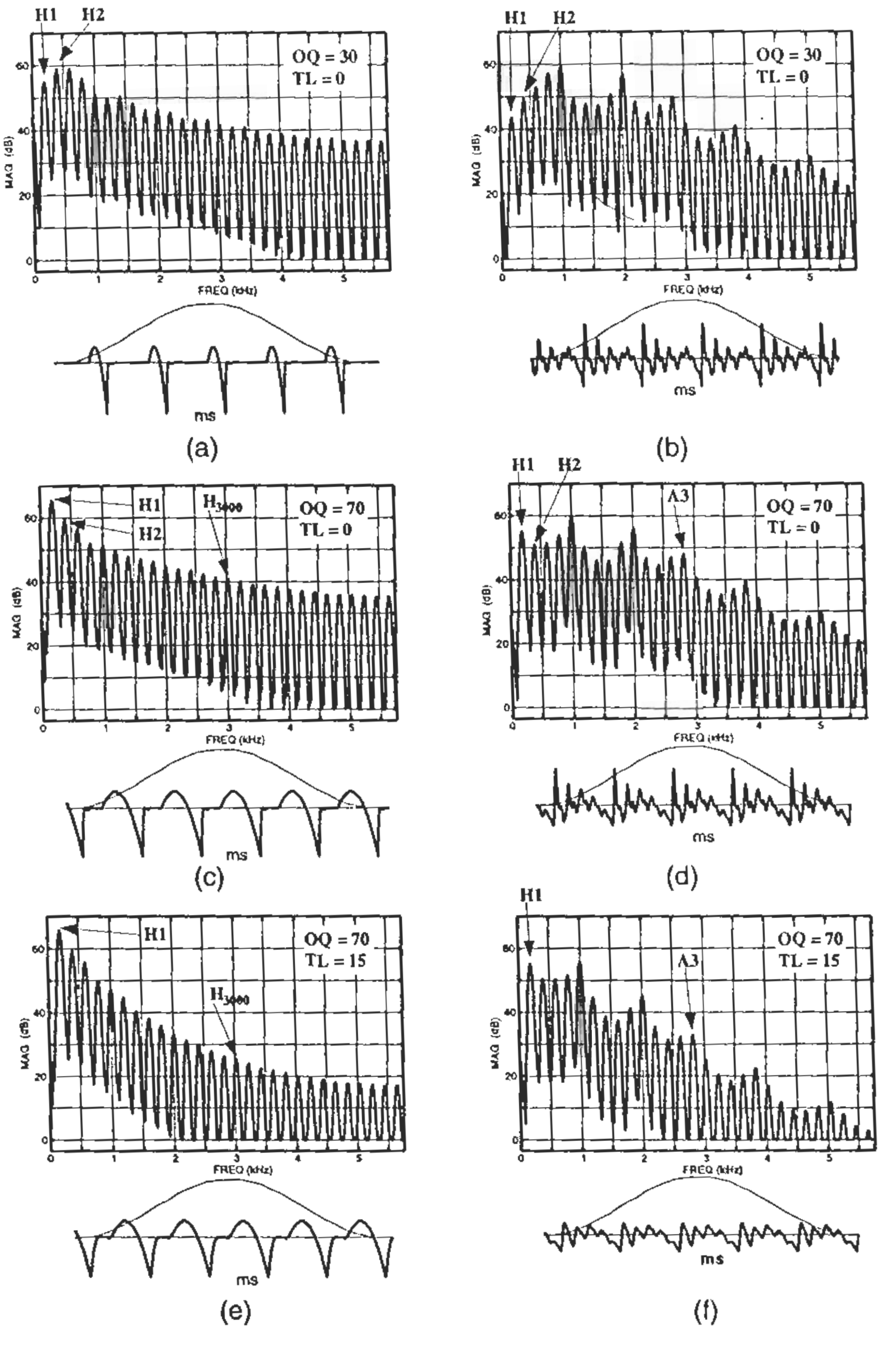


Figure 3.1: *Waveforms and spectra of the periodic glottal volume-velocity source corresponding to various manipulations of the glottis. The fundamental frequency is in the range for an adult female speaker. Panels (a), (c), and (e) show spectra and derivatives of the volume-velocity sources, while panels (b), (d), and (f) show the spectra of the vowel /æ/ synthesized using those volume-velocity sources. (a)-(b) Open quotient (OQ) is 30%, spectral tilt (TL) is zero; (c)-(d) OQ is 70%, TL is 0; (e)-(f) OQ is 70%, TL is 15 dB (i.e., spectrum is 15 dB lower at 3 kHz). These waveforms and spectra were generated by the KLSYN88 synthesizer (Klatt and Klatt, 1990) which contains several glottal sources, including a representation of the source proposed by Fant et al. (1985).*

waveform shows a more gradual downward slope near the time of closure. The effect on the spectrum is to introduce an additional downward tilt in the spectrum at high frequencies. If we define T_D as the time from initiation of the anterior closure to the time of closure at the posterior end, and if we approximate the gradual cutoff as an exponential, then we can say that the time constant T of this exponential is roughly one-half of the time of the sliding closure, i.e.,

$$T \approx \frac{T_D}{2}$$

The breakpoint for the change in spectral slope is then given by

$$f_T = \frac{1}{2\pi T} = \frac{1}{\pi T_D} \tag{3.1}$$

Above this frequency, the slope of the spectrum increases to 12 dB/octave if an exponential approximation is assumed. For f_T less than about 2000 Hz, the resulting increase in the tilt at 2750 Hz, an average location of F3 for female speakers, is

$$20 \log_{10} \frac{2750}{f_T} \tag{3.2}$$

For example, if T_D is 0.5 ms, f_T is 637 Hz and the increase in tilt at 2750 Hz is 13 dB. Likewise, if T_D is 1.0 ms, f_T is 318 Hz and the increase in tilt at 2750 Hz is 19 dB. The glottal waveform of Fig. 3.1(e) is synthesized with a tilt of 15 dB, or an exponential return phase having a time constant T of about 0.65 ms. The spectrum of a vowel synthesized using the glottal waveform of Fig. 3.1(e) is shown in Fig. 3.1(f). Note that the amplitude of the third formant (A3) drops by about 15 dB compared with the spectrum in Fig. 3.1(d). Thus, the value of A3 relative to H1 appears to be a reasonably accurate measure of spectral tilt, except if H1 is weak, as in Fig. 3.1(b). Holmberg et al. (in press) have also used this measurement as an indication of how abruptly airflow is cut off.

The abruptness at closure can also be influenced by the rate of decrease of flow at the instant of closure. For a given open quotient, this rate of closure depends on the amount of skewness of the glottal pulse, as shown in Figs. 3.2(a) and 3.2(b). As the slope of the closing phase becomes faster relative to the slope of the opening phase (keeping OQ about constant), the spectrum amplitude at middle and high frequencies increases relative to the amplitude at low frequencies. In this case, the difference between the amplitudes of the first harmonic (H1) and the harmonic at 3000 Hz (H_{3000}) increases by about 10 dB with the change in speed quotient. Speech spectra corresponding to these glottal waveforms are shown in Figs. 3.2(c) and 3.2(d), and again, the difference in A3 relative to H1 for the two spectra seems to accurately reflect the change in tilt.

The amplitude of the third formant is also influenced by the locations of F1 and F2. Unless a correction is made for this effect (see Appendix A.2), values of A3 compared across vowels or speakers must be interpreted with caution. Another complication in comparing A3 across vowels is that the bandwidth of F3 is influenced by the radiation characteristic to a greater extent than are the lower formants. House and Stevens (1958) measured third formant bandwidths (B3) for male speakers of English, and found B3 of the vowel /æ/ to be 103 Hz and that for /ʌ/ to be 64 Hz. This result predicts that the amplitude of the third formant of /æ/ will be about 4 dB less than that of /ʌ/. Measures of tilt based on A3, then, should also be corrected for this effect when they are to be compared across vowels.

A minimum H1 – A3 value that can be expected due to the 6 dB/octave dropoff in the source spectrum can be estimated using glottal waveforms synthesized with either the LF model (Fant et al., 1985) or the KLGLOTT88 model (Klatt and Klatt, 1990). When the Klatt model is used to synthesize a source waveform with a fundamental frequency of 200 Hz, an open quotient of 50 percent, and no additional tilt, the difference between the amplitudes of the first harmonic and the harmonic at 3000 Hz is about 21 dB. Subtract from this the effect of the third formant, which, given a bandwidth of about 170 Hz for a neutral vocal tract setting (Fant, 1972), is about 12.6 dB, and the resulting H1 – A3 is 7.4 dB. To this we can add 1–2 dB because F3 will, on average, not be centered exactly on a harmonic, bringing the minimum H1 – A3 measure to about 9 dB.

In summary, when the glottis is adjusted so that there is complete closure during the closed phase, one might expect that different individuals exhibit glottal waveforms that differ either in the low-frequency region or in the high-frequency region, or both. Theoretically, these individual differences may show a large range. This discussion has suggested that H1 – H2 may be a good measure of OQ, and that H1 – A3 may serve as a measure of spectral tilt, which is related to the abruptness of glottal closure. The value

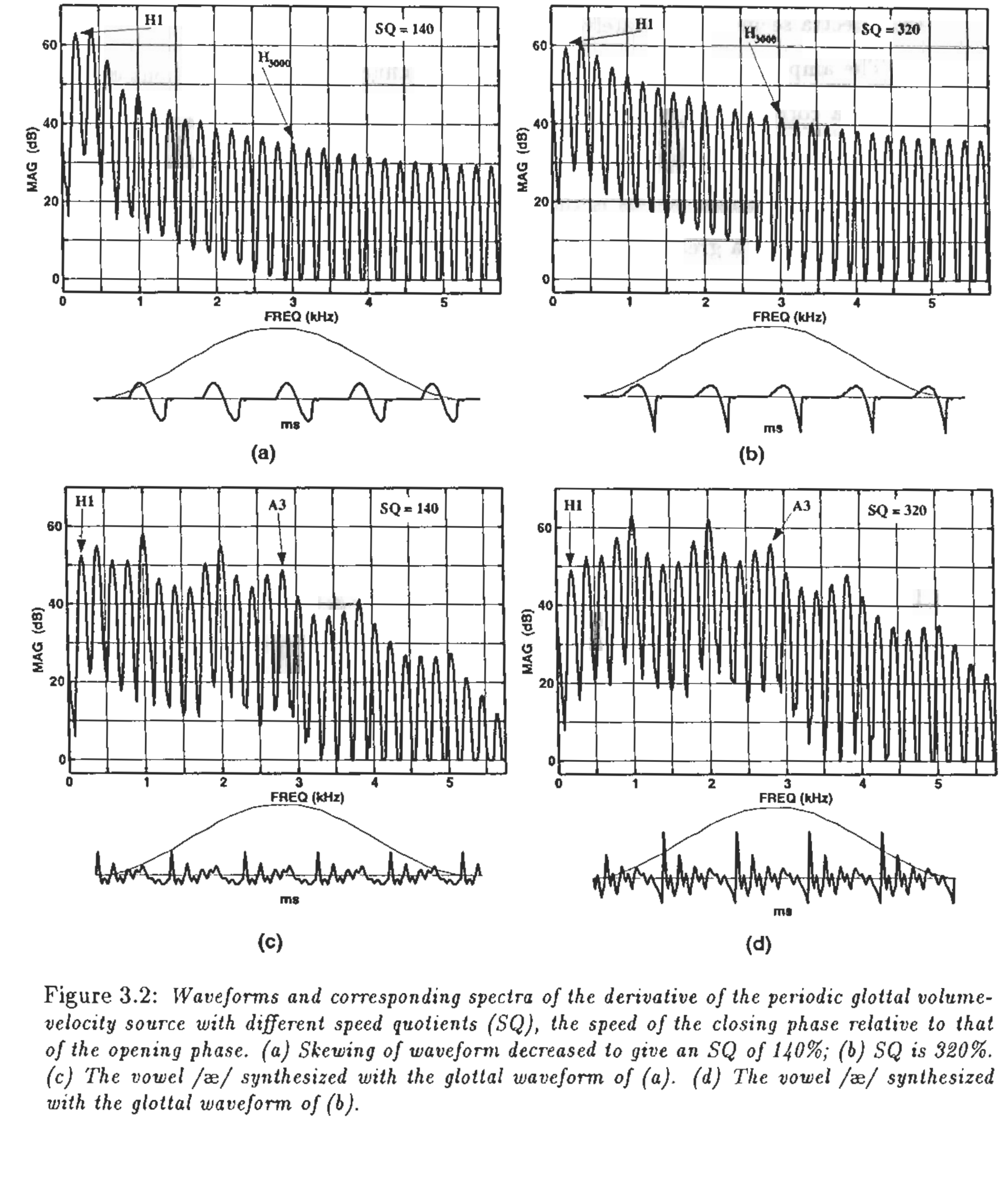


Figure 3.2: *Waveforms and corresponding spectra of the derivative of the periodic glottal volume-velocity source with different speed quotients (SQ), the speed of the closing phase relative to that of the opening phase. (a) Skewing of waveform decreased to give an SQ of 140%; (b) SQ is 320%. (c) The vowel /æ/ synthesized with the glottal waveform of (a). (d) The vowel /æ/ synthesized with the glottal waveform of (b).*

of this tilt measure is expected to exceed 9 dB for most speakers.

3.2.2 Incomplete glottal closure during a vibration cycle

Many speakers, however, configure the vocal folds so that the glottis is never completely closed during a cycle of vibration (see, for example, Holmberg et al., 1988; Södersten and Lindestad, 1990). In one such configuration there is a fixed opening, or glottal chink, between the arytenoid cartilages, but the vocal folds remain approximated at the posterior end, that is, at the vocal processes. Another glottal state is one in which the vocal processes also remain abducted throughout the glottal cycle. These glottal configurations exhibiting an airflow bypass can modify the basic spectrum of the glottal waveform in several ways relative to the spectrum that would exist for the configuration in which the entire glottis is closed over part of the cycle of vibration. Among the modifications introduced by this glottal chink are: (1) an increase in the bandwidth of the first (and possibly the second) formant; (2) an increased tilt in the glottal spectrum at high frequencies; and (3) emergence of a turbulence noise source in the vicinity of the glottis that is comparable in amplitude (at high frequencies) to the spectrum amplitude of the periodic source. Each of these attributes usually can be observed in the sound and hence can provide evidence for the existence of a glottal opening that is maintained throughout a cycle of vibration.

All of these acoustic manifestations of a glottal opening occur whether the bypass is a glottal chink or a widening at the vocal processes. However, the second of these properties, an increase in tilt, is expected to be more marked when there is also abduction of the arytenoids at the vocal processes. We consider now each of these acoustic correlates of a spread glottal configuration.

3.2.2.1 Effect on first-formant bandwidth

Formant bandwidths are related to the rate of energy loss in the vocal tract. The energy losses in the frequency range of the first formant come from several sources, including the resistance of the yielding walls of the vocal tract, and heat conduction and frictional losses at the walls. These energy losses in the vocal tract lead to a first-formant bandwidth of 40-95 Hz for female speakers in the closed glottis condition (Fujimura and Lindqvist, 1971; Fant, 1972). When the glottis is not closed and there is airflow through it, the glottal resistance can contribute further energy loss, particularly at low frequencies, thus adding significantly to the first-formant bandwidth. In fact, measurement of the F1 bandwidth can provide an indirect indication of the degree to which the glottis fails to close completely during a cycle of glottal vibration.

In earlier work (House and Stevens, 1958; Fant, 1962; Fujimura and Lindqvist, 1971) bandwidths were measured by exciting a subject's vocal tract while the subject held his or her glottis closed. This method measures the bandwidths due to vocal-tract losses. House and Stevens (1958) also measured bandwidths for the open glottis condition for their male subjects, and found that bandwidth did indeed increase under this condition. Bandwidths can also be estimated from the speech waveform. If the F1 oscillation is assumed to be of the form $e^{-\alpha t}\cos 2\pi f t$, that is, a damped sinusoid, where f is the frequency of the first formant, then the constant α (in sec^{-1}) is related to the bandwidth B1 hy the equation $\mathrm{B1} = \alpha/\pi$ Hz. Then by measuring the decay rate of the first formant waveform during the early part of the glottal period, where the glottal area is expected to be smallest, one can estimate the first-formant bandwidth, according to the following formula

$$\mathrm{B1} = \frac{1}{\pi}\frac{\ln\left(\frac{\mathrm{X1}}{\mathrm{X2}}\right)}{\frac{1}{2}(\mathrm{T1}+\mathrm{T2})} \tag{3.3}$$

where X1 and X2 represent the amplitudes of the peak-to-peak oscillations, and T1 and T2 represent the time between maximum and minimum amplitudes of the oscillations. These variables are illustrated in Fig. 3.3.

As an example of how F1 bandwidth is manifested in the acoustic sound pressure, waveforms of the radiated sound pressure for the vowel /æ/ produced by two different female speakers are shown in Figs. 3.4(a) and 3.4(c). The sound-pressure waveform during the initial part of each glottal period is a damped oscillation, the largest component of which is at the frequency of the first formant. The rate of decay of the amplitude of this oscillation is related to the F1 bandwidth. The increased rate of decay of the F1 oscillation during the last part of each cycle (particularly in Fig. 3.4(c)) reflects the increased losses at the glottis, and hence the increased bandwidth during the open phase (Fant, 1979). If the glottis remains open throughout the cycle of vibration, then the decay rate during the first part of the glottal cycle will also be increased relative to that for the closed-glottis

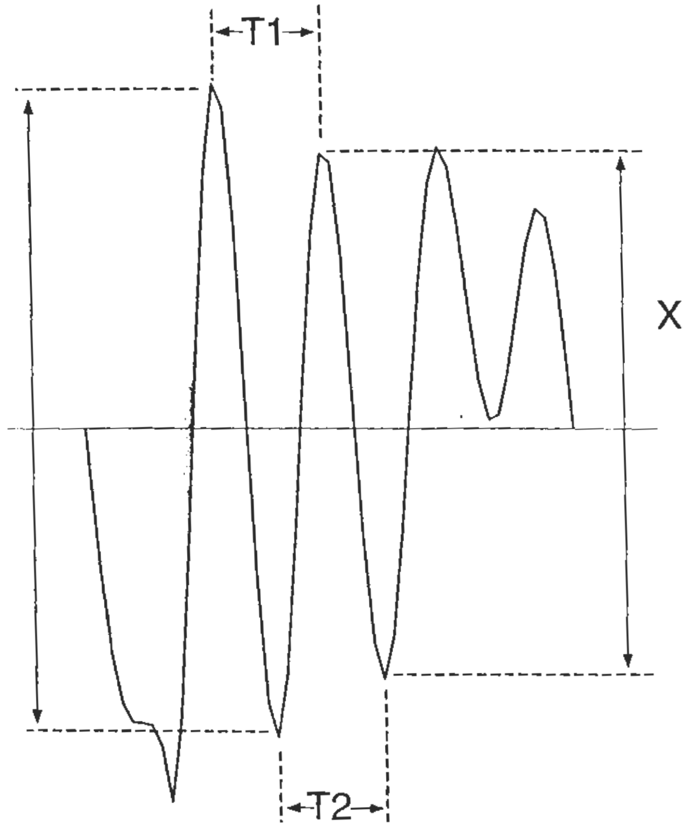


Figure 3.3: *An F1 waveform illustrating the measures X1, X2, T1, and T2 used to compute the decay over the first two oscillations. These measures are substituted into Eqn. 3.3 to compute the F1 bandwidth.*

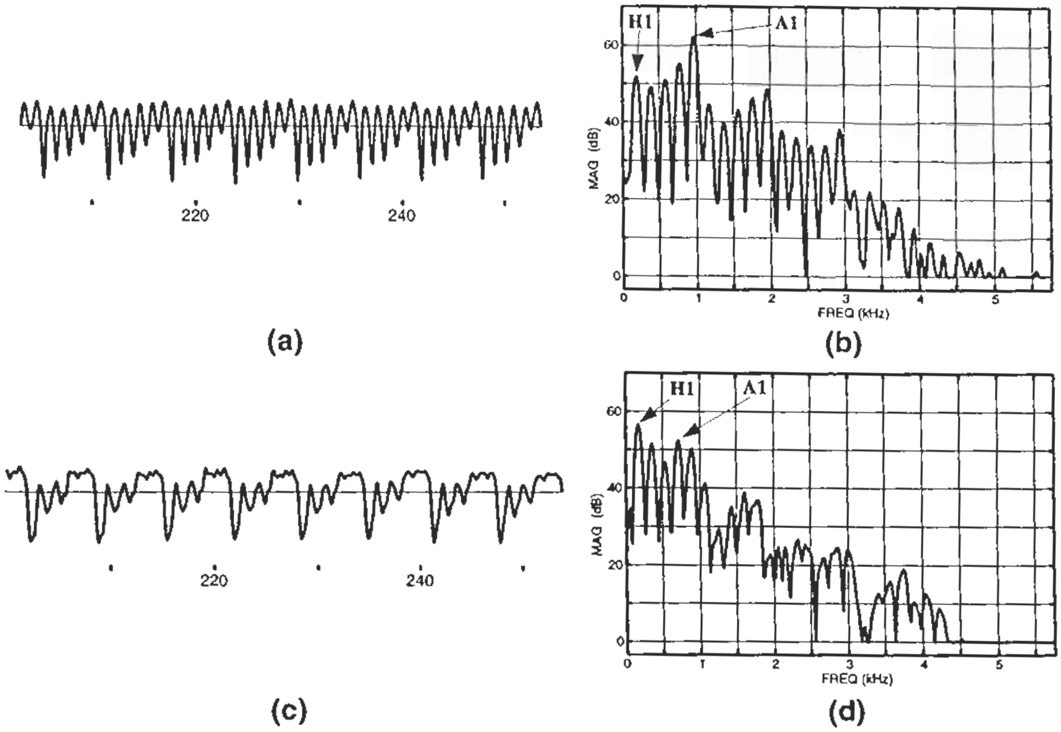


Figure 3.4: *Examples of waveforms and spectra of the vowel /æ/ produced by two different adult female speakers. The waveforms illustrate decay rates that are (a) slow and (c) rapid, corresponding to narrow and wide first-formant bandwidths, respectively. As estimated using Eqn. 3.3, the bandwidths during the first part of the cycle are about 60 Hz for (a) and 275 Hz for (c). From the corresponding spectra in (b) and (d), we see that a narrow first-formant bandwidth results in a stronger, more prominent first-formant peak.*

condition. For the waveforms in Fig. 3.4, the bandwidths during the first part of the cycle, as estimated using Eqn. 3.3, are about 60 Hz for (a) and 275 Hz for (c).

To get an accurate measure with this method, there must be a high enough F1 frequency and long enough pitch period to get at least two oscillations during the closed part of the cycle. Otherwise, the first oscillations will be affected by the extra losses during the open part of the cycle, and the B1 measure may be too large. The vowel /æ/ usually has a high first formant, making it a good candidate for this type of analysis. However, for females with high fundamental frequencies or relatively low F1 frequencies, the measure may be inaccurate.

Another measure of F1 bandwidth is the amplitude of the F1 peak in the speech spectrum. As predicted by theory, this amplitude should decrease by 6 dB when the bandwidth is doubled. Given the bandwidths measured for the waveforms in Fig. 3.4, we might expect a difference in relative amplitudes of the F1 peaks in the corresponding spectra to be about 12 dB. From the spectra in Fig. 3.4(b) and (d), we see that the larger bandwidth in (c) results in a reduced F1 peak amplitude, making the peak less prominent relative to the amplitude of the first harmonic. The values of H1 – A1 for these two spectra are about −10 dB for (b) and 4 dB for (d), resulting in a difference of about 14 dB between the two spectra, close to that predicted. This difference is mainly due to difference in first-formant amplitude (A1), but is also partially due to the variation in H1 across the two speakers.

This example suggests that the difference in amplitude of the first harmonic and the amplitude of the F1 peak (H1 – A1) may also be a suitable measure of bandwidth. Holmberg et al. (in press) have used this measure as an indicator of increased first-formant bandwidth resulting from increased subglottal coupling. However, this method gives an average bandwidth over the entire glottal cycle, including those times when the glottis is open. Thus, this method may give a larger value of bandwidth than that estimated from the waveform near the beginning of the glottal cycle. Variation across speakers in the relative amplitude of the first harmonic will also add some uncertainty to this measure.

For the measure H1 – A1, we can predict a minimum in the same way we did for the tilt measure H1 – A3 (see Section 3.2.1). From the glottal waveform synthesized using the KLGLOTT88 model, assuming the first formant is at about 600 Hz, the difference between the first and third harmonics is about 8 dB. The contribution of the first formant is about 23 dB, assuming that the formant is centered on a harmonic and has a bandwidth of 50 Hz, bringing the minimum H1 – A1 to −15 dB. However, on average, F1 will not be centered on a harmonic, increasing the minimum value somewhat. This increase will be greater for narrow bandwidths than for wide bandwidths, and on average may be about 4 dB, bringing the minimum to −11 dB. For a first-formant bandwidth of 250 Hz, this will increase to about 5 dB, giving a 16 dB range of H1 – A1. If the tilt breakpoint is low in frequency, the range of H1 – A1 may be even greater.

These measurements of bandwidth can be used to make estimates of the area of the glottis during the maximally constricted part of the cycle. Theoretical estimates of the

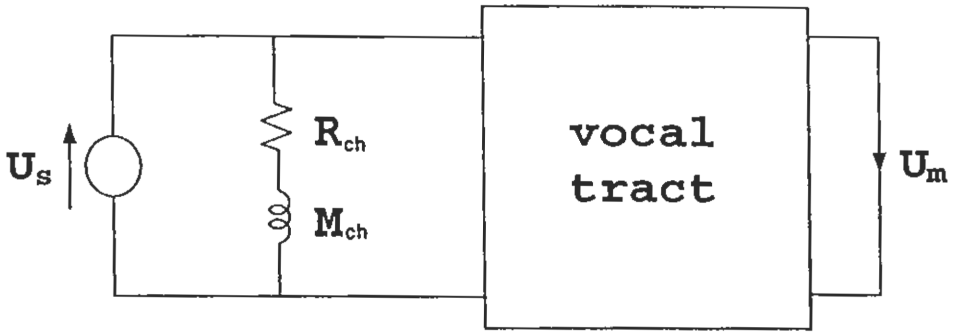


Figure 3.5: *Model of speech production when the membranous part of the folds have come together, but an opening remains at the arytenoid cartilages, the vocal processes, or both.* R_{ch} *and* M_{ch} *represent the resistance and mass of the glottal opening,* U_s *the volume velocity at the source, and* U_m *the volume velocity at the mouth.*

contribution of the glottal opening to the bandwidth B1 of the first formant can be made by calculating the value of the resistive termination at the glottis and determining the acoustic energy loss in this resistance (Fant, 1960). An equivalent circuit for calculating the losses is given in Fig. 3.5. The glottal impedance is represented by an acoustic resistance and an acoustic mass (Stevens, in preparation). If we assume that the glottis terminates a uniform tube of length ℓ_v and cross-sectional area A_v, the contribution B_g to the bandwidth of a formant is

$$B_g = \frac{\rho c^2}{\pi A_v \ell_v R_{ch}(1 + \frac{4\pi^2 f^2 M_{ch}^2}{R_{ch}^2})} \qquad (3.4)$$

where

ρ = density of air in vocal tract (g/cm^3)

c = speed of sound (cm/sec)

f = frequency (Hz)

R_{ch} = glottal resistance due to the chink (dyne-sec/cm^5)

M_{ch} = acoustic mass of the glottal chink (gm/cm^4)

(Stevens, in preparation). From this equation we see that as f increases, B_g decreases, so a glottal opening has its greatest effect on the bandwidth of F1. The pressure drop across the glottis can be approximated by

$$\Delta P \approx \frac{\rho U_{ch}^2}{2A_{ch}^2} \qquad (3.5)$$

(van den Berg et al., 1957; Stevens, in preparation) where U_{ch} is the airflow through the glottal chink and A_{ch} is the area of the chink. The glottal acoustic resistance R_{ch} is the derivative of the pressure drop ΔP with respect to volume velocity U_{ch}, or

$$R_{ch} = \frac{d\Delta P}{dU_{ch}} \approx \frac{\rho U_{ch}}{A_{ch}^2} \tag{3.6}$$

Assuming that the pressure drop across the glottis is equal to the subglottal pressure P_s, from Eqn. 3.5 we can write

$$P_s = \frac{\rho U_{ch}^2}{2A_{ch}^2} \tag{3.7}$$

from which we obtain

$$U_{ch} = A_{ch}\sqrt{\frac{2P_s}{\rho}} \tag{3.8}$$

and, substituting into Eqn. 3.6,

$$R_{ch} \approx \frac{\sqrt{2P_s\rho}}{A_{ch}} \tag{3.9}$$

M_{ch} can be expressed as

$$M_{ch} = \frac{\rho \ell_g}{A_{ch}} \tag{3.10}$$

where ℓ_g is the thickness of the glottis. Thus, for a given subglottal pressure, we can calculate B_g and U_{ch} as functions of A_{ch}, using Eqns. 3.4 and 3.8–3.10. Table 3.1 lists a range of A_{ch} values and the corresponding values of B_g, $B1$, and U_{ch}, where $B1 = B_g + B_v$, B_v being the F1 bandwidth due to vocal-tract losses (with a closed glottis). For /æ/, B_v is approximately 50 Hz for female speakers of Swedish (Fujimura and Lindqvist, 1971; Fant, 1972). From this table we see that bandwidth increments up to 200 Hz might be expected for glottal openings in the range up to 8 mm^2, when the subglottal pressure P_s is assumed to be 5500 dynes/cm^2. This minimum opening corresponds to a minimum flow of about 249 cm^3/sec, which is about the upper limit observed by Holmberg et al. (1994a) for 15 female speakers of American English.

3.2.2.2 Effect on spectral tilt

When there is a glottal chink with the arytenoid cartilages approximated at the vocal processes, the pattern of mechanical vibration of the vocal folds should be approximately the same as it is when there is no glottal chink. The shape of the airflow waveform will,

Table 3.1: *Range of glottal chink areas (A_{ch}) and corresponding glottal contribution to first formant (B_g); bandwidth of first formant (B1); flow through chink (U_{ch}); time constant (T) of flow cutoff; and resulting increment in spectral tilt at 2750 Hz. A subglottal pressure of 5500 dynes/cm^2 is assumed.*

A_{ch}	B_g	$B1$†	$20\log_{10} B1$	U_{ch}	T	*Tilt*
(cm^2)	(Hz)	(Hz)	(dB)	(cm^3/sec)	(ms)	(dB)
0.00	0	50	34	0	0	0
0.01	25	75	38	31	0.13	7
0.02	50	100	40	62	0.16	9
0.03	76	126	42	93	0.20	11
0.04	101	151	44	124	0.23	12
0.05	126	176	45	155	0.27	13
0.06	151	201	46	186	0.30	14
0.07	176	226	47	217	0.33	15
0.08	202	252	48	249	0.37	16
0.09	227	277	49	280	0.40	17
0.10	252	302	50	311	0.43	18

†Assuming vocal tract losses contribute 50 Hz.

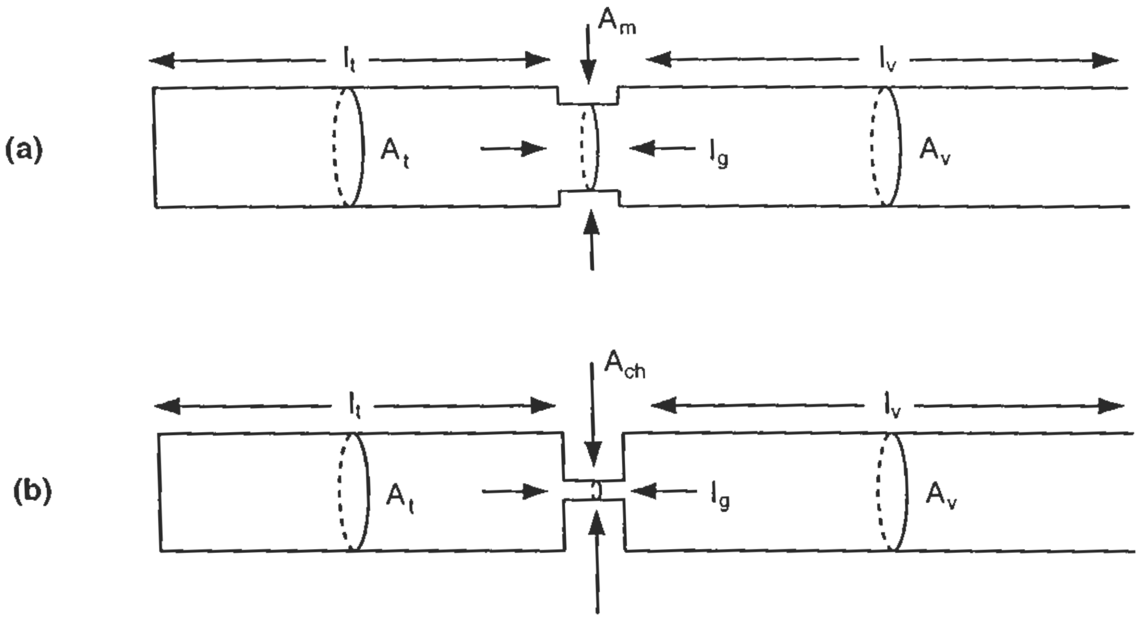


Figure 3.6: *Schematic drawings of the the glottal configuration (a) before, and (b) after the membranous part of the folds have come together. The area of the glottal opening changes abruptly from A_m, in (a), to A_{ch}, in (b) at closure. A_t and ℓ_t are the cross-sectional area and length of the trachea, while A_v and ℓ_v are those of the vocal tract. The parameter ℓ_g is the effective vertical length of the glottis. A_m is the cross-sectional area of the opening at the membranous part of the folds and A_{ch} is the cross-sectional area of the glottal chink.*

however, be influenced by the bypass through the interarytenoid space, particularly at the time when the vocal folds come together, because the acoustic mass of the airway and the presence of the bypass path prevent the modulated portion of the glottal airflow from being abruptly terminated.

Figure 3.6 shows schematic drawings of the situation before and after the folds come together. In (a) the tube representing the glottal opening just prior to closure has cross-sectional area A_g. At the time that the folds come together, the area of this tube changes abruptly to A_{ch}, as shown in (b). The situation can be modeled as in Fig. 3.7. When the switch is closed, the circuit represents the case where the glottis is open at the folds. The closure of the folds is modelled by the sudden opening of the switch, with the result that the glottal resistance and acoustic mass of the circuit change abruptly. The effect of this abrupt change at the vocal folds is that of applying a step excitation to the circuit. Thus, the flow at the instant of closure is limited by the time constant $T = M/R_{ch}$, where M

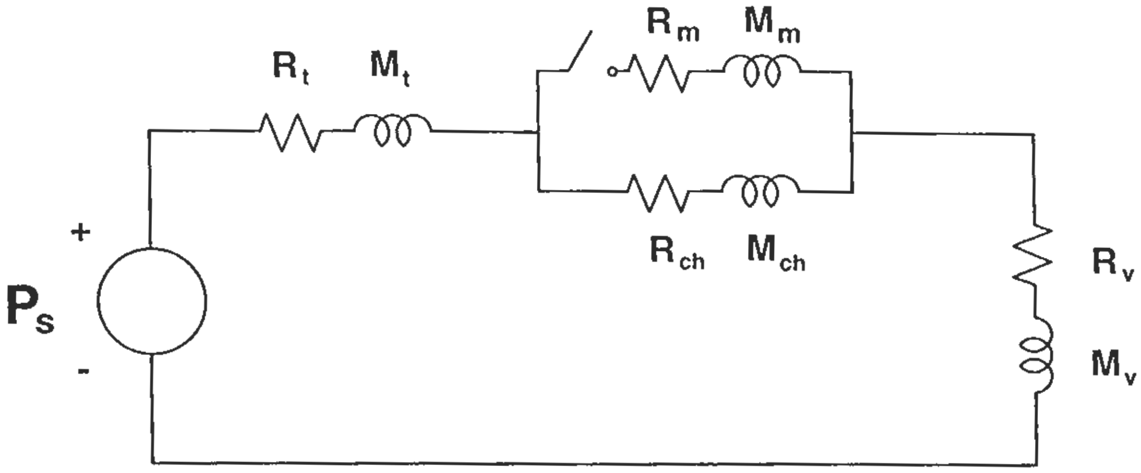


Figure 3.7: *Model of the speech production system when a fixed opening remains at the arytenoid cartilages during the "closed" phase of the glottal cycle. When the switch is closed the situation just prior to elosure is modeled. At closure, the switch opens. This abrupt change corresponds to a step excitation, which leads to a gradual, rather than abrupt, cutoff in flow. In this figure, P_s represents the subglottal pressure. R_t and M_t are the acoustic resistance and mass of the trachea, while R_v and M_v are those of the vocal tract. R_m and M_m are the acoustic resistance and mass of the opening at the membranous part of the vocal folds and R_{ch} and M_{ch} are those of the glottal chink.*

represents the acoustic mass of the air in the trachea, glottal chink, and vocal tract,

$$M = M_{\text{trachea}} + M_{\text{chink}} + M_{\text{vocal tract}} \tag{3.11}$$

(Stevens, in preparation). Since $M = \frac{\rho \ell}{A}$, we have

$$M = \rho \left(\frac{\ell_t}{A_t} + \frac{\ell_g}{A_{ch}} + \frac{\ell_v}{A_v} \right) \tag{3.12}$$

where

ℓ_t = length of the trachea

A_t = cross-sectional area of the trachea

ℓ_g = effective vertical length of the glottis

A_{ch} = cross-sectional area of the glottal chink

ℓ_v = length of the vocal tract

A_v = cross-sectional area of the vocal tract

The length and cross-sectional area of the trachea are about 11 cm and 2 cm^2, respectively, for females (Zemlin, 1988), and ℓ_g, the effective acoustic vertical length of the glottis, is about 0.3 cm (based on data from Titze, 1989a, 1989b). If we assume a vocal tract in a neutral setting, with a length ℓ_v of 15 cm and cross-sectional area A_v of 3 cm^2, then

$$\begin{aligned} M &\approx \rho\left(\frac{11}{2}+\frac{0.3}{A_{ch}}+\frac{15}{3}\right) && (3.13) \\ &= \rho\left(10.5+\frac{0.3}{A_{ch}}\right) && (3.14) \end{aligned}$$

From Eqn. 3.9

$$R_{ch} \approx \frac{\sqrt{2P_s\rho}}{A_{ch}} \qquad (3.15)$$

Then the time constant is

$$T = \frac{M}{R_{ch}} = \sqrt{\frac{\rho}{2P_s}}(10.5A_{ch}+0.3) \qquad (3.16)$$

This time constant leads to an additional 6 dB/octave tilt in the spectrum at high frequencies, with the extra tilt beginning at a frequency

$$f_T = \frac{1}{2\pi T}$$

This breakpoint can be translated into a measure of the number of decibels reduction in spectrum amplitude at 2750 kHz, which is approximately the frequency of the third formant for a female speaker. Table 3.1 summarizes some time constants T and the corresponding increases in spectral tilt that might be expected for a range of glottal chink areas. Based on minimum airflows measured from inverse-filtered waveforms (Holmberg et al., 1994) which have a range up to 256 cm^3/sec, the maximum increase in tilt that one should expect due to a glottal chink is about 16 dB.

When the arytenoid cartilages remain abducted at the vocal processes throughout the glottal vibration cycle, the membranous part of the folds does not close abruptly, but rather nonsimultaneously along the length of the glottis. As discussed in Section 3.2.1, this nonabrupt closing can contribute significantly to the spectral tilt at mid- to high frequencies, depending on the time it takes for the folds to close. With a glottal configuration that has both a fixed space between the arytenoids and some separation at the vocal processes, the effect on the spectral tilt in the F3 frequency region could be considerable. Thus, depending on the positioning of the arytenoids, including the vocal processes, during phonation, one might expect variations in the F3 range of the high-frequency spectrum that are substantially greater than 16 dB.

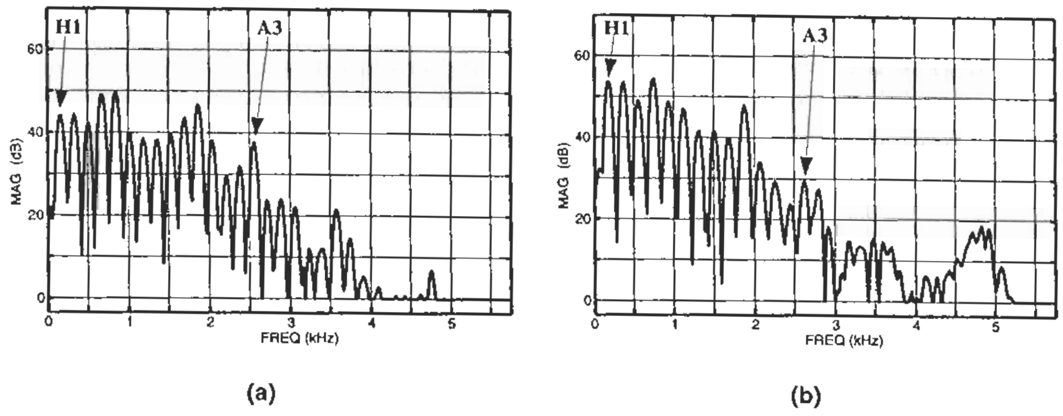


Figure 3.8: *Spectra of the vowel /æ/ produced by two adult female speakers with different amounts of spectral tilt. Time window for calculating spectrum is 22.3 ms.*

Examples of spectra for the vowel /æ/ produced by two different female speakers are displayed in Fig. 3.8. These spectra illustrate two extremes of spectral tilt. As discussed in Section 3.2.1, the difference (in dB) between the amplitude H1 of the first harmonic and the spectrum amplitude A3 of the third formant peak may be a suitable measure of spectral tilt, if certain corrections are made. The values of H1 − A3 in these two examples are 6 and 23 dB.

3.2.2.3 Turbulence noise at the glottis

Another acoustic consequence of a glottal opening is the generation of turbulence noise in the vicinity of the glottis. Based on theoretical analysis and experimental evidence, it is possible to make some estimates of the amplitude and the spectrum of the turbulence noise source at the glottis when the glottal area and the transglottal pressure are known (Shadle, 1985; Stevens, 1993). From this kind of analysis we can compare the spectrum of the periodic glottal source to the effective spectrum of the noise source due to turbulence in the vicinity of the glottis.

When there is modal vocal-fold vibration, with complete glottal closure during half of the cycle, the comparison of the periodic and noise source spectra is shown by the solid lines in Fig. 3.9. Both of these sources are filtered by essentially the same vocal-tract transfer function to yield formant prominences. These source spectra are the result of calculations based on theoretical and experimental data from turbulence noise sources and from periodic glottal sources (Shadle, 1985; Stevens, 1993). The ratio of the amplitude of the harmonics at 3 kHz to the noise amplitude in a 50-Hz band at the same frequency is 17 dB. Over the entire frequency range up to 5 kHz the noise spectrum is well below the spectrum of the periodic source, so that the combined spectrum is expected to show well-defined harmonics.

When the glottal area does not decrease to zero over a cycle of vibration, the spectra given by solid lines in Fig. 3.9 change in two ways. The spectrum amplitude of the periodic component becomes weaker at high frequencies, as noted above, and the amplitude of the turbulence noise increases because of the increased flow. For a given subglottal pressure, the amplitude of the turbulence noise source at the glottis is expected to increase approximately in proportion to $A_g^{0.5}$, where A_g is the average glottal area during a cycle of vibration (Stevens, 1971). For example, the average glottal area during modal glottal vibration in which the glottis is closed during a portion of the cycle is approximately 0.03 cm^2 for an adult female. If a fixed glottal chink of 0.05 cm^2 is added to this area, the amplitude of the turbulence noise is expected to increase by about 4 dB. As noted earlier in Table 3.1, however, the spectral amplitude of the periodic glottal source decreases by about 13 dB at 2750 Hz, giving a 17 dB decrease in harmonics-to-noise ratio in this frequency range. The two spectra now have the form given as dashed lines in Fig. 3.9, with the noise spectrum being comparable to the periodic spectrum at high frequencies.

Numerous researchers have developed objective measures of the noise present in the speech waveform during glottal vibration (see, for example, Yumoto et al., 1982; Ladefoged and Antoñanzas-Barroso, 1985; Kasuya and Ogawa, 1986; Klingholz, 1987; de Krom, 1993; Hillenbrand et al., 1994; Mori et al., 1994). Usually these methods involve isolating the periodic component of the speech waveform from the noisy component. This can be done through spectral- or cepstral-based analysis, or through comparing the pitch periods in the time domain, measuring the differences between pitch periods that result from the statistical variability of noise. However, as pointed out by Ladefoged and Antoñanzas-Barroso (1985), these methods do not measure just the noise that is due to an aspiration source, but rather the noise that results from a combination of factors. These other factors include jitter (changes in pitch) and shimmer (changes in amplitude of excitation). Their

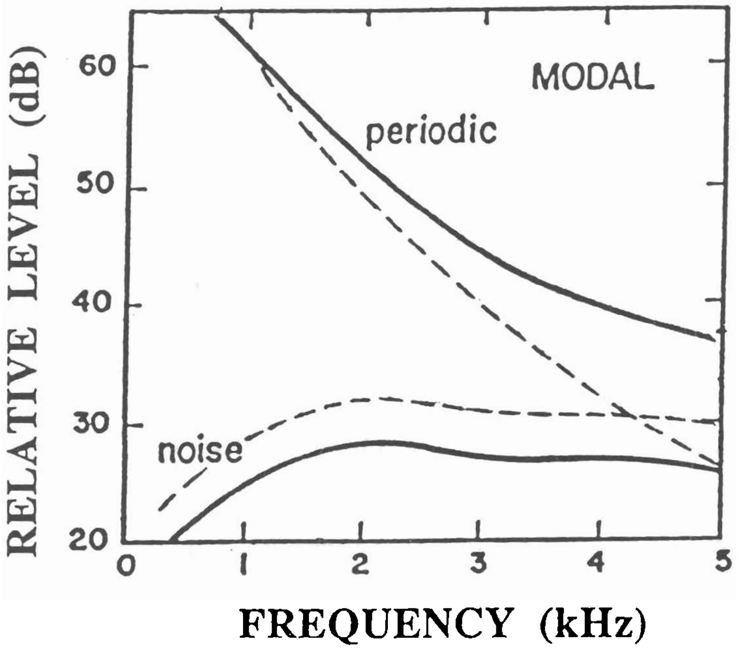


Figure 3.9: *Calculated spectra and relative amplitudes of periodic volume-velocity source and turbulence-noise source for two different glottal configurations: a modal configuration in which the glottis is closed over one-half of the cycle (solid lines), and a configuration in which the minimum glottal opening is 0.1 cm^2 (dashed lines). The spectrum for the periodic component gives the amplitudes of the individual harmonics. The noise spectrum is the spectrum amplitude in 50 Hz bands. The calculations are based on theoretical models of glottal vibration and of turbulence noise generation (Stevens, 1993; Shadle, 1985). (From Stevens and Hanson, 1995 and Stevens, in preparation)*

solution was to use only *part* of a vibratory cycle and compare it with the corresponding part of the next cycle.

Klatt and Klatt (1990) suggest two problems with this waveform-based measure. First, the waveform is dominated by the lower formants because they have a greater amplitude, particularly F1, while aspiration noise occurs primarily at high frequencies. This problem can be reduced by highpass or bandpass filtering. Second, unless the fundamental frequency is an exact multiple of the sampling period, even a perfectly periodic waveform will appear aperiodic, due to frequency components near the Nyquist frequency that are represented by only a few samples. This can only be remedied by significant oversampling.

To quantify the noise component in relation to the periodic component, we have chosen to define a harmonics-to-noise ratio as the ratio of the level of the harmonic with the greatest amplitude in the third-formant region (for a nonretroflexed vowel) to the level of the aspiration noise in the same region, both levels being measured from the spectrum calculated with a 22.3 ms hamming window (bandwidth of about 90 Hz (Rabiner and Schafer, 1978)). Of course, it is not possible to separate the noise from the periodic component and to measure each separately. However, the harmonics-to-noise ratio *can* be determined for vowels synthesized with a formant synthesizer that contains a periodic glottal source and an aspiration noise source.

Figure 3.10(b) shows the spectrum of a synthesized vowel /æ/ with formant frequencies and fundamental frequency at values appropriate for an adult female speaker, but with no aspiration noise. Above this spectrum, in Fig. 3.10(a), is the spectrum of the same vowel when the sound source is continuous aspiration noise with a suitably shaped spectrum. The level of this aspiration at 3 kHz, the frequency of the third formant, is 8 dB below the level of the highest harmonic in the F3 region in Fig. 3.10(b), also at 3 kHz, in a 90-Hz band. When the two are mixed, the result is the spectrum in Fig. 3.10(d). The harmonics-to-noise ratio for this composite spectrum is defined to be 8 dB. (In the synthesizer, the noise amplitude is modulated by the glottal source, so that the harmonics-to-noise ratio as just defined refers to the peak level of the noise during the glottal cycle.) Fig. 3.10(c) displays the spectrum of the same vowel synthesized with an additional tilt (10 dB) in the periodic glottal spectrum. The level of aspiration (Fig. 3.10(a)) at 3 kHz is now about 2 dB *above* the level of the highest harmonic in the F3 region in Fig. 3.10(c). The spectrum of the vowel synthesized with both sources is shown in Fig. 3.10(e), and the harmonics-to-noise ratio for this combined spectrum is defined to be −2 dB.

Figure 3.8 shows the effect of turbulence noise at the glottis in the spectrum of a natural vowel. The harmonic structure of the spectrum in Fig. 3.8(b), which has a more extreme tilt, becomes less apparent at high frequencies (2.5 kHz and above), presumably because of the effect of the aspiration noise.

The influence of aspiration noise can also be seen by examining a vowel waveform when it is bandpass filtered at F3, with a bandwidth of 600 Hz. The two F3 waveforms corresponding to Figs. 3.10(d) and 3.10(e) are shown in Figs. 3.10(f) and 3.10(g). The effect of a 10 dB difference in the harmonics-to-noise ratio is clear. The waveform in Fig. 3.10(f), while showing signs of noise excitation, still has a periodic nature. However, the waveform in Fig. 3.10(g) shows mainly noise, with much less evidence of periodic excitation.

The technique of estimating the amount of noise in relation to the periodic component by examining the bandpassed waveform in the F3 region, such as those in Figs. 3.10(f) and 3.10(g), has been used by Klatt and Klatt (1990). It is also possible for an observer to make estimates of the amount of noise in a spectral representation, such as those of Fig. 3.8. The observer makes estimates of the amount of noise on a scale from 1 to 4, where 1 means there is essentially no evidence of noise interference and 4 means that there is little evidence of periodicity. Separate estimates are made from the waveform and from the high-frequency part of the spectrum.

To relate these scaling methods to the physical characteristics of the stimuli, we have made a set of judgments for a series of synthesized vowel stimuli. These synthetic vowels were generated with known amplitudes of aspiration noise in relation to the periodic glottal source, so that the harmonics-to-noise ratio of the stimuli are known. Stimuli of the type shown in Figs. 3.10(d) and Fig. 3.10(e) were synthesized with several amplitudes of the aspiration noise source and with several amounts of spectral tilt. The spectrum for each vowel was generated, and two judges independently rated the noisiness of these spectra on a scale from 1 to 4, following the procedure described by Klatt and Klatt (1990).

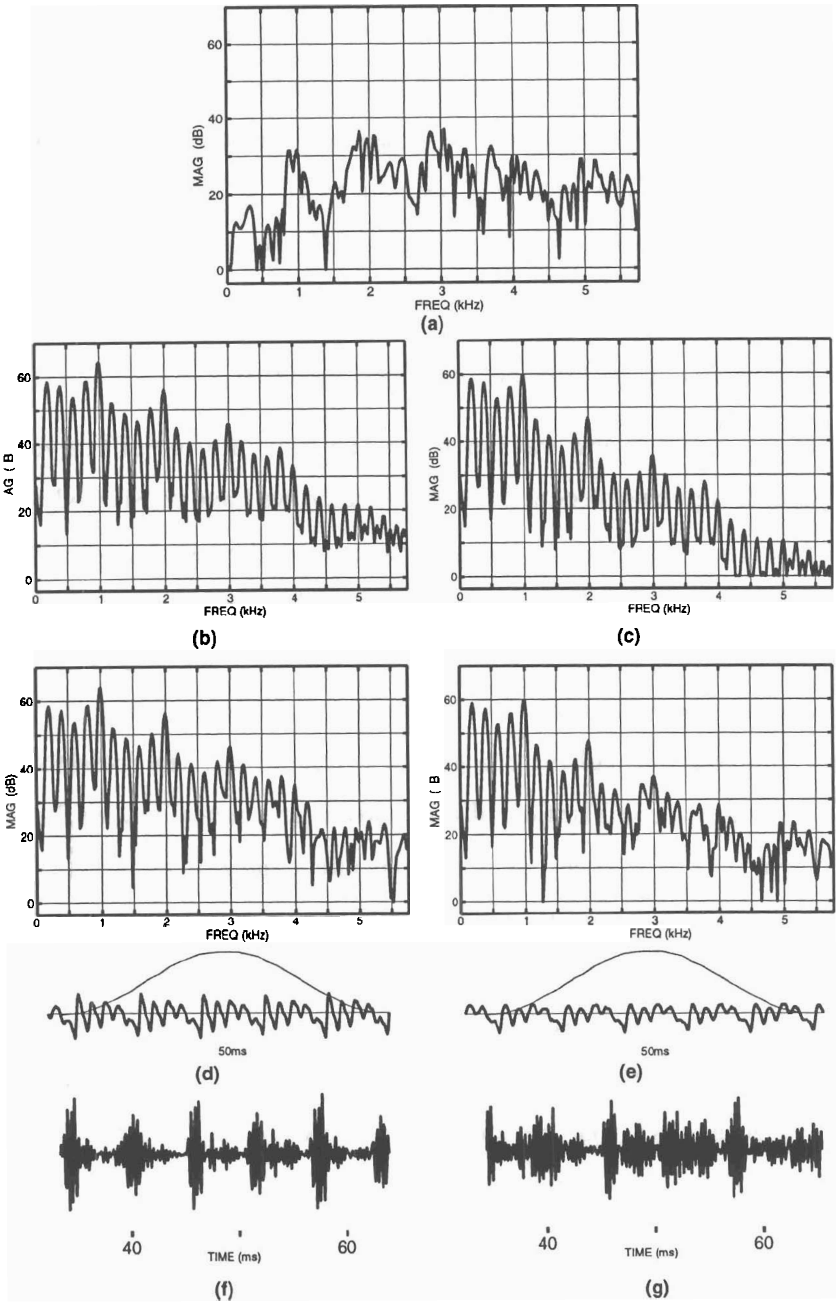


Figure 3.10: *Waveforms and spectra of the synthesized vowel /æ/ illustrating how aspiration noise influences the waveforms and spectra. Panel (a) shows the spectrum when the only source is aspiration noise. The spectra in (b) and (c) give the spectrum when the only source is the periodic glottal source, but with two different values of source spectral tilt (TL). The spectra in (d) and (e) show the result of mixing the aspiration and periodic components of the source. The waveforms of the two vowels are displayed immediately below these spectra. The waveforms (f) and (g) at the bottom were generated by bandpass filtering the waveform with a filter having a center frequency of 3 kHz and a bandwidth of 600 Hz. The harmonics-to-noise ratio (at 3 kHz) is 8 dB for the vowel in the left column and −2 dB for the vowel in the right column.*

Thus for each stimulus we have a measure of the harmonics-to-noise ratio and we have average judgments from the observers based on the spectrum. Figure 3.11 shows a plot of the harmonics-to-noise ratio vs. average noise judgments for these synthesized vowels, including a straight line that has been fit to the data. Using this plot, judgments for synthetic stimuli can be related to similar judgments for spoken vowels, as discussed in Section 3.3.3.

3.2.3 Summary of theoretical background

We have discussed several ways in which the configuration of the vocal folds and glottis may vary during vowel production. Specifically, we have considered four types of configurations: (1) the arytenoids are approximated and the membranous part of the folds close abruptly; (2) the arytenoids are approximated, but the membranous folds close nonsimultaneously along the length of the folds; (3) there is a fixed bypass airway, or "chink," at the arytenoids, but the folds close abruptly; (4) both the vocal processes and arytenoids remain abducted throughout the glottal cycle, forcing the folds to close nonsimultaneously. Through a combination of observation and modeling, we have suggested several ways in which these various configurations affect the glottal airflow and are manifested in the speech spectrum or waveform. Note that there may be other glottal configurations in addition to the four that we have considered.

As a result of the theoretical discussion, we have suggested several measures that can be made directly on the spectra and waveforms of natural vowels and that may give some indication of the vocal fold and glottal configuration during vowel production. A summary of these measures follows:

- A change in open quotient affects the spectrum mainly at low frequencies, so the difference in amplitude of the first two harmonics, H1 – H2, should give some measure of OQ.

- There are several sources of change in the spectral tilt of the voicing source: increases in speed quotient, or skewness of the glottal pulse, presence and size of posterior glottal chinks, and nonsimultaneous closure of the membranous part of the vocal folds all lead to decreases in the abruptness with which the airflow through the

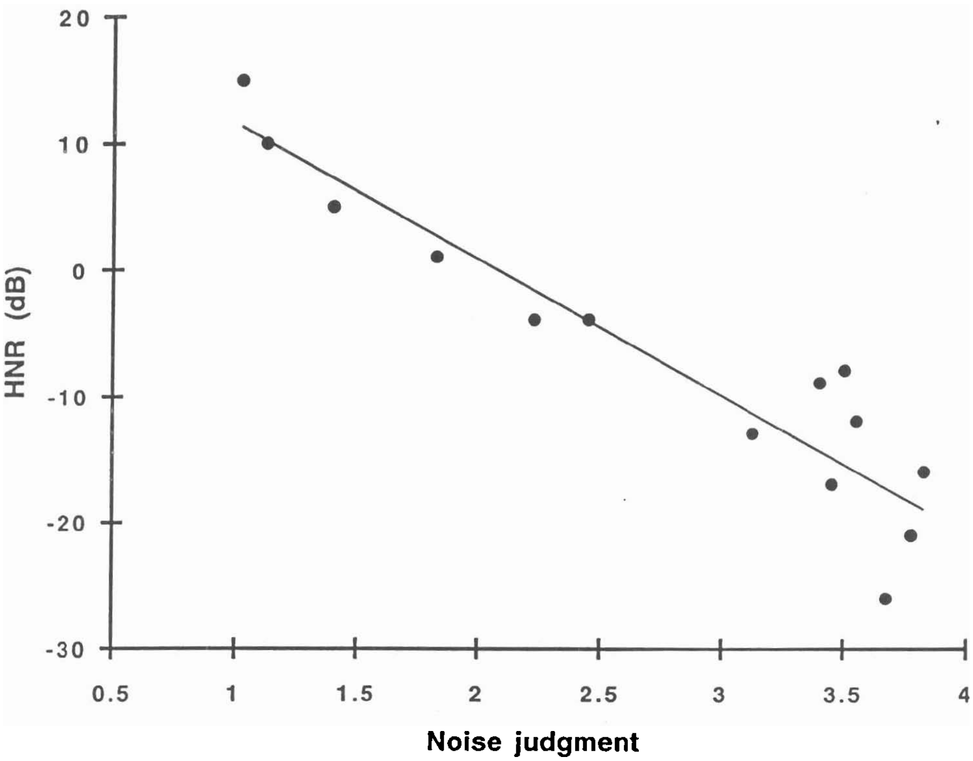


Figure 3.11: *Harmonics to noise ratio vs. noise rating for spectra of synthesized vowels.*

glottis is cut off. Decreases in this abruptness lead to increases in spectral tilt. These increases in the tilt of the glottal source spectrum are most evident at mid- to high frequencies, so we will use the difference between the amplitude of the first harmonic and the amplitude of the third formant peak, H1 – A3, as a measure of spectral tilt.

- The presence and size of a posterior glottal opening affects the first-formant bandwidth. These increases may be observed in both the speech waveform and spectrum. In the waveform the oscillations due to the first formant damp out more rapidly, and in the spectrum the amplitude of the F1 peak is reduced. Thus, we will use two measures of F1 bandwidth: one an estimate of the decay rate of the F1 waveform oscillation, and the other the difference between the amplitude of the first harmonic and the amplitude of the first formant peak, H1 – A1.

- Finally, the high-frequency noise content of the speech waveform and spectrum will increase as the size of a posterior glottal opening increases. This noise will be estimated using subjective ratings of noise in the F3 waveforms (Klatt and Klatt, 1990) and in the spectrum. These ratings can be related to harmonics-to-noise ratios using Fig. 3.11.

The theory predicts relationships between these measures in some cases, particularly under conditions where the glottis does not close completely during some part of the vibration cycle. For example, we see in Table 3.1 that as the area of the glottal chink increases, both the F1 bandwidth and the spectral tilt are expected to increase, and we also expect the strength of the noise source to increase.

In the remainder of this chapter we describe some data that were collected for 22 female speakers, and we attempt to interpret these data in terms of the theoretical models.

3.3 Experimental data

3.3.1 Speakers and speech material

We collected recordings of a number of utterances from 22 adult female subjects in the age range 22 to 49 years. The speakers showed no evidence of voice or hearing problems, and all were native speakers of American English. The utterances consisted of three nonhigh vowels, /æ, ɛ, ʌ/, embedded in the carrier phrase "Say bVd again." Each utterance was repeated five times, with the 15 sentences presented in random order during a single session. All the utterances were low-pass filtered at 4.5 kHz, digitized with a sampling rate of 11.4 kHz, and stored for further analysis.

3.3.2 Measurements

The acoustic measurements summarized in Section 3.2.3 were extracted from these utterances in the following manner:

First-formant bandwidths. For all repetitions of the vowel /æ/ the first-formant bandwidth during the initial part of the glottal cycle was estimated from the rate of decay of the waveform. The rate of decay was determined from the change in the peak-to-peak amplitude in the first two cycles of the F1 oscillation, using Eqn. 3.3. Estimates were made for eight consecutive pitch periods in a relatively stable portion of the vowel, generally at the middle. To reduce interference by the second formant, the waveforms were bandpass filtered with a filter having a bandwidth of 600 Hz centered at the first formant frequency. These 40 estimates were then averaged to obtain a mean value for each speaker. This analysis was restricted to the vowel /æ/ because for this vowel, the first formant is usually high enough so that two oscillations of the formant waveform occur during the closed part of the glottal vibratory cycle, and the second formant is well separated from the first.

H1* – H2*. The difference between the amplitudes of the first and second harmonics was measured for all repetitions of all three vowels. For /æ/, H1 – H2 was measured from the spectrum obtained by centering a 22.3 ms Hamming window during the initial part of the glottal cycle, at the eight points where the F1 bandwidth was estimated. For /ʌ/ and /ɛ/, the measurements were taken at three points in midvowel, 20 ms apart, where the formants were relatively stable. Corrections were made for the amounts by which H1 and H2 are "boosted" by the first formant,[1] yielding the measure H1* – H2*. This corrected measure can be compared across vowels and

[1] Correction given in Appendix A.1

across speakers. The values for each repetition were averaged to obtain a mean value for each vowel for each speaker.

H1* – A1. The difference between the (corrected) amplitude of the first harmonic and the amplitude of the first formant peak (A1) was measured. A1 was estimated by measuring the amplitude of the strongest harmonic of the F1 peak. The measurements were taken at the same points as those for H1* – H2*, and similarly, average values were computed for the three vowels for each speaker.

H1* – A3*. The difference between the amplitudes of the first harmonic and the third formant peak (A3) was measured. As was done for A1, A3 was estimated using the strongest harmonic of the F3 peak. H1 was corrected as above, and A3 was corrected for the effect of F1 and F2 on the spectrum amplitude of the third formant.[2] For this normalization F1 and F2 were set to 555 and 1665 Hz, respectively, based on the average F3 measured for all speakers. As mentioned earlier, A3 is also dependant on the bandwidth of F3. House and Stevens (1958) measured F3 bandwidths of male speakers for /æ, ʌ, ɛ/ to be 103, 64, and 88 Hz, respectively. In dB this means that /æ/ is expected to have an F3 amplitude that is 4 dB less than that of /ʌ/, while that for /ɛ/ is 3 dB less. For females speakers, the bandwidth values will be higher, but because data are not available for these vowels for female speakers, we made corrections based on the male data. This use of male data should result in minimal error because the ratio between the bandwidths is used to compute the difference in dB and this ratio is not expected to be very different across gender. Thus the value of A3 measured for each token of /æ/ and /ɛ/ was increased by 4 and 3 dB, respectively. The combination of these two corrections, for the location of F1 and F2, and for the F3 bandwidth, yields a normalized H1* – A3*.

Noise ratings. All repetitions of the three vowels were bandpass filtered around F3 using a filter having a bandwidth of 600 Hz. The bandpass filtered waveforms and the speech spectra corresponding to the speech segments used in the previously described measures were given ratings for noise, as described in Section 3.2.2.3. These judg-

[2] Correction given in Appendix A.2

ments were made independently by two judges, who did not know which waveforms or spectra corresponded to which speaker. Their average ratings were highly correlated ($r > 0.92$) and were averaged to obtain two noise judgments for each speaker, one based on the waveforms and the other on the spectra. The waveform-based ratings were found to be well correlated with the spectrum-based ratings. Analysis of variance showed a significant difference between the two methods ($F = 64$, $p = 8.1 \times 10^{-8}$), for the vowel /ɛ/. For /ʌ/ the results for the two measures were almost the same ($F = 4.9$, $p = 0.04$). For /æ/ there was no significant difference ($F = 0.08, p = 0.39$).

3.3.3 Results

3.3.3.1 Mean values

The mean values of the acoustic measurements for each speaker are summarized in Tables 3.2-3.4. Minimum and maximum values for each measure across speakers are given in boldface in these tables. H1* – H2* has a range of about 10 dB, corresponding roughly to a 40 percent range in open quotient (see Fig. 3.1). H1* – A3* has a range of about 26 dB, indicating a wide variation in spectral tilt among the subjects. This large range of spectral tilt is assumed to be a consequence of the presence of a glottal chink or a nonsimultaneous closure along the length of the glottis, or both, for some speakers. The minimum value of tilt is 8.6 dB, about what might be expected for the case where there is complete, abrupt glottal closure during some part of the glottal cycle (see Section 3.2.1). The range of H1* – A1 is 16 dB, as predicted earlier, and the minimum and maximum values are very close to those predicted in Section 3.2.2.1, −11 and 5 dB. The range of values obtained suggests that first formant peaks vary from being very prominent for some speakers to being highly damped for others, although part of this range can be due to variation in the amplitude of H1 and how well F1 is centered on a harmonic across speakers. This range of first-formant amplitudes presumably arises in part due to a range of F1 bandwidths and in part due to differences in the degree to which spectral tilt extends to the low frequency harmonics.

The first-formant bandwidth estimates for /æ/ vary from 53 Hz to 280 Hz. For the

Table 3.2: Average acoustic measures for the vowel /æ/, 22 female speakers, where H1* − H2*, H1* − A1, and H1* − A3* are given in dB, N_w and N_s are the waveform- and spectra-based noise judgements, and B1 is the bandwidth of the first formant, given in Hz. Numbers in boldface represent maxima or minima for each measure across speakers.

Subject	H1* − H2*	H1* − A1	H1* − A3*	N_w	N_s	B1
F1	3.5	−0.2	30.7	3.0	2.8	194
F2	1.7	0.4	32.2	2.8	2.9	244
F3	4.4	−8.0	32.1	2.7	2.8	94
F4	1.6	−5.7	**13.0**	1.6	1.6	209
F5	5.4	2.2	**35.0**	**3.8**	2.7	245
F6	2.4	−5.5	23.0	**1.1**	**1.1**	153
F7	3.8	−1.3	31.3	3.1	**3.1**	150
F8	2.1	−3.7	32.6	2.9	2.7	97
F9	2.8	−7.2	16.8	1.2	1.2	104
F10	5.0	**3.9**	26.4	2.2	2.6	184
F11	4.5	−4.4	19.5	1.8	2.1	158
F12	0.7	−5.6	31.3	2.4	2.2	217
F13	3.8	−8.9	19.4	1.7	1.2	**53**
F14	5.2	**−11.3**	16.3	**1.1**	1.2	78
F15	6.2	0.3	33.7	3.1	2.4	256
F16	**6.8**	1.2	30.4	2.3	2.5	132
F17	1.6	−2.6	22.0	2.0	1.8	**280**
F18	4.5	−2.2	21.8	2.0	2.5	163
F19	5.4	−0.5	24.3	2.0	2.0	166
F20	0.9	−6.2	14.7	1.7	1.6	178
F21	0.8	−8.5	17.9	1.5	1.4	124
F22	**0.6**	−9.2	20.8	1.4	1.2	149
Mean	*3.4*	*−4.2*	*24.1*	*2.1*	*2.1*	*165*

Table 3.3: *Average acoustic measures for the vowel /ʌ/, 22 female speakers, where* $H1^* - H2^*$, $H1^* - A1$, *and* $H1^* - A3^*$ *are given in dB, and* N_w *and* N_s *are the waveform- and spectra-based noise judgements. Numbers in boldface represent maxima or minima for each measure across speakers.*

Subject	$H1^* - H2^*$	$H1^* - A1$	$H1^* - A3^*$	N_w	N_s
F1	4.8	2.8	26.4	3.0	2.8
F2	1.2	−0.3	25.2	2.7	2.9
F3	3.6	−1.7	26.0	2.7	2.7
F4	**−0.7**	−9.0	**10.9**	1.8	1.3
F5	3.7	1.5	29.1	2.3	2.4
F6	3.0	−6.6	18.9	**1.4**	1.2
F7	1.8	−1.0	28.3	3.2	**3.5**
F8	3.0	−2.7	29.2	2.5	2.3
F9	1.5	−6.4	20.6	1.7	1.8
F10	3.1	2.8	24.7	2.4	2.3
F11	3.9	−2.9	22.0	1.7	2.1
F12	2.2	−5.8	22.9	2.2	1.9
F13	2.7	−4.4	15.5	**1.4**	**1.1**
F14	5.1	−11.9	15.1	**1.4**	1.3
F15	3.6	−4.0	27.2	2.9	2.3
F16	**5.8**	**3.5**	24.6	2.0	2.3
F17	1.5	−4.0	22.7	2.4	1.7
F18	3.5	−2.8	18.5	1.7	2.0
F19	5.0	1.3	**34.1**	**3.5**	3.2
F20	−0.2	−9.9	14.9	1.6	1.7
F21	0.1	−6.8	20.5	2.5	1.6
F22	0.3	**−12.1**	14.8	2.1	1.2
Mean	*2.6*	*−4.1*	*22.0*	*2.2*	*2.0*

Table 3.4: *Average acoustic measures for the vowel /ɛ/, 22 female speakers, where* $H1^* - H2^*$, $H1^* - A1$, *and* $H1^* - A3^*$ *are given in dB, and* N_w *and* N_s *are the waveform- and spectra-based noise judgements. Numbers in boldface represent maxima or minima for each measure across speakers.*

Subject	$H1^* - H2^*$	$H1^* - A1$	$H1^* - A3^*$	N_w	N_s
F1	6.3	1.7	28.8	3.2	2.9
F2	1.3	−2.0	27.4	2.8	2.1
F3	3.5	−3.1	31.9	3.2	3.1
F4	0.9	−11.0	**8.6**	1.7	1.1
F5	5.4	**3.7**	30.6	3.2	3.0
F6	3.3	−9.0	17.3	1.4	**1.0**
F7	3.1	−2.5	27.3	**3.6**	**3.3**
F8	2.6	−3.8	29.8	2.4	2.2
F9	3.0	−4.3	19.9	2.2	1.5
F10	6.5	2.5	22.6	2.7	2.5
F11	4.6	−5.8	18.0	1.8	1.7
F12	1.9	−5.7	26.0	2.1	1.6
F13	3.0	−5.3	16.0	**1.6**	1.1
F14	4.0	**−12.4**	16.6	1.9	1.2
F15	4.0	−1.1	30.2	2.5	1.9
F16	**6.9**	−1.6	29.4	2.9	1.9
F17	2.4	−5.3	27.1	2.7	2.3
F18	4.2	−3.7	16.5	**1.6**	1.6
F19	5.1	−3.9	**32.8**	2.8	2.5
F20	−0.8	−10.3	13.7	1.9	1.4
F21	1.5	−5.5	20.4	1.9	1.2
F22	**−2.6**	−6.7	15.5	1.7	1.2
Mean	*3.1*	*−4.7*	*22.5*	*2.3*	*1.9*

Table 3.5: *Results of analyses of variance (ANOVAs) performed to examine differences in acoustic measures across vowels.*

Measure	*F*	*p*
H1* – H2*	4.035	†0.025
H1* – A1	0.848	0.435
H1* – A3*	5.255	†0.009
Waveform-based noise	1.970	0.152
Spectra-based noise	2.237	0.119

†In pairwise analysis, only /æ/ and /ʌ/ are significantly different.

speaker with the lowest value of bandwidth (53 Hz), this estimate is about what is expected for the closed-glottis condition (Fant, 1972). For speakers with higher values of bandwidth, losses must exist at the glottis. Theoretical analysis of glottal losses indicates that a first-formant bandwidth of 280 Hz corresponds to a minimum glottal opening of about 0.09 cm^2 (see Table 3.1), while 75 Hz corresponds to about 0.01 cm^2, so we have a range of glottal chink cross-sectional areas of about 0.08 cm^2. The noise judgments range from 1.0 to 3.8; that is, some of our speakers show little to no noise in the high frequency range, while other speakers have substantial noise.

3.3.3.2 Statistical analysis

Analysis of variance was performed for all measures (except B1) to examine differences in parameter values among the different vowels. The results are summarized in Table 3.5. As seen in the table, across all vowels H1* – H2* and H1* – A3* were found to be significantly different ($p < 0.05$). However, post-hoc analysis of variance for each vowel pair showed that the differences were significant only when comparing /æ/ and /ʌ/. Thus, it would seem that the corrections made to H1, H2, and A3 for vowel quality (see Section 3.3.2) were largely successful in minimizing differences across vowels. However there may be some effects of vocal-tract configuration on the glottal waveform that would lead to differences across vowels (Bickley and Stevens, 1986, 1987).

Table 3.6 shows Pearson product moment correlation coefficients for the various measures for each vowel, while Table 3.7 shows the correlation coefficients for the three vowels combined. In the following discussion we consider a correlation with r greater than or equal to 0.70 to be strong. The strongest correlation was found between the high-frequency noise ratings and the tilt measure, H1* – A3*. As mentioned earlier, this is not unexpected given that both tilt and noise are expected to increase with the area of a fixed glottal opening (see Table 3.1 and the discussion in Section 3.2.2). H1* – A1 also has a strong correlation with the spectra-based noise ratings. Again, this is predicted from earlier discussion (see Table 3.1 where B1 increases with A_{ch}). For the vowels /ʌ/ and /ɛ/, H1* – A3* is well correlated with H1* – A1, but the correlation is only moderate for /æ/. Finally, the correlation between H1* – A1 and estimated F1 bandwidth for /æ/ is moderate.

It is striking that H1* – H2* is not well correlated with any other measure ($r < 0.59$). One might expect a larger open quotient to lead to greater losses and noise due to an increase in average glottal area. Although one might interpret this to mean that H1* – H2* is not a good measure of open quotient, Holmberg et al. (in press) have found H1* – H2* to be well correlated with open quotient in simultaneous observations of airflow and acoustic spectra for female speakers. Therefore it may be that open quotient is nearly independent of other glottal parameters. For example, a speaker may adjust her glottal configuration in such a way that a larger open quotient results while rate of decrease of flow at glottal closure remains nearly the same. Thus H1* – H2* increases, but the tilt may stay nearly the same, changing only a small amount due to a change in the skewness of the glottal pulse (speed quotient).

For the combined vowels, the noise measures are strongly correlated ($r > 0.70$) with the tilt measure, and the spectra-based noise measure is strongly correlated with the H1* – A1 (BW) measure. In addition, H1* – A1 has a fairly good correlation ($r = 0.68$) with the tilt measure H1* – A3*.

Table 3.6: *Pearson product moment correlation coefficients (r) for the various acoustic measures for each of the three vowels /æ, ʌ, ɛ/. Numbers in boldface represent strong correlations (r >* 0.70*). The notation* n.s. *indicates that a correlation was not significant.*

/æ/	H1* – H2*	H1* – A1	H1* – A3*	N_w	N_s	B1
H1* – H2*	1					
H1* – A1	0.47	1				
H1* – A3*	*n.s.*	0.62	1			
$\mathrm{N_W}$	*n.s.*	0.67	**0.87**	1		
$\mathrm{N_S}$	0.38	**0.72**	**0.82**	**0.88**	1	
B1	*n.s.*	0.61	*n.s.*	0.45	*n.s.*	1

/ʌ/	H1* – H2*	H1* – A1	H1* – A3*	N_w	N_s
H1* – H2*	1				
H1* – A1	0.57	1			
H1* – A3*	0.51	**0.78**	1		
$\mathrm{N_W}$	*n.s.*	0.56	**0.81**	1	
$\mathrm{N_S}$	*n.s.*	**0.75**	**0.84**	**0.83**	1

/ɛ/	H1* – H2*	H1* – A1	H1* – A3*	N_w	N_s
H1* – H2*	1				
H1* – A1	0.59	1			
H1* – A3*	0.49	**0.70**			
$\mathrm{N_W}$	0.45	**0.71**	**0.82**	1	
$\mathrm{N_S}$	0.48	**0.73**	**0.79**	**0.94**	1

Table 3.7: *Pearson product moment correlation coefficients (r) for the various acoustic measures for the three vowels /æ, ʌ, ɛ/ combined. Numbers in boldface represent strong correlations (r >* 0.70*).*

	H1 – H2**	*H1* – A1*	*H1* – A3**	N_w	N_s
H1* – H2*	1				
H1* – A1	0.53	1			
H1* – A3*	0.46	0.68	1		
N_W	0.30	0.63	**0.80**	1	
N_S	0.40	**0.73**	**0.80**	**0.86**	1

3.3.3.3 Interpretation of acoustic measurements

In order to gain a better understanding of the correlations reported in Table 3.7, and to perhaps be able to interpret the acoustic measurements in terms of glottal configurations, we examined scatterplots of measures that were well correlated with each other.

Figure 3.12(a) plots H1* – A3* against H1* – A1. Almost all of the data points with H1* – A1 less than about –6 dB have an H1* – A3* measure less than about 23 dB, while all of the data points with H1* – A1 greater than about –2 dB have an H1* – A3* measure greater than about 23 dB. Note that the highest H1* – A3* measure expected for speakers with a posterior glottal opening and simultaneous closure of the membranous part of the folds is about 25 dB (see Section 3.2.2.2). Based on this observation, we divided the data points into two groups, depending on whether H1* – A3* was less than or equal to 23 dB (Group 1) or greater than 23 dB (Group 2). Analysis of the two groups revealed that for 19 speakers, all three data points fell into either one group or the other, but not both. Data points for the other three speakers (**F10**, **F12**, **F17**) fell into both groups. Because subjects **F10** and **F12** had only one point each in Group 1, they were assigned to Group 2. Speaker **F17** had two points in Group 1, so she was assigned to that group.

Figure 3.12(b) shows a second version of Fig. 3.12(a) where data points for Group 1 speakers are represented by closed circles and those for Group 2 are represented by open circles. From Fig. 3.12(b), we see that the 11 speakers in Group 1 have relatively low

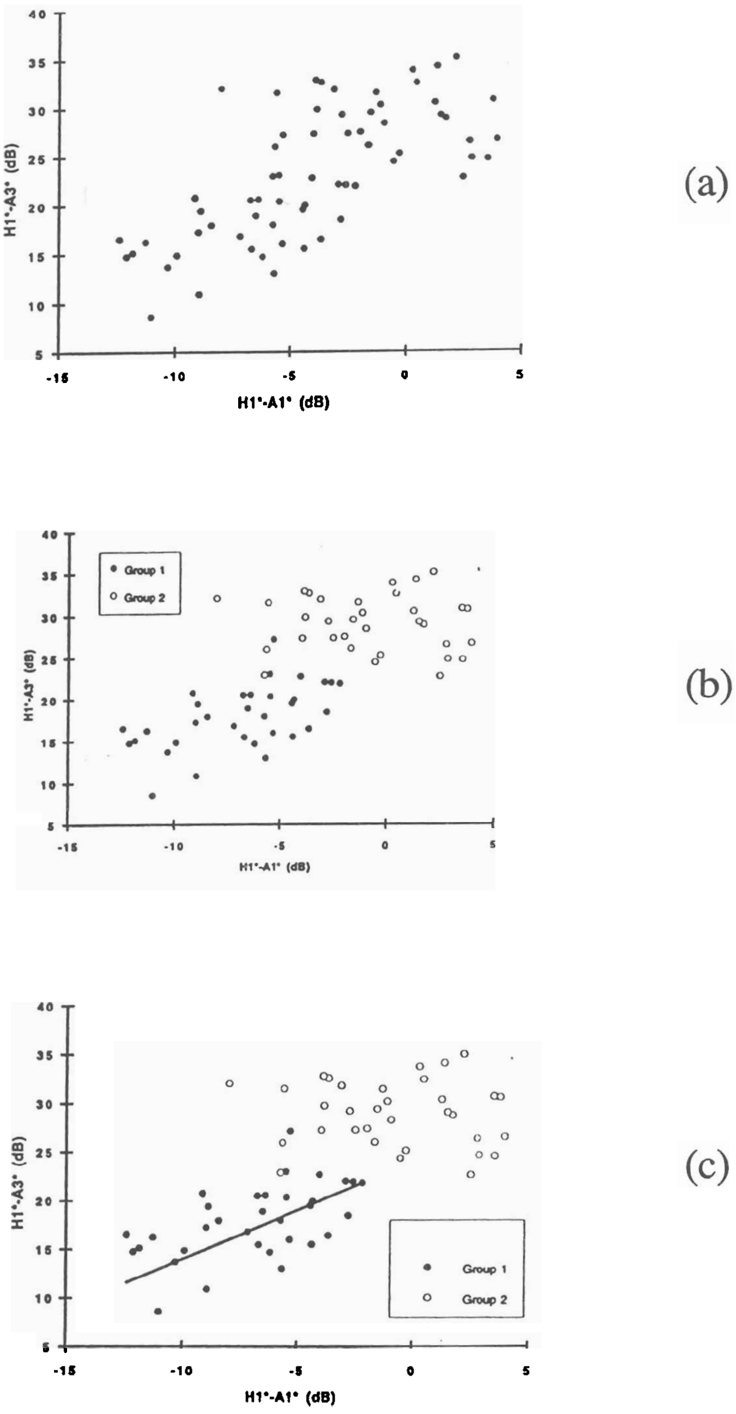


Figure 3.12: *(a) Relation between H1*− A3* and H1*− A1. (b) Same as (a), but data points for Group 1 are displayed as closed circles and data points for Group 2 are displayed as open circles (see text). (c) A line of slope one has been drawn through the data points for Group 1, showing the theoretically predicted relationship between spectral tilt and the amplitude of the first formant.*

values of H1* – A3* and H1* – A1. That is, speakers in this group have shallow spectral tilts and prominent first-formant peaks. Therefore, this group can be hypothesized to have abrupt glottal closures. Some speakers may also have posterior glottal chinks, which would account for the range of H1* – A3* (about 15 dB) and H1* – A1 (about 11 dB) that is present.

Speakers in Group 2, indicated by open circles, have much higher values of H1* – A3*, that is, steeper spectral tilts. From these values, we surmise that the glottal closure is not simultaneous along the length of the membranous part of the vocal folds. This nonsimultaneous closure is probably due to the glottis being spread at the vocal processes, although the folds could also close nonabruptly when the vocal processes are approximated. The higher values of H1* – A1 for Group 2 speakers are due to two influences on A1: (1) the first formant has an increased bandwidth because there are greater losses associated with the glottal configuration in which the vocal processes are spread, and (2) the spectral tilt is so steep that its influence extends down into the first-formant range. There is no upward trend between H1* – A1 and H1* – A3* for Group 2. This may be because for these speakers, the source spectral tilt and the prominence of the first-formant peak are influenced by both posterior glottal opening and nonsimultaneous closure, but the effect of the nonsimultaneous closure is independant of the effect of the posterior glottal opening.

From Table 3.1 we see that if the bandwidth of the first formant (B1) is expressed on a log (dB) scale, then B1 and H1* – A3* should increase together with a slope of 1 for speakers who have abrupt glottal closure. In Fig. 3.12(c) a line with slope 1 has been drawn through the data and is seen to fit nicely with the Group 1 points. This result is evidence that Group 1 speakers have abrupt glottal closure and posterior glottal openings that range in size across speakers.

Figure 3.13 shows the relation between the two types of noise judgments and the tilt parameter H1* – A3*. Recall that there was a high correlation between these quantities. This figure is also divided into the two groups of speakers of the previous figures. Speakers with greater degrees of tilt show greater amounts of noise in their speech signals, as predicted from the theoretical discussion earlier in this chapter. From Fig. 3.11, we see that noise ratings of 2 and 3 correspond to harmonics-to-noise ratios of about 2 and

-10 dB, respectively. For about half of our female speakers, then, the harmonics-to-noise ratio in the third-formant range was greater than 2 dB. A regression line ($r^2 = 0.62$) has been drawn through the points in Fig. 3.13.

In Fig. 3.14 the parameter H1* – A1 is plotted against F1 bandwidth (on a log scale) as measured in the first part of the glottal cycle for the 22 speakers producing the vowel /æ/. The data are presented to indicate which points belong to Group 1 and Group 2 speakers. A line of slope 1 is drawn through the data to represent the relationship expected based on the theoretical development. There seems to be a trend toward a decrease in F1 prominence (that is, a decrease in A1) as the F1 bandwidth increases, but the correlation is only moderate ($r = 0.61$, $p < 0.01$). The relatively weak correlation may be due to the fact that the prominence of A1 depends on the entire glottal cycle, whereas the bandwidth measure is based only on the closed (or minimum glottal area) part of the glottal cycle. Thus, A1 is influenced by the open quotient and the glottal aperture during the open phase, but the F1 bandwidth measure is not. In addition, other factors, such as spectral tilt, may reduce A1. In fact, given these influences, it is not surprising that the Group 1 data in Fig. 3.14 appears to be better correlated than the Group 2 data.

For one speaker (**F13**) the bandwidth is sufficiently small (53 Hz) that complete glottal closure can be assumed during a portion of the glottal cycle. This speaker is from Group 1. For speakers with higher bandwidth and H1* – A1 measures, it is reasonable to assume that the source of loss is an incomplete glottal closure. Two speakers from Group 2 (**F3** and **F8**) have fairly narrow bandwidths (94 and 97 Hz), although this would not be expected given our hypothesis that Group 2 members have abduction at the vocal processes. The H1* – A1 measure for these speakers indicates that A1 is indeed quite prominent, consistent with the narrow bandwidth. The findings for these speakers may indicate that their glottal closure is characterized by adducted vocal processes with no posterior glottal chink, but nonsimultaneous closure within the membranous portion. This interpretation might explain the narrow first-formant bandwidths, and consequently, high first-formant amplitudes, and steep spectral tilts that these two speakers exhibit.

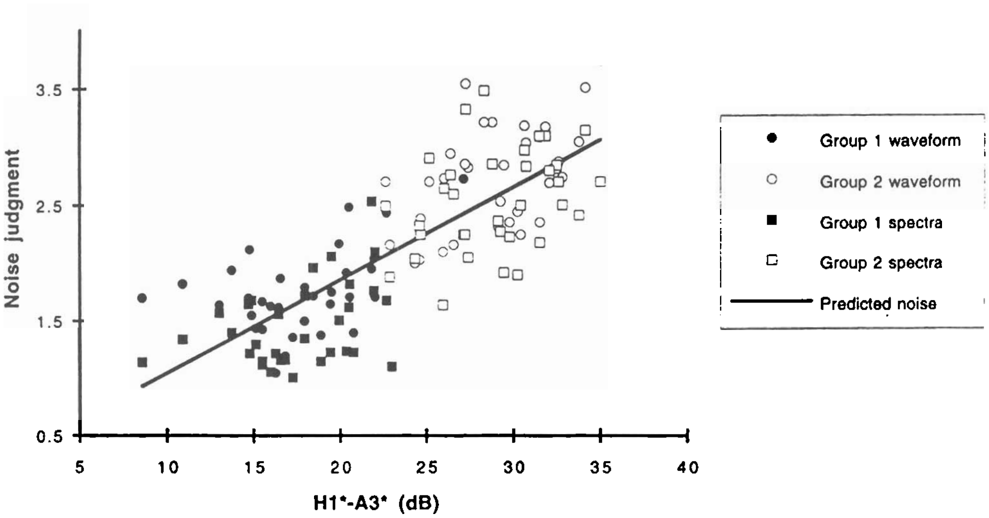


Figure 3.13: *Relation between noise judgments and H1*− A3*, together with a regression line (r^2 = 0.62). Points represented as circles are judgments based on waveforms and the squares are based on spectra. Closed points represent Group 1 data, while open points represent Group 2 data.*

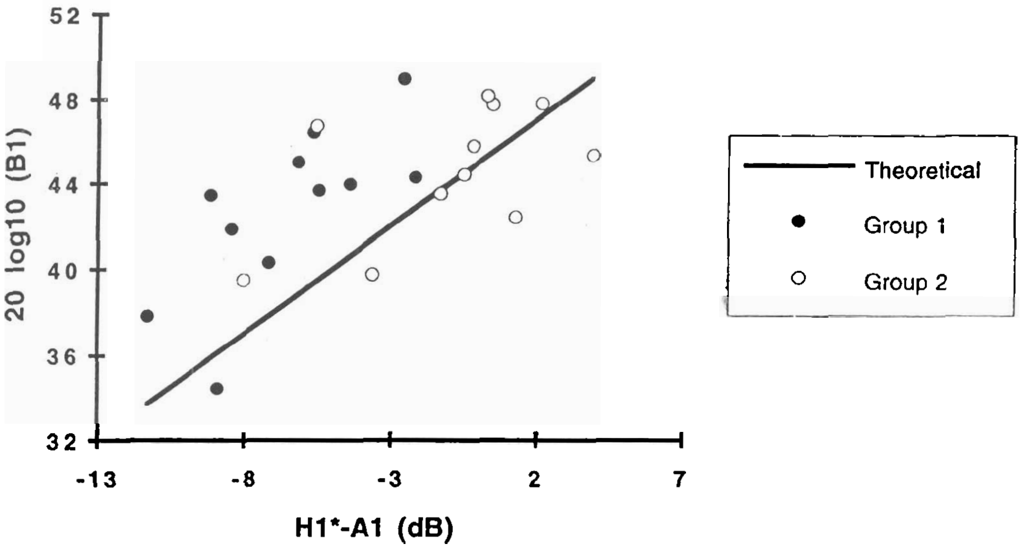


Figure 3.14: *Relation between H1* − A1 and F1 bandwidth (on a log scale) as measured from the waveform. The data are from speakers producing the vowel /æ/. Data points for Group 1 members are represented by closed circles, while those for Group 2 members are represented by open circles. A straight line representing the theoretical relationship has been drawn through the data.*

3.4 Summary

In the earlier part of this chapter we gave theoretical background describing how glottal characteristics may be manifested in the speech spectrum or waveform. As a result of this theoretical development, we suggested several measures to be made on the spectrum and waveform that might be suitable for obtaining glottal parameters. We also predicted how some of these measures might be related, and gave ranges of values that might be expected in natural speech of females. These measures were then used to analyze the steady state portion of vowels excised from the speech of 22 female subjects.

The results show substantial individual differences in several of the parameters. These differences are in line with the ranges that were predicted in the theoretical development. In particular, minimum values of the tilt measure H1* – A3* and the waveform-based bandwidth measure B1 are very close to those predicted. The maximum value of B1 is close to that derived from minimum (DC) airflow measures that have been reported (Holmberg et al., 1994), and the maximum value of H1* – A3* measured seems reasonable given our earlier discussion. The range of values obtained for the spectrum-based bandwidth measure H1* – A1 is the range that was predicted, and the minimum and maximum values are within 1 dB of those predicted. In addition, several of the acoustic measures are correlated as predicted from theory. The tilt measure H1* – A3* and the noise ratings N_W and N_S are strongly correlated. H1* – A3* is also relatively strongly correlated with one of the first-formant bandwidth measures, H1* – A1, and the noise ratings also tend to have a good to strong correlation with H1* – A1.

Using the acoustic measures, we were able to divide the 22 subjects into two hypothetical groups. Group 1, with 11 speakers, is hypothesized to have abrupt glottal closure. Based on the measure B1, one speaker in this group seems to have complete closure during some part of the glottal cycle. The other speakers have larger B1 values, and thus are thought to have some losses at the glottis due to glottal chinks. The ranges of values obtained for the two bandwidth measures, the tilt measure, and the noise ratings, suggest that the glottal losses, and thus the size of these glottal chinks, vary from subject to subject. In Section 3.2.2.2 we suggested that 16 dB might be a maximum value expected for additional tilt due to a glottal chink, and, in fact, the additional tilt observed for speakers at the extreme for this group is about 15 dB. The maximum B1 that would be predicted given this amount of additional tilt is about 225 Hz (see Table 3.1), while the maximum B1 measured for this group is about 210 Hz.

Group 2 also includes 11 speakers, and due to their higher values of additional tilt, we assume that these speakers have both glottal chinks and nonsimultaneous closure of the membranous part of the folds. The generally higher B1 measures suggest greater losses at the glottis, probably due to a fixed opening that extends to the vocal processes, which would cause the nonsimultaneous closure. However, two members of this group have fairly narrow first-formant bandwidths and lower $H1^* - A1$ measures, suggesting that these two speakers may have a glottal configuration consisting of approximated vocal processes, nonsimultaneous closure, and, possibly, a glottal chink.

Our results are satisfying in that the ranges of observed values and the relationships between these values are in line with the predictions based on our theoretical development. However, these results and our interpretation of the data have raised additional questions, prompting further investigation. First, we have made hypotheses about the glottal configurations of our subjects, splitting them into two groups. The question arises as to how valid this classification is. In an attempt to answer this question, we have performed physiological measures on a subset of the subjects. These measures include glottal waveform parameters obtained by inverse filtering of vocal tract airflow, and observation of the vocal folds during phonation, via fiberscopy. This experiment and its results are reported in Chapter 4. Second, the hypothesized difference in vocal fold configuration would predict that members of Group 2 have a breathier voice quality than do members of Group 1. We have performed a listening test to investigate this possibility. This test is described in Chapter 5.

Finally, the wide ranges of parameter values that we have observed suggest that consideration of glottal characteristics has great importance for describing female speech and, in addition to formant frequencies and fundamental frequency, should be taken into account for applications such as synthesis and recognition of speech and speakers. We have performed a synthesis experiment using our measures of glottal characteristics to guide the synthesis of the vowels /ʌ, ɛ/ of six of our speakers. The success of this synthesis was judged by a number of subjects in a listening test. This experiment and the results are also presented in Chapter 5.

Chapter 4

Physiological measures

4.1 Introduction

In Chapter 3 we made acoustic measurements on the speech waveforms and spectra of a group of 22 female speakers, and from these measurements we made hypotheses about their glottal configurations and waveforms. In this chapter we turn to more direct, physiological measures of glottal characteristics in order to gain some insight into the acoustic measurements and, perhaps, validate our hypotheses. One method is based on oral airflow and intraoral pressure. These are measured during speech production via a Rothenberg mask (Rothenberg, 1973), shown earlier in Fig. 2.5. The glottal waveform is obtained by inverse filtering of the oral airflow measured during phonation; that is, the effects of the formants are removed, and glottal parameters can be extracted from this waveform and its derivative. Figure 4.1 shows a schematic of a glottal waveform and its derivative. Glottal waveform parameters that are of special interest are illustrated. In the second method, a fiberscope is inserted through the nasal cavity and positioned above the vocal folds so that the folds can be observed during phonation. The fiberscope system is schematicized in Fig. 2.6. As we discussed in Chapter 2, these two methods are well established and have been used in many studies to measure characteristics of vocal-fold vibration (see, for example, Karlsson, 1986, 1988; Holmberg et al., 1988, in press; Gauffin and Sundberg, 1989; Södersten and Lindestad, 1990; Kiritani et al., 1990).

Our subjects for this additional analysis came from both groups of speakers, those assumed to have abrupt glottal closure and those assumed to have nonsimultaneous closure. Based on these groupings, we had some expectations about the results. For one, we ex-

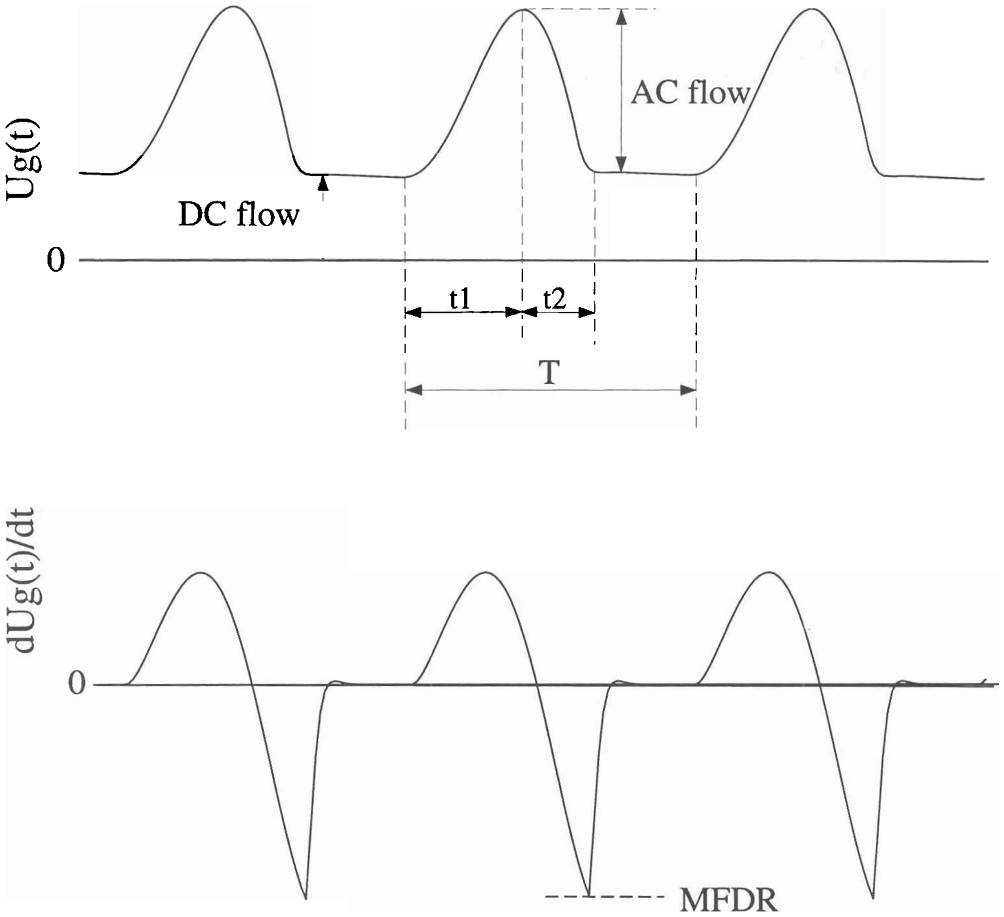


Figure 4.1: *Schematic of a glottal waveform $U_g(t)$, and its derivative dU_g/dt, synthesized using the KLSYN88 formant synthesizer (Klatt and Klatt, 1988). The glottal parameters AC flow, DC flow, MFDR, and the pitch period T are indicated. Speed quotient is defined as* $t1/t2$ *(ratio of rise time to fall time), and open quotient is defined as* $(t1+t2)/T$ *(ratio of open time to pitch period).*

pected that the airflow measures might show that Group 2 speakers have higher minimum (DC) flows than Group 1 speakers, due to larger openings at the glottis. First-formant bandwidth can be estimated from the minimum (DC) flow and the transglottal pressure (see Eqns. 3.4 and 3.8–3.10), and if the acoustic measures actually reflect glottal configurations, this estimated bandwidth should be close to the bandwidth measured from the acoustic sound pressure.

Other studies of glottal waveforms derived from oral airflow signals have related certain measures made on these waveforms to spectral measures (Gauffin and Sundberg, 1989; Fant et al., 1994). For example, an increase in AC flow leads to an increase in amplitude of the first harmonic (H1); and also, a higher maximum flow declination rate (MFDR) will correspond to greater sound pressure level (SPL) and thus a higher first-formant amplitude in the spectrum (A1).

With the fiberscopy we expected to see that the Group 1 speakers would have smaller glottal chinks relative to those of Group 2 speakers, and that the Group 2 speakers would have nonsimultaneous closure. One of the Group 2 speakers chosen is a subject who had a narrow first-formant bandwidth, despite having a large spectral tilt. We expected that she might have approximated vocal processes, despite the nonsimultaneous closnre.

As discussed in Chapter 2, the aerodynamic method has shortcomings (Holmberg et al., in press). For one, unless a tight seal between the mask and the subject's face is achieved, mask leak will occur, with the result that minimum (DC) flow will appear to be less than it is. On the other hand, Hertegård et al. (1992) found that even with complete glottal closure during the phonatory cycle, some speakers show minimum flows up to 90 cm^3/sec (depending on F0), perhaps due to vertical movements of the folds during phonation. This offset can *inflate* the values of minimum flow. Another problem is that failure to correctly filter out the first formant will result in F1 residual in the glottal waveform. This residual can interfere with the data extraction algorithms, leading to incorrect measures of open quotient. Finally, the required lowpass filtering of the flow at 1100 Hz (see below) means that information about abruptness of closure may be lost, because this information is present at mid- to high frequencies in the glottal source spectrum, as discussed in Section 3.2.1.

The outline of this chapter is as follows. We first describe the subjects and the speech material. Next, the method and results for each of the three types of analysis (acoustic, aerodynamic, and fiberscopic) are presented. Finally we discuss the results of the three analyses and give our conclusions.

4.2 Subjects and speech material

Our choice of subjects for this further analysis was limited to those of the original 22 who were available and willing to participate. Four subjects were chosen, two from Group 1 (**F9** and **F14**) and two from Group 2 (**F3** and **F5**). We chose subjects who were well separated according to the acoustic measures described in Chapter 3, for example, the Group 2 speakers, **F3** and **F5**, had very high tilt measures (H1* – A3*) compared to the Group 1 speakers, **F9** and **F14**. In addition, speaker **F3** was of particular interest because she had the somewhat unusual combination of a high tilt measure (H1* – A3*) and narrow first-formant bandwidth estimates (B1 and H1* – A1).

We collected data using two types of utterances. One group of utterances was comprised of the vowels /æ, ʌ, ɛ/ embedded in the carrier phrases "may bVb again" and "may pVp again." These phrases were chosen to avoid coronal speech segments, which might interfere with the collection of the iutraoral pressure signal. We were primarily interested in the /bVb/ context because we used the voiced stop environment in our earlier acoustic analysis (cf. Section 3.3), and we wanted to compare those results with the results of the aerodynamic analysis. However, in the aerodynamic analysis, subglottal pressure is estimated from the intraoral pressure measured during the stop occlusions on either side of the vowel. This procedure can only be assumed to be valid for a voiceless stop, when the vocal folds are fully spread, so we included the /p/ as well as the /b/ context. The other type of utterance was a string of five repetitions of the syllable /pVː/, where V is again one of the vowels /æ, ʌ, ɛ/. Holmberg et al. (1988, 1994a, 1994b, in press) and Perkell et al. (1994) used the latter type of utterance with the vowel /æ/ for their aerodynamic measures, and consequently we can compare our results with theirs. The two types of utterances were mixed so that each carrier phrase was followed by a syllable string with the same vowel. These blocks of a carrier phrase followed by a syllable string were each

Table 4.1: *Average acoustic data for three vowels (/æ, ʌ, ɛ/) for four female speakers. Data in columns labelled 'R1' are from the first recording, previously reported in Chapter 3. 'R2' refers to acoustic data collected in conjunction with the physiological data.*

	F14		F9		F3		F5	
Recording	R1	R2	R1	R2	R1	R2	R1	R2
H1* – H2* (dB)	4.5	1.6	2.4	−2.0	3.7	−1.5	4.8	−3.0
H1* – A1 (dB)	−11.9	−11.5	−6.0	−10.8	−4.3	−9.6	2.5	−7.9
H1* – A3* (dB)	16.0	17.2	19.1	9.0	30.0	28.8	31.6	21.6
B1 estimate (/æ/) (Hz)	78	89	104	116	94	82	245	222
F0 (Hz)	196	195	192	193	201	226	178	204

recorded five times, in random order. Thus, for each vowel we obtained five tokens each of /bVb/ and /pVp/, and 10 syllable strings.

The recordings were made 12 months after the first recording for speakers **F3** and **F9**; 14 months later for speaker **F14**; and 18 months later for speaker **F5**. Our procedure was to collect all the data from each speaker in one day. First, the speech was recorded in a sound-proof room, to be used for the acoustic analysis. Next, the aerodynamic data were collected. Finally, the fiberscopy was performed. While the three types of data were not collected simultaneously, the results should be closely related, because the recordings were completed within a short time period of each other in one day. We now discuss in more detail the collection of these three types of data, and the results.

4.3 Acoustic measures

Although acoustic data were collected for both kinds of utterances, only the vowels embedded in the carrier phrase "may bVb again" were analyzed. The analysis techniques and the methods for normalizing the data were the same as in Chapter 3, except that noise ratings were not collected. The results are presented in Table 4.1, along with the earlier results of Chapter 3, where the values have been averaged across vowels. In addition to the spectral measures, we include fundamental frequency (F0). These data are also presented graphically in Fig. 4.2, where each graph represents one of the acoustic measures.

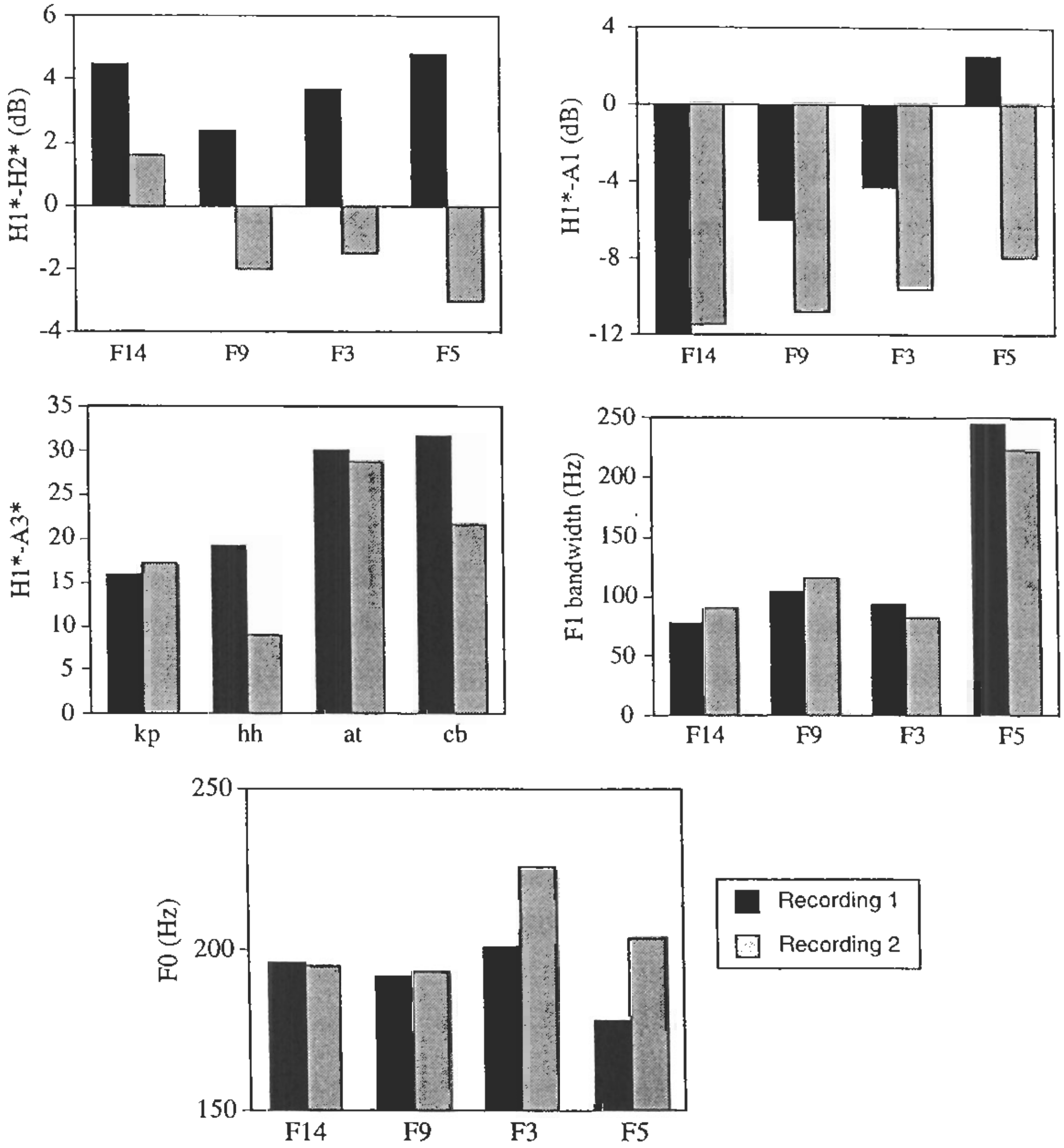


Figure 4.2: *Graphical representation of acoustic data extracted from syllable strings and given in Table 4.1. Each graph represents one acoustic measure. Average values for each of the two recording sessions are indicated for four female speakers. Data in black-colored columns are from the first recording (previously reported in Chapter 3). Light grey-colored columns represent acoustic data collected in conjunction with the physiological data.*

For all speakers, the bandwidth estimated from the waveform, B1, is about the same for both recordings. All of the subjects have smaller open quotients ($H1^* - H2^*$) for the second recording, suggesting a more pressed phonation (Fant et al., 1994), so it is possible that they spoke more loudly than they had during the original recordings. For the remaining measures, one of the speakers, **F14**, has results quite similar to those presented in Chapter 3, but the other subjects show some differences. These three subjects have more prominent first formant peaks, that is, $H1^* - A1$ is reduced. This reduction may be related to the smaller open quotient, which reduces both H1 and the extra damping that occurs during the open part of the glottal cycle. The latter reduction can boost the amplitude of the first-formant peak. An increase in subglottal pressure P_s may also result in an increased first-formant peak (Fant, 1993; Gauffin and Sundberg, 1989), and thus a reduced $H1^* - A1$.

Two of the speakers (**F3** and **F5**) show substantial increases in fundamental frequency (F0), further suggesting that they were speaking more loudly because most speakers increase F0 when they increase SPL. Speakers **F9** and **F5**, have reduced $H1^* - A3^*$ measures, possibly due to several factors, one of which could be a smaller opening at the glottis (see Section 3.2.2.2). An increased subglottal pressure could also reduce the tilt caused by a chink at the glottis (cf. Eqn. 3.16). Therefore, for these two speakers an increase in P_s could be the source of the reduction in both $H1^* - A1$ and $H1^* - A3^*$.

The implication of this acoustic analysis is that the subjects show variation in their acoustic measures across recording sessions that may be due to changes in glottal configuration. Thus, in our analysis of the physiological data, we may not see the patterns that we had predicted based on the acoustic analysis of Chapter 3. In particular, speaker **F5** shows such a large drop in the tilt measure $H1^* - A3^*$ and the bandwidth measure $H1^* - A1$ for the second set of recordings that she would be assigned to Group 1 according to the criterion used to classify the speakers in Chapter 3.

4.4 Glottal airflow measures

4.4.1 Data collection and parameter extraction

Our method for measuring the oral airflow and intraoral pressure, and for extracting the glottal characteristics from these signals was that used by Holmberg et al. (1988) and Perkell et al. (1991, 1994). We will give only an overview here. Briefly, the data were collected in a sound-proof room, and oral airflow and intraoral pressure were measured during speech production via a Rothenberg mask (Rothenberg, 1973), modified to include a pressure transducer, and recorded onto DAT tape. The airflow signal is lowpass-filtered at 1100 Hz because the frequency response of the mask is flat only up to about that frequency. The acoustic sound pressure was recorded simultaneously with the airflow and pressure. Calibration signals for oral airflow, intraoral pressure, and sound pressure level were also collected and recorded. The data were then digitized and subjected to various signal processing techniques, following which the signals could be viewed. The oral airflow signal was inverse filtered to obtain the glottal waveform, and glottal characteristics were extracted interactively from the glottal waveform.

Acoustic and aerodynamic parameters were extracted from the signals. The acoustic parameters were SPL (dB) and F0 (Hz), and the aerodynamic parameters were intraoral pressure (cm H_2O), AC flow (cm^3/sec), minimum (DC) flow (cm^3/sec), maximum flow declination rate (MFDR) (ℓ/sec^2), open quotient (open time/T, where T is the pitch period), and speed quotient (rise time/fall time). Some of these measures are illustrated in Fig. 4.1.

For the vowels embedded in carrier phrases, the parameters were measured from four consecutive pitch periods at the center of the vowel. For the syllable strings, the parameters were extracted from the third syllable, again using four pitch periods at the midportion of the vowel. The parameters from the four pitch periods were averaged. For two of the speakers (**F3** and **F5**) some of the data for certain tokens had to be discarded. In some cases this was due to evidence of mask leak (for example, negative flow signals). In other cases, the vowels in the carrier phrases, particularly the phrase with /pʌp/, had very short durations and the effects of the CV and VC transitions resulted in poor performance of the data extraction algorithms.

Table 4.2: *Pearson product moment correlation coefficients (r) between MFDR, and SPL and AC flow, from the aerodynamic data for four female speakers. The correlations are significant ($p < 0.025$).*

		F14	F9	F3	F5
Syllable strings:	*MFDR v. SPL*	0.73	0.64	0.74	0.83
	MFDR v. AC flow	0.45	0.77	0.50	0.70
Carrier phrases:	*MFDR v. SPL*	0.69	0.81	0.69	0.90
	MFDR v. AC flow	0.85	0.67	0.62	0.80

4.4.2 Statistical analysis

In a previous study, Holmberg et al. (1988) found several of the measured parameters to be well correlated. Thus, Pearson product moment correlation coefficients were computed for each speaker for all pairs of parameters. These correlations are summarized in Table 4.2. Consistent with Holmberg et al. (1988), we found moderate to strong[1] relationships between MFDR and SPL in both speech conditions. An explanation for this correlation is that the MFDR represents the main excitation of the vocal tract, thus being the primary determinant of the amplitude of the formants (Fant et al., 1994). The amplitude of F1 is about the same as the SPL, except for very soft voice (Gauffin aud Sundberg, 1989). Thus, MFDR is expected to be correlated with SPL. Some of the speakers also show strong correlations between MFDR and AC flow, as may be expected for changes in amplitude of a periodic waveform.

In pilot work, Holmberg and her colleagues found that amplitudes of glottal airflow measures extracted from vowels in carrier phrases had higher values than those extracted from syllable strings.[2] Therefore, the two groups of data will be reported and discussed separately in the next section. However, it was not certain whether vowel quality and consonant context (/p/ vs. /b/) would have any effect, because previous analyses using this method have relied on one vowel in syllable-string or sustained-vowel speaking conditions (see, for example, Karlsson, 1986, 1988; Holmberg et al., 1988, in press; Gauffin and

[1] We define a moderate correlation to be one with $0.60 \leq r < 0.70$, and a strong correlation to be one with $r \geq 0.70$.

[2] Holmberg, personal communication.

Sundberg, 1989). For each speaker, then, analysis of variance was performed on each parameter to determine if the effects of vowel quality and (in the case of the vowels embedded in carrier phrases) consonant context were significant. For the vowels in carrier phrases, intraoral pressure was excluded from the analysis because it was only measured for the /p/ context. A significance level of $p < 0.0016$ was used for the vowels in carrier phrases, based on the Bonferroni/Dunn correction for multiple univariate comparisons, leading to an overall significance level of $p < 0.01$ (because $0.01/(6 \text{ variables}) \approx 0.0016$). A level of $p < 0.0014$ was used for the vowels in syllable strings, again based on the Bonferroni/Dunn correction. This resulted in an overall significance level of $p < 0.01$ ($0.01/(7 \text{ variables}) \approx 0.0014$).

The results for the analysis of variance for the vowels in syllable strings showed almost no vowel effects. For one speaker, **F9**, speed quotient showed an effect, although a post-hoc analysis showed that this was only for /æ/ compared to /ɛ/ and the size of the difference was comparable to those shown by the other three speakers. Another speaker, **F5**, showed a vowel effect for pressure. In post-hoc comparisons, /æ/ was different from /ʌ, ɛ/ and again, this pattern was also evident for the other speakers. Because these vowel effects were not consistent, it was decided that it was safe to average the results across vowels.

The analysis of variance results were somewhat more complex for the vowels embedded in carrier phrases. Two speakers (**F14** and **F5**) showed significant consonant context (/p, b/) effects for the variables MFDR, SPL, and AC flow. The strong correlations between these parameters that these two subjects show (see Table 4.2) probably explain why all three simultaneously show the effect. One other speaker, **F9**, also showed a context effect for AC flow and **F14** had a vowel effect for MFDR (/æ/ was significantly different from /ʌ, ɛ/). No interaction was found between vowel and consonant context.

The results of the analysis of variance suggest that the neighboring consonants affected the glottal vibratory pattern during vowel production. This finding is not surprising given that the glottal configuration is expected to be more abducted for /p/ production than for /b/, and is in agreement with earlier studies by Gobl and Ní Chasaide (1988) and Ní Chasaide and Gobl (1993). In their study of 'CVCV nonsense utterances they found that for some speakers of English, a vowel preceding a voiceless stop becomes increasingly

Table 4.3: *Data from syllables. Means and standard deviations of glottal and acoustic measures extracted from aerodynamic data recorded from four female speakers. The speech materials were the vowels /æ, ʌ, ɛ/ in /pVː/ syllable strings. Also given are group means and standard deviations for 20 female speakers in normal voice, extracted from /pæː/ syllable strings (Holmberg et al., in press). Numbers in boldface are measures that fall outside the range of the Holmberg et al. (in press) mean plus or minus one standard deviation.*

Measure	*F14*	*F9*	*F3*	*F5*	*Holmberg et al.*
SPL (dB)	77 (1)	79 (1)	**82** (2)	77 (1)	75 (4)
Pressure (cm H_2O)	5.7 (0.5)	**7.2** (0.7)	**9.4** (1.1)	†6.6 (0.3)	5.5 (1.1)
DC flow (cm^3/sec)	69 (15)	45 (8)	100 (7)	106 (12)	97 (42)
AC flow (cm^3/sec)	149 (13)	147 (12)	**231** (27)	**229** (14)	147 (45)
Open quotient (%)	54 (7)	**43** (1)	**60** (8)	49 (5)	‡50 (6)
Speed quotient	**2.9** (0.6)	§2.1 (0.3)	**3.1** (0.7)	1.6 (.4)	(*unavailable*)
MFDR (ℓ/sec^2)	247 (33)	**272** (24)	**420** (69)	**366** (34)	191 (76)
F0 (Hz)	180	208	222	189	(*unavailable*)

†In analysis of variance, /æ/ is significantly different from /ʌ, ɛ/.
‡Holmberg et al. (in press) report adduction quotient, or 100 minus open quotient.
§In analysis of variance, /æ/ is significantly different from /ɛ/

breathy throughout the course of the vowel, suggesting that glottal abduction is anticipated very early in the vowel production, and that evidence of this anticipation included a weakening MFDR (excitation). They also found that at vowel onset the effect of voiceless stops was comparatively small, with full excitation being achieved almost immediately. The /p, b/ context effect that we found suggests that the measures involved should be looked at separately according to the voiced/voiceless context. In the next section, we will give these measures (MFDR, SPL, and AC flow) both separately and averaged across context.

4.4.3 Results

4.4.3.1 Syllable strings

The data for the syllable strings are summarized for the four speakers in Table 4.3 as mean values across vowels, along with the standard deviations. Measures that were averaged across vowel despite showing a significant difference in an analysis of variance are indicated. Included in this table are the group mean values and standard deviations reported in Holmberg et al. (in press) for 20 female speakers in normal voice. Numbers in boldface represent values that fall outside the range of the mean plus or minus one standard deviation of Holmberg et al. (in press). These data are perhaps more easily absorbed by viewing them in bar graph form, given in Fig. 4.3. Each graph represents one measure. Some correlations across speakers can be seen in these graphs, for example the positive relationship between SPL and subglottal pressure.

As expected, Group 2 speakers, **F3** and **F5**, have higher DC (minimum) flow values than do the Group 1 speakers, although they are average compared to the Holmberg et al. (in press) data. However, these two speakers displayed signs of mask leak, as mentioned in Section 4.4.1, and so it is possible that their minimum flow values should be higher than the measured values. The trend for them to have higher minimum (DC) flows than the Group 1 speakers suggests that these two speakers may have greater losses at the glottis due to larger glottal openings.

Three subjects, **F14**, **F9**, and **F5** have comparable values of SPL, while **F3**'s SPL is significantly higher. These results for SPL may seem counterintuitive given that **F5** and **F3** may have greater losses at the glottis. However, note that the Group 2 speakers have higher MFDRs, relative to those of the Group 1 speakers, and SPL has been found to be strongly correlated with MFDR (cf. Table 4.2, and Holmberg et al., 1988; Gauffin and Sundberg, 1989). Holmberg et al. (1994b) suggested that some speakers with glottal configurations that lead to greater losses at the glottis may raise subglottal pressure in an effort to increase SPL (Ladefoged, 1962; Isshiki, 1964; Gauffin and Sundberg, 1989), and consequently increase the speed of closure, which could lead to increased MFDR. Thus, it is possible that the relatively high values of MFDR for **F3** and **F5** indicate efforts by them to compensate for losses at the glottis. Together, their higher values of DC flow and MFDR might indicate breathy/hyperfunctional phonations. Continuing this line of reasoning, speaker **F3** has a considerably higher subglottal pressure and slightly higher speed quotient compared to the other speakers, which may further contribute to her very high SPL.

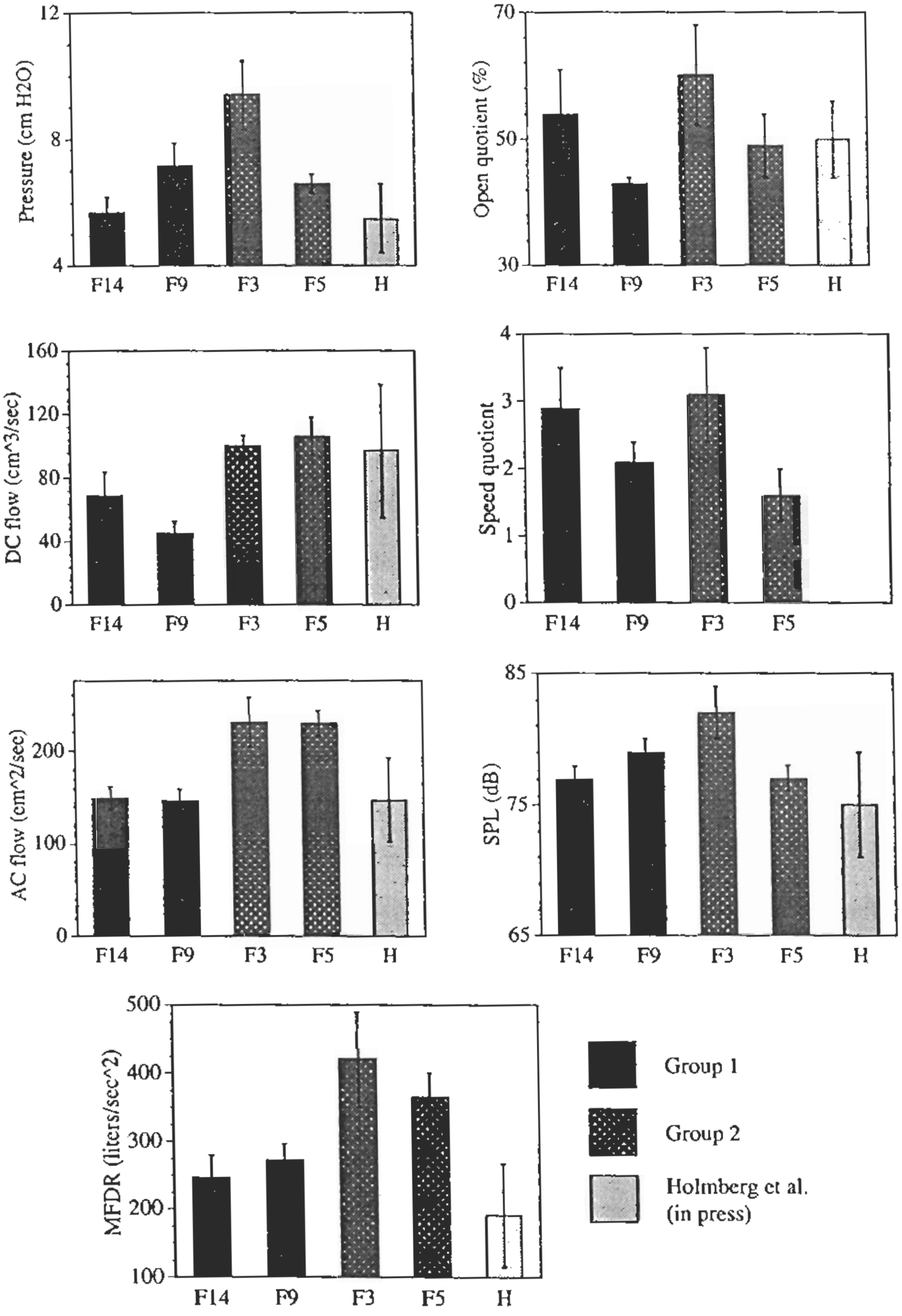


Figure 4.3: *Graphical representation of aerodynamic data extracted from syllable strings, given in Table 4.3. Each graph represents one aerodynamic parameter. Average values and standard deviations are indicated for four female speakers. Speakers* **F14** *and* **F3** *are from Group 1, while speakers* **F3** *and* **F5** *belong to Group 2. Data labeled with the letter* **H** *are from group data given in Holmberg et al. (in press).*

Gauffin and Sundberg (1989) found that pressed phonation resulted in a glottal waveform with a lower AC flow than would result from a phonation associated with a breathy voice quality; however, this analysis was limited to one speaker. As expected from their result, our Group 1 speakers, **F14** and **F9**, have smaller AC flows than **F3** and **F5**. Subjects **F14** and **F9** have higher values of speed quotient than **F5**, suggesting lower spectral tilt measures (see Section 3.2.1), not surprising for members of Group 1. However, **F3** has an even higher speed quotient, yet her tilt measures were found to be quite high compared to **F14** and **F9**. It is not clear how to explain this result, but **F3**'s generally high tilt measures may be due to nonsimultaneous glottal closure (as we hypothesized in Chapter 3) in spite of her higher speed quotient.

To summarize the results for this section, the Group 2 speakers, **F3** and **F5**, have higher minimum (DC) flows, MFDRs, and AC flows than the Group 1 speakers, **F9** and **F14**, the combination of which may indicate a phonation that might result in a breathy/hyperfunctional voice quality. Based on our acoustic analysis in Chapter 3, we had hypothesized that Group 2 speakers had glottal configurations that would result in breathier voice qualities when compared with Group 1 speakers.

4.4.3.2 Carrier phrases

Table 4.4 summarizes the aerodynamic data extracted from the vowels embedded in the carrier phrases for the four speakers. The acoustic data given in Table 4.1 are repeated for comparison. Again, mean values and standard deviations are presented. The data were averaged across vowels and consonant context, but for MFDR, SPL, and AC flow, tbe data are also given separately for both the /b/ and /p/ contexts, due to the results of the analysis of variance reported in Section 4.4.2. Those measures that were averaged despite showing a vowel or context effect are indicated. These data are also presented graphically in Fig. 4.4.

Values of the aerodynamic measures for these vowels tend to be slightly higher than for those in the syllable strings (cf. Table 4.3). One explanation may be that all four speakers used higher fundamental frequencies for the vowels in carrier phrases than for the vowels in syllable strings. F0 has been found to be positively correlated with SPL

Table 4.4: *Carrier phrase data. Means and standard deviations of glottal and acoustic measures extracted from aerodynamic data recorded from four female speakers. The speech material were the vowels /æ, ʌ, ɛ/ embedded in carrier phrases. Means for SPL, MFDR, and AC flow are given within and across context (/bVb/ vs. /pVp/). An estimate for first-formant bandwidth based on DC flow and intraoral pressure is included. Note that this bandwidth estimate assumes a vocal tract of uniform cross-sectional area. The acoustic data from Table 4.1 are repeated for convenience.*

	F14	**F9**	**F3**	**F5**
SPL (dB)	†84.8 (1.6)	83.0 (1.0)	88.2 (1.8)	†83.2 (1.2)
/bVb/	85.6 (1.6)	83.1 (0.8)	88.5 (2.0)	83.8 (1.0)
/pVp/	84.0 (1.3)	82.8 (1.2)	87.7 (1.6)	82.4 (0.9)
Pressure (cm H_2O)	5.5 (0.5)	6.3 (0.3)	10.3 (1.1)	7.4 (0.4)
DC flow (cm^3/sec)	99 (20)	62 (16)	126 (104)	130 (41)
AC flow (cm^3/sec)	†184 (27)	†153 (19)	286 (61)	†289 (45)
/bVb/	202 (25)	167 (13)	338 (37)	314 (37)
/pVp/	165 (15)	140 (11)	238 (34)	260 (25)
Open quotient (%)	45 (2)	42 (2)	57 (8)	42 (4)
Speed quotient	3.6 (0.7)	2.7 (0.4)	3.2 (1.0)	2.3 (0.8)
MFDR (ℓ/sec^2)	†,‡458 (105)	349 (49)	636 (98)	†566 (61)
/bVb/	503 (127)	364 (47)	676 (101)	599 (45)
/pVp/	412 (55)	333 (48)	593 (82)	532 (49)
F0 (Hz)	227	218	262	201
Flow-based B1 estimate (Hz)	126	98	109	141
H1* – H2* (dB)	1.6	−2.0	−1.5	−3.0
H1* – A1 (dB)	−11.5	−10.8	−9.6	−7.9
H1* – A3* (dB)	17.2	9.0	28.8	21.6
F0 (Hz)	195	193	226	204
B1 estimate (Hz)	89	116	82	222

†In analysis of variance, /bVb/ context is significantly different from /pVp/ context.
‡In analysis of variance, /æ/ is significantly different from /ʌ, ɛ/

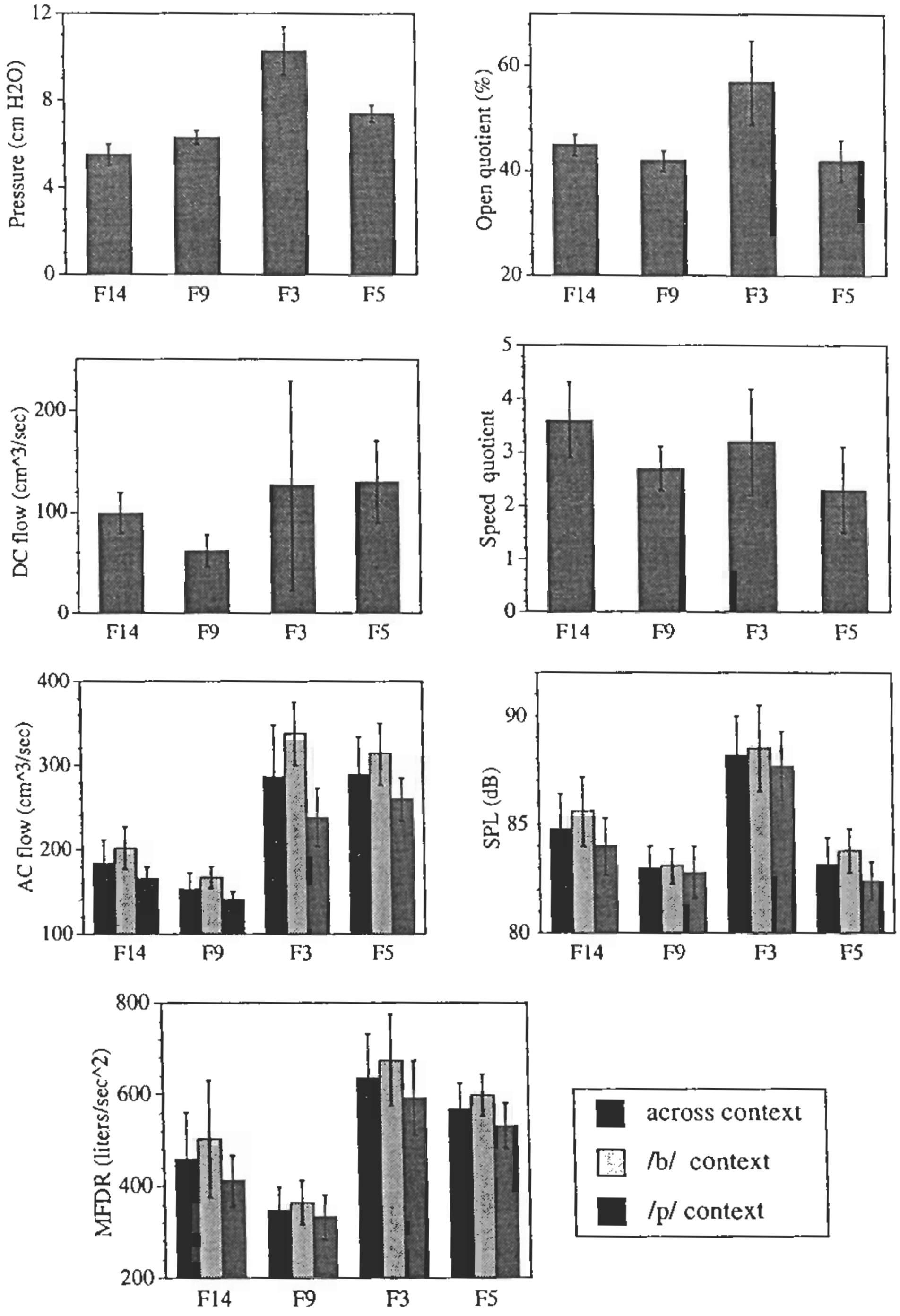


Figure 4.4: *Graphical representation of aerodynamic data extracted from vowels in carrier phrases, given in Table 4.4. Each graph represents one aerodynamic parameter. Average values and standard deviations are indicated for four female speakers. Speakers* **F14** *and* **F3** *are from Group 1, while speakers* **F3** *and* **F5** *belong to Group 2. For three measures (MFDR, SPL, and AC flow), the averages are given both across context* (/bVb/ vs. /pVp/) *and separately.*

(Gauffin and Sundberg, 1989) and MFDR and AC flow have also been found to increase with F0 (Holmberg et al., 1989; Fant et al., 1994).

Nonetheless, the carrier phrase vowel data show trends similar to those of the syllable strings: speaker **F3** has higher SPL and pressure values compared to those of the other three speakers, and the Group 2 speakers have greater values for AC flow, DC flow, and MFDR. As we discussed in the previous section, the latter result suggests the Group 2 speakers may have breathy/hyperfunctional phonations, while Group 1 speakers may have more pressed phonations. Speed quotient also shows a similar pattern to that observed for the syllable string data.

Subjects **F9** and **F5** have both the smallest open quotient and $H1^* - H2^*$ measures. Values for **F14** are slightly higher. Subject **F3** has an open quotient that is much greater than **F14**'s but her $H1^* - H2^*$ measure is about 3 dB less than **F14**'s. Again, it is unclear how to explain this result. However, if subglottal resonances are present, as they may be for a speaker with a large opening at the glottis, the acoustic measure of open quotient may be modified.

If the acoustic data accurately reflect the glottal configuration and the minimum flow data are not corrupted by mask leak, then the bandwidths estimated from the acoustic sound pressure should be in line with those predicted from the minimum flow and intraoral pressure data. These flow-based bandwidth estimates are given in Table 4.4. While those for **F9** and **F3** are within 20 Hz of those measured, the value for **F14** is higher by about 35 Hz and that for **F5** is lower by about 80 Hz. For the latter case, the error could be due to mask leak which leads to an underestimation of minimum flow.

The aerodynamic measures AC flow and MFDR are expected to be proportional to the amplitudes of the first harmonic H1 and the first formant A1, respectively (Fant, 1993; Fant et al., 1994; Gauffin and Sundberg, 1989). Therefore, the difference between these two measures in dB should be proportional to the acoustic measure $H1^* - A1$. Figure 4.5 shows a plot of the average values of $20\log_{10}(\text{AC flow/MFDR})$ for the vowels in /b/ context in relation to $H1^* - A1$ for the four speakers (the vowels in /p/ context were not included because the acoustic analysis was done only on the vowels in /b/ context). The relation between these two quantities is indeed monotonic, but due to the small amount

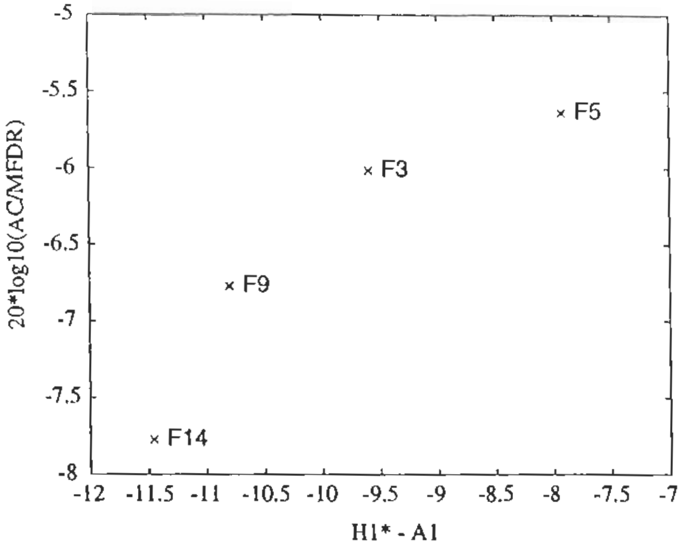


Figure 4.5: *Plot of the aerodynamic measure* $20\log_{10}(AC\ flow/MFDR)$ *vs. the acoustic measure* $H1^* - A1$*. Each point represents the average across the vowels /æ, ʌ, ɛ/ in /b/ context for one of four female speakers.*

of data, we cannot say anything about its linearity or its slope. There are several factors that affect these measures. One such factor is the size of the larynx, which has much individual variation and influences AC flow, and thus H1 (Gauffin and Sundberg, 1989). In addition, Fant et al. (1994) found that the ratio of AC flow to MFDR decreases with increasing F0. As we have seen, F0 can vary quite a bit for certain individuals across recording sessions, adding to the variance of AC flow/MFDR. Thus, although the trend that we found between $H1^* - A1$ and $20\log_{10}(\text{AC flow/MFDR})$ in Fig. 4.5 is encouraging, it must be interpreted with caution.

4.5 Fiberscopy

A schematic of the fiberscope system was shown earlier in Fig. 2.6. For this procedure three of the four subjects were first treated with a topical anaesthetic, lidocaine. The fiberscope was inserted through the nasal cavity and positioned above the vocal folds. Two of the subjects, **F3** and **F14**, repeated the speech material used for the acoustic and aerodynamic data analysis (this material is described in Section 4.2). However, when the other two subjects, **F9** and **F5**, produced the carrier phrases and syllable strings, the view of the glottis was blocked by the epiglottis. Consequently, recordings were only made of **F9** producing the sustained vowel /i/ (two tokens), and of **F5** producing sustained /i, ʌ, ɛ/ (two tokens each). Video recordings were made of these sessions.

An observer experienced in viewing and evaluating fiberscopic images looked at the recordings for each subject. There was some difficulty in evaluating the images because it was necessary to view them in slow motion and that resulted in blurring. The conclusions were drawn from different frames because a clear view of the glottis was not available for each utterance, and are as follows:

- **F14** appeared to have a tiny opening at the arytenoid cartilages, and to have abrupt closure along the membranous part of the vocal folds.
- **F9** seemed to have complete closure at both at the arytenoid cartilages and along the membranous part of the folds in one token. For the other token, there appeared to be a small opening at the arytenoid cartilages and abrupt closure along the membranous part of the vocal folds.
- **F3** showed evidence of an opening at the arytenoid cartilages extending beyond the vocal processes and into the membranous part of the folds. Her opening and closing movements seemed to involve a "rolling" motion, that is, they appeared to be nonsimultaneous.
- **F5** had a large opening at the arytenoid cartilages, possibly extending into the membranous part of the folds. Her glottal opening movements seemed very gradual, but her closing movements appeared to be abrupt.

Figure 4.6 shows schematic drawings of the glottal configurations during the so-called "closed" phase of the glottal cycle that correspond to these descriptions.

These observations were made somewhat informally, and given the small amount of data, they must be considered to be tentative. Nevertheless, there was a difference between the Group 1 speakers and the Group 2 speakers, the latter having more incomplete glottal closures. There were also differences between the two Group 2 speakers, **F3** and **F5**: **F3**'s opening extended nearly to the anterior tip of the glottis, and closure was over a small portion of the vocal folds, while **F5**'s opening appeared to be mostly between the

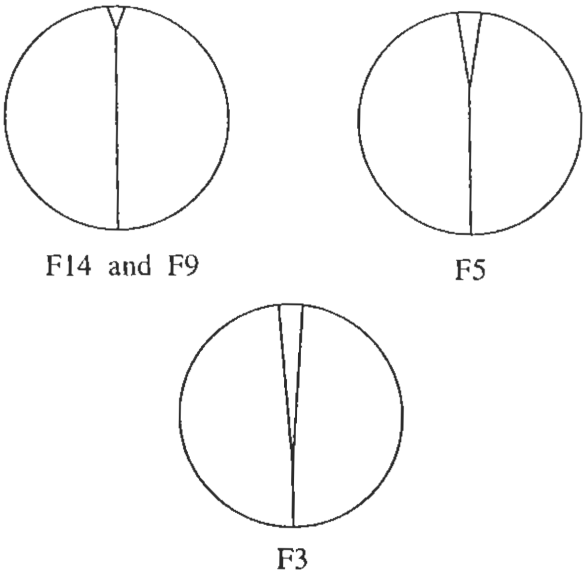


Figure 4.6: *Schematic drawings of the glottal configurations during the "closed" portion of the vibratory cycle for the four subjects, as described by an observer. In these images, the posterior end of the folds is at the top. Subjects* **F14** *and* **F9** *have small openings at the arytenoid cartilages;* **F5** *has a large opening at the arytenoid cartilages, possibly extending into the membranous part of the folds; and* **F3** *has a large opening that extends well into the membranous part of the folds.*

arytenoid cartilages, perhaps extending beyond the vocal processes. In addition, **F5**'s closure seemed abrupt, while **F3**'s appeared to be nonsimultaneous.

The observations made for **F3** and **F5** somewhat contradict the hypotheses that we made about these speakers in Chapter 3. We expected that **F5** would have nonsimultaneous closure, yet she did not appear to in the fiberscopy. However, reference back to the acoustic analysis reported earlier in this chapter shows that this subject had a lower tilt measure H1* – A3* compared to that reported in Chapter 3 (both are given in Table 4.1). The results of the second recording would have placed **F5** in Group 1, according to our criterion given in Chapter 3, so the fiberscopy results are actually in line with the acoustic data.

The results for **F3** are more complicated. Based on her narrow first-formant bandwidth measure, we hypothesized that she would show only a small glottal chink, but due to her high tilt measure, she would have nonsimultaneous closure. However, the fiberscopy showed an opening that extended along most of the length of the folds. The large opening may explain her large tilt measure (cf. Table 4.1) but it is not clear why this large opening does not result in a wide bandwidth. However, recalling the results of the aerodynamic analysis (Tables 4.3–4.4), this speaker had higher subglottal pressure, compared to the other speakers, and also higher MFDR and SPL. The high pressure may serve as a compensation for the losses due to the large opening, with reduced F1 bandwidth and increased SPL as a result, in spite of the losses due to large subglottal coupling.

4.6 Discussion

There were several factors complicating the aerodynamic data collection and analysis. These included mask leak, consonant context, and simultaneous changes in fundamental frequency (F0), maximum flow declination rate (MFDR), AC flow, and sound pressure level (SPL). In addition, there were only four subjects, one of which, **F3**, did not always follow the trends found for the other subjects. Nevertheless, we are able to partially validate our hypotheses that are based on the acoustic data of Chapter 3.

Analysis of the aerodynamic data showed several trends expected from the acoustic analysis. Values of glottal parameters extracted from vowels in syllable strings were somewhat lower than those from vowels embedded in carrier phrases, possibly due to differences in fundamental frequency (Holmberg et al., 1989; Gauffin and Sundberg, 1989; Fant et al., 1994). The speakers from Group 2 had higher minimum flows than speakers from Group 1, supporting our theory that Group 2 speakers have greater losses at the glottis. Group 2 speakers (**F3** and **F5**) also had higher AC flows than the Group 1 speakers (**F14** and **F9**), also a sign of a more open glottal configuration that may result in a breathier voice quality (Gauffin and Sundberg, 1989). In addition, they had higher MFDR values which may indicate that their phonations are also hyperfunctional (Holmberg et al., 1994b).

As predicted from their relatively low values of the acoustic measure H1* – A3*, Group 1 speakers, **F14** and **F9**, have higher speed quotients than Group 2 speaker **F5**. However, their speed quotients are lower than **F3**'s, despite **F3**'s high values of H1* – A3*. This combination of a steep spectral tilt and high speed quotient may be due to a glottal closure that is fast, but nonsimultaneous, supporting our earlier hypothesis about this speaker. Also comparing **F14**, **F9**, and **F5**, there is a tendency for speakers with a larger open quotient to have higher H1* – H2* values. However, **F3** does not follow this trend, either.

Hertegård et al. (1992) found in simultaneous fiberscopy and airflow measures that female speakers of Swedish having a chink extending only to the posterior tips of the vocal processes have lower minimum (DC) flows than speakers with openings that extend into the membranous part of the folds. Therefore, based on our measures of minimum flow, Group 2 speakers are more likely than Group 1 speakers to have openings at the glottis that extend to the membranous part of the folds. However, minimum flow may not be a reliable measure. First-formant bandwidths estimated from minimum flow and subglottal pressure did not line up well with the first-formant bandwidths measured from the speech waveforms. Given the potential for mask leak and the difficulty in detecting such a leak during the recording, this result is not so surprising. Mask leak was detected for at least two of the subjects (due to negative flow values), and may have been present, but undetected, for the other two subjects. In addition, Hertegård et al. (1992) found that male speakers with complete glottal closure had minimum flow measures ranging from 0 to 90 cm^3/sec. They theorized that this flow was due to vertical movements of the vocal folds during closure, and that for normal ranges of F0, the result could be a DC flow offset of up to 20–30 cm^3/sec for both male and female speakers. Thus, there are two independent sources of error for DC flow. Holmberg et al. (1988, 1994a, 1994b, in press) have repeatedly found that minimum flow measures do not show expected correlations with other measures, and that minimum flow values vary greatly across repeated recordings on the same subject, further suggesting that this measure is problematic and somewhat unreliable.

The relation between the bandwidth measure H1* – A1 and the ratio of the AC flow to the MFDR also supports our contention that glottal characteristics can be estimated from measurements on the speech spectrum. Speakers with larger H1* – A1 values have greater AC flow to MFDR ratios, as predicted. Another interesting result is that consonant context can affect glottal characteristics, in particular MFDR, AC flow, and SPL, during vowel production. This finding is in agreement with earlier studies by Gobl and Ní Chasaide (1988), and Ní Chasaide and Gobl (1993).

The fiberscopy for two subjects, **F9** and **F5**, was performed using sustained vowels as opposed to the vowels in carrier phrases and syllable strings that the other two subjects produced, so care must be taken when comparing their results to those of the other speakers. There are differences between the two groups, however: the Group 1 speakers, **F9** and **F14** seemed to have either complete closure along the entire length of the vocal folds, or small openings at the arytenoid cartilages, and they have opening and closing movements that seem to be simultaneous along the membranous part of the folds. However, Group 2 speakers seemed to have large openings at the posterior end of the glottis extending into the membranous part of the folds. For one Group 2 speaker, the closing movement seemed to be nonsimultaneous along the length of the folds. These observations generally agree with the acoustic analysis presented in Section 4.3: **F9** and **F14** have low tilt measures and narrow first-formant bandwidths; **F5** has a large first-formant bandwidth and relatively low tilt; and **F3** has a large tilt. However, **F3**'s larger opening at the posterior end of the glottis was not predicted by the acoustic analysis.

Although we can only make tentative conclusions based on the physiological data, the trends that we found for the four subjects examined support our hypotheses based on the acoustic data collected and analyzed in Chapter 3. Therefore, it seems possible that the acoustic measures, on which these hypotheses were based, capture information about glottal vibration and configuration.

Chapter 5

Voice quality perception tests

5.1 Introduction

The results of the acoustic analysis of Chapter 3 have shown some differences between speakers, and we have used those results to categorize our subjects according to glottal configurations that we believe are the bases of these differences. Our results from Chapter 4 have suggested that there is some basis to our hypothesis that the acoustic measures can be used to classify speakers according to their glottal configurations. Specifically, a group of speakers that we refer to as Group 2 have steep spectral tilts, significant noise excitation at high frequencies, and strongly damped first formants, compared to the speakers in Group 1. We have suggested that these measures imply glottal configurations that have large openings during the closed part of the glottal cycle, probably extending to the vocal processes. In Chapter 4 we found physiological evidence to support this hypothesis.

The question arises as to whether or not the acoustic differences that we found are perceptually meaningful. The glottal configuration that we have hypothesized for Group 2 speakers has been associated with breathy voice (Klatt and Klatt, 1990; Stevens and Hanson, 1995). If this association is correct, and the acoustic measures accurately reflect glottal configuration, then listeners should perceive Group 2 speakers as being breathier than the speakers in Group 1. Another question is whether these acoustic differences contribute to what we call speaker characteristics.

full voice —————————————————— breathy

Figure 5.1: *Continuum used by listeners to make breathiness judgments on vowels for the test described in Section 5.2. A mark on the left end of the line indicates that the vowel was perceived to be fully voiced, while a mark on the right end of the line means that the vowel was perceived to be breathy.*

We will explore these two questions through two listening tests. In the first test, we will examine the correlation of the acoustic measures to perceived breathiness. In the second test, we will attempt to determine if the glottal characteristics contribute to the successful synthesis of a given person's voice, and if so, which parameters are of particular importance.

5.2 Breathiness perception test

5.2.1 Method

For our first listening test, the hypothesis was that Group 2 speakers would be perceived to be breathier than Group 1 speakers. The stimuli for the test were the vowels /æ, ʌ, ɛ/ excised from the carrier phrases recorded for Chapter 3 (see Section 3.3). These vowel segments ranged in duration from about 100 ms to about 275 ms. One token of each vowel for each speaker was used. These tokens were chosen arbitrarily to be the third token recorded. For each test item, an excised vowel was repeated three times in succession. The test items were randomized and each was presented three times during the test.

There were four listeners, three female and one male, all speech researchers with experience doing listening tests. Listeners made ratings on breathiness for all 22 subjects. They were asked to make ratings along a continuum from "breathy" to "full voice", as shown in Fig. 5.1. This method of rating voice quality is due to Gelfer (1993). All listeners found the task to be very difficult, saying that it was not easy to ignore strong perceptual factors such as vowel quality and pitch.

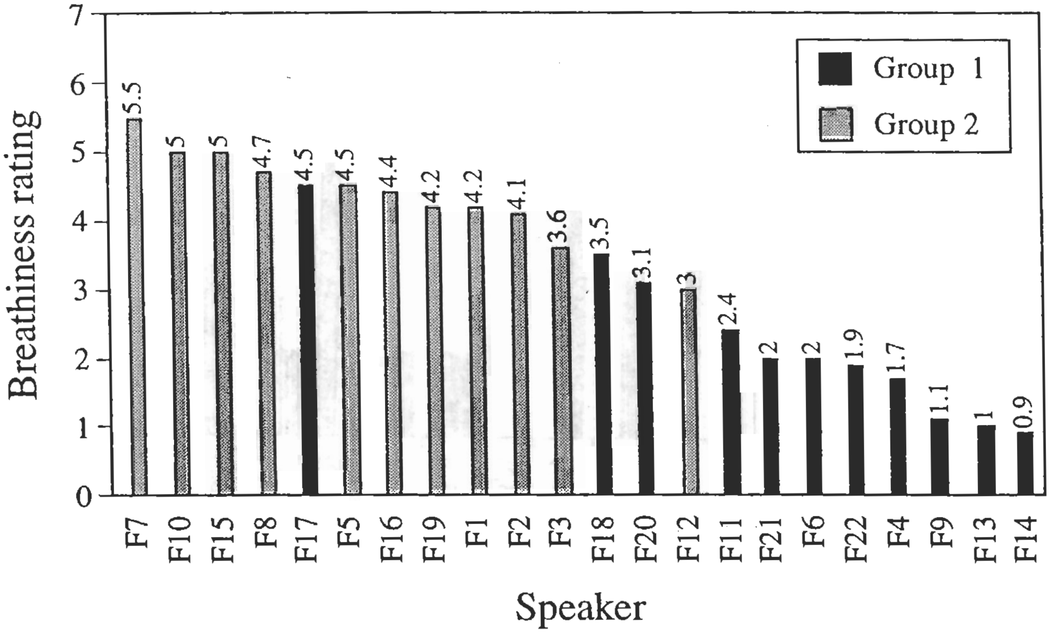


Figure 5.2: *Breathiness ratings for 22 female subjects. Each rating represents an average across four listeners. Group 2 speakers tend to be perceived as breathier than Group 1 speakers.*

5.2.2 Results

Following the tests, the ratings were converted to scores on a scale from 0.0 to 7.0, where 0.0 corresponds to "full voice" and 7.0 corresponds to "breathy". They were then averaged across vowels and listeners, resulting in a mean breathiness rating for each speaker. These ratings are given in Figure 5.2, where the speakers are ranked according to breathiness. For the most part, Group 2 speakers were judged to be breathier than Group 1 members. A histogram of the results is shown in Fig. 5.3, illustrating that, despite some overlap, the two groups are fairly well separated. The average rating for Group 1 was 2.2, while for Group 2 it was 4.4.

Pearson product moment ~~d~~correlation coefficients were computed for pairwise comparisons of the ratings given by the different listeners. These were found to be strong ($0.71 \leq r \leq 0.87$ for five out of six comparisons), so there seems to be a general agreement among listeners as to which voices are perceived to be breathy. Correlation coefficients for comparisons between the mean breathiness ratings for each vowel for each speaker

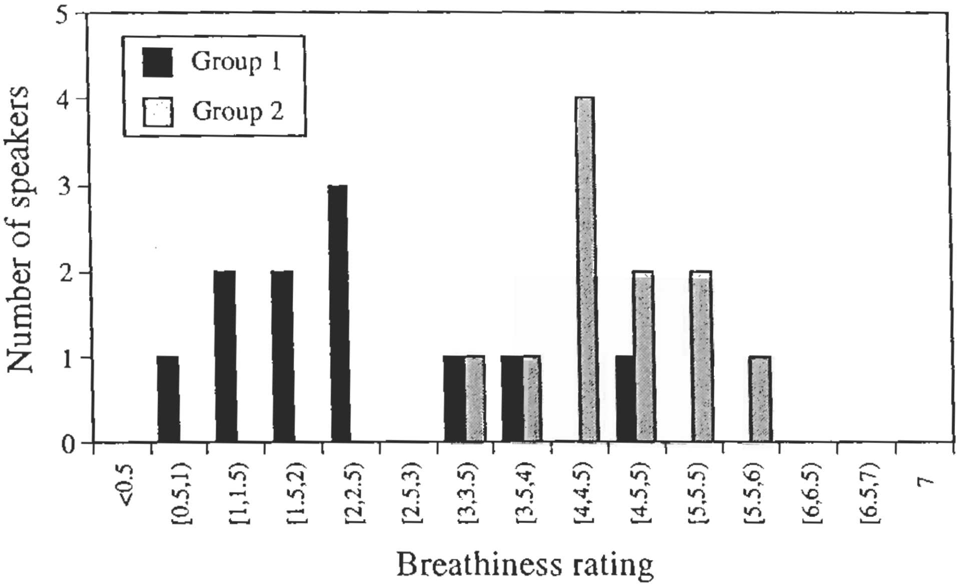


Figure 5.3: *Histogram of breathiness ratings given in Fig. 5.2. The two groups of speakers can be seen to be separated, with Group 2 speakers perceived to be breathier than Group 1 speakers. The average breathiness rating for Group 1 speakers was 2.2, while for Group 2 speakers it was 4.4.*

and the corresponding acoustic measures described in Section 3.3.3 were computed for each listener, and also for the mean breathiness ratings across listeners. These correlations are summarized in Table 5.1. Many researchers have suggested H1 – H2 as a measure of breathiness (see Klatt and Klatt, 1990, for a review), but H1* – H2* was not a good correlate of our mean perceived breathiness ratings ($r = 0.25$). However, H1* – A1 and H1* – A3* were strongly correlated with mean perceived breathiness ($r = 0.74$ and $r = 0.69$, respectively). This result is not surprising given that the division of tbe subjects into two groups was largely based on these two measures (cf. Section 3.3.3), and Group 2 is perceived to be breathier than Group 1. The spectra-based noise ratings (N_S) are also well correlated to the mean breathiness ratings ($r = 0.75$), as might be expected given the strong correlation that it has to the measures H1* – A1 and H1* – A3* (cf. Tables 3.6–3.7 in Section 3.3.3). The first-formant bandwidth (B1) and the waveform-based noise ratings (N_W) are only moderately correlated to mean perceived breathiness ($r = 0.50$ and $r = 0.64$, respectively). These results are in agreement with results reported in Klatt and Klatt (1990), who used synthesized speech to examine how glottal parameters affect perceived breathiness. They found that an increased H1 – H2 did not by itself introduce

Table 5.1: *Pearson product moment correlation coefficients* r *between breathiness ratings and the acoustic measurements made on these vowels in Chapter 3. The first four lines give* r *individually for the four listeners. The last line gives* r *for the comparison between the mean breathiness ratings across listeners and the acoustic measures. An insignificant coefficient is indicated by the notation 'n.s.'*

Listener	H1* − H2*	H1* − A1	H1* − A3*	B1	N_W	Ns
L1	0.26	0.66	0.62	0.41	0.61	0.67
L2	0.22	0.70	0.54	0.63	0.58	0.58
L3	0.26	0.75	0.71	0.40	0.63	0.74
L4	*n.s.*	0.57	0.63	0.50	0.50	0.71
mean	0.25	0.74	0.69	0.50	0.64	0.75

a breathy quality to synthesized vowels, but that increasing both spectral tilt and the aspiration noise at high frequencies did increase perceived breathiness. They also found that increasing formant bandwidths alone did not increase perceived breathiness.

Comparing the correlations for different listeners, we see that there is some variation as to which acoustic measures seem to provide strong perceptual cues to breathiness. This variation suggests that the four listeners had somewhat different criteria for judging breathiness. Klatt and Klatt (1990) also found intersubject differences in the cues used to perceive breathiness.

5.2.3 Discussion

Due to the small number of listeners and the difficulty that they had in making subjective ratings on stimuli of such short duration, one should not read too much into our results. However, the results do seem to support our method of classifying glottal configurations based on acoustic measurements. The Group 2 speakers, hypothesized to have a phonation with a greater average glottal opening than Group 1 speakers, were perceived to be breathier by our listeners. In addition, three of the spectral measures made on these vowels and used to categorize the speakers in Chapter 3 were found to be strongly correlated with the breathiness ratings. These measures are the bandwidth measure H1* − A1, the tilt $H1^* - A3^*$, and the high-frequency noise rating N_S.

Because there were some individual differences between listeners regarding which acoustic measures were correlates of a breathy voice quality, more testing should be done with more listeners to determine if these differences are merely statistical artifacts or if they have some significance for the perception of breathiness. Cleaner results might be obtained using sustained vowels instead of vowels excised from carrier phrases. Sustained vowels have a longer duration, and thus it should be easier for listeners to make a judgment about them. Another option is to use synthesized speech, which would allow us to carefully control glottal parameters, while eliminating the effects of vowel quality and pitch that contributed to the difficulty reported by listeners.

5.3 Voice quality synthesis test

In the next listening test to be described, we wanted to determine if glottal parameters are important for describing a given person's voice quality, or if other factors such as formant frequencies and fundamental frequency are overwhelming factors. If glottal parameters did turn out to be important, we wanted to get some idea as to *which* parameters are most influential. Another goal was to determine if the acoustic measurements made in Chapter 3 could be applied successfully to the synthesis of a given person's voice. We will begin this section by giving an overview of the test, and then turn to a description of the test stimuli. This description is followed by a discussion of the method used to construct the stimuli. Next, we describe the test administration and discuss the results.

5.3.1 Overview

To study the influence that glottal parameters have on a person's perceived voice quality, we asked listeners to compare natural vowels to synthesized vowels for a subset of six of our group of 22 female speakers. The voice quality of the synthesized vowels were intended to vary along a continuum from *pressed* to *breathy*.[1] Different voice qualities were obtained

[1] This is actually very simplistic because there are several types of breathy voice, including breathy/hyperfunctional phonation which is both breathy and pressed.

by varying synthesizer parameters in a way that we believed would result in the desired continuum. This process will be described in more detail below.

One option for the test was to have stimuli that were pairs of natural and synthesized speech tokens, and to ask listeners to make ratings of the synthesis. However, the listeners for the test described in Section 5.2 found it difficult to make these subjective ratings, so we chose to make the test of the form AXB. In this type of listening test, X is a reference item (in our case, the natural speech) and A and B are test items (in our case, synthesized speech). The listeners are then asked to choose which test item, A or B, is closest to the reference item, X. This type of test also requires a subjective rating of the test items, but is much easier for the listeners to do. A diagram of the AXB test is shown in the upper part of Fig. 5.4. The details of this diagram will be explained in the following text.

Another concern was how best to ensure that the listeners' judgments would be based on voice quality, rather than on vowel quality, consonant quality, or prosodic effects. In an effort to reduce these effects, the pitch and formant contours used for the vowel synthesis were obtained by copying measured data from the vowel segments of the reference tokens. We also avoided the influences of the neighboring stop consonant occlusions and bursts by not synthesizing the occlusions and releases of the /b/ and /d/ segments of the test items to match those of the reference items. Rather, two /b/-/d/ occlusion/release pairs were obtained by excising them from one token each of /bʌd/ and /bɛd/. The criteria in choosing these /b/-/d/ tokens was that there be little pre-voicing during the occlusion, and that the release not be too strong (and thus distracting). These were concatenated with both the synthesized test vowels and the vowels excised from the natural speech. The same two pairs were used for all six speakers.

5.3.2 Stimuli

5.3.2.1 Description

As reference items, we chose one token each of the words 'bud' and 'bed' for each of the six speakers. Single words were chosen over a longer phrase to avoid the influence of prosody and consonant quality in determining the goodness of synthesis.

The test items were also the words 'bud' and 'bed'. The vowels were synthesized using

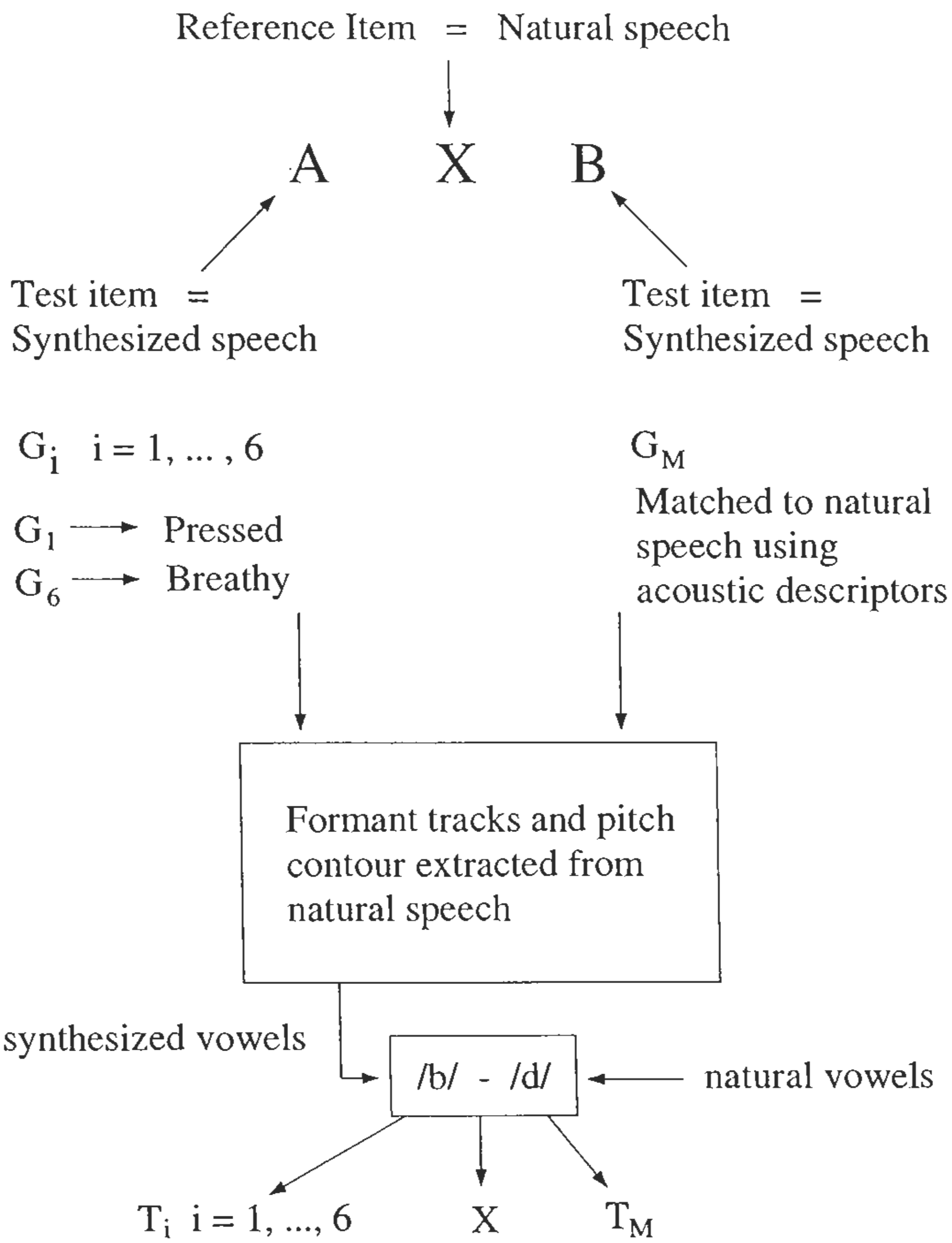


Figure 5.4: *Diagram of the voice quality synthesis test and the construction of its stimuli. For details, see the accompanying text.*

Table 5.2: *The glottal parameter sets $G_i, i = 1 \ldots 6$, used to synthesize the test items referred to as T_i. The parameters vary such that set G_1 contains parameters appropriate for a pressed voice quality, while set G_6 is appropriate for a breathy voice quality.*

	Glottal parameter set					
Synthesizer parameter	G_1	G_2	G_3	G_4	G_5	G_6
Open quotient (OQ) (%)	57	60	63	65	68	70
Spectral tilt (TL) (dB)	0	5	10	15	20	25
First-formant bandwidth (B1) (Hz)	60	90	120	150	180	200
Aspiration noise (AH) (dB)	48	45	43	40	37	35

the KLSYN88 formant synthesizer (Klatt and Klatt, 1990). This synthesizer has three voicing sources that allow a great deal of control over voice quality. We chose to use the KLGLOTT88 voicing source. The synthesizer parameters that control this source are OQ (open quotient), TL (additional spectral tilt), and AH (amplitude of aspiration noise). Strictly speaking, the first-formant bandwidth (B1) contributes to the vocal-tract transfer function, but because the cross-sectional area of the glottis contributes to B1, we will treat it loosely as a glottal parameter. In general, as these parameters increase, the voice quality should become breathy, as the experiments described above in Section 5.2 have shown. The F0 and formant contours of the test items were based on those of the natural speech, as described above, and all other parameters were held constant.

There were two types of test items. One type consisted of words synthesized using the sets of glottal parameters presented in Table 5.2. These sets, which we shall refer to as $G_i, i = 1, \ldots, 6$, were chosen to vary so that the first set, G_1, contained parameters thought to be appropriate for a pressed voice quality, while the last set, G_6, contained parameters thought to result in a very breathy voice quality. Note that the parameters OQ, TL, and B1 increase with breathiness, but parameter AH decreases, because as tilt increases less noise is necessary to get the appropriate signal-to-noise ratio. By varying the glottal parameters across test items, we can determine if listeners prefer only a small range of glottal parameters for a given speaker, or if an arbitrary set of glottal parameters will do. In the latter case, voice quality would seem to be unimportant compared to vowel quality and pitch for distinguishing a speaker.

The second type of test item was one synthesized to obtain a good match to the natural speech; the results reported in Section 3.3.3 were used to guide this synthesis. The set of glottal parameters used to synthesize this type of test item will be referred to as G_M, where the subscript refers to the fact that this set of parameters yields synthesized speech that matches the natural speech. The inclusion of this type of test item allowed us to determine the usefulness of the acoustic measurements in guiding the synthesis of female voice, and also provided a means to determine which parameters are most influential for voice quality.

Test items synthesized using the glottal parameter sets G_i will be referred to as items T_i, while a test item synthesized using the glottal parameter set G_M to closely match the natural speech will be referred to as item T_M. Each AXB sequence had one test item, A or B, that was an item T_i, while the other test item was always the item T_M. As noted above, item X in the sequence was the naturally spoken word. It is important to note that the only difference between the items T_i and T_M is the set of glottal parameters, because the pitch and formant contours, and the neighboring consonants /b/ and /d/ are always the same between the two test items A and B in a stimulus. Thus, listeners' judgments should only be based on voice quality.

5.3.2.2 Construction

Six of our group of 22 subjects were chosen for the synthesis tests. Three were from Group 1 (**F9**, **F13**, and **F18**) and the other three were from Group 2 (**F2**, **F10**, and **F15**). Subjects **F9** and **F13** had breathiness ratings that were among the lowest for the entire group of speakers; **F10** and **F15** had two of the highest breathiness ratings; and **F18** and **F2** had ratings that were at about the middle of the range.

We constructed two tests, one each for the words 'bud' and 'bed'. Reference items were obtained by excising the vowels /æ, ʌ, ɛ/ from the carrier phrases recorded for the acoustic analysis described in Section 3.3. One token was obtained for each speaker. These vowel tokens were then concatenated with the appropriate /b/-/d/ pair to form the words 'bud' and 'bed'.

The vowel portions of the test items were synthesized as follows. First, each reference item was analyzed to obtain fundamental frequency (F0) and formant tracks for the vowel segments.[2] The formant and F0 tracks so obtained were used as the basis for synthesizing the test items. The test items T_i were synthesized using the six sets of glottal parameters G_i defined in Table 5.2. Although there is evidence that glottal parameters vary throughout the course of vowel production, due to the effects of adjacent speech segments (Gobl and Ní Chasaide, 1988; Ní Chasaide and Gobl, 1993), these parameters were held constant throughout the duration of the vowel to maintain simplicity. It may seem like this would contribute some unnaturalness to the quality of the synthesized vowels, but informal observations implied that this was not the case. This procedure resulted in vowels that ranged from having a pressed quality to a breathy quality. For speakers from Group 1, sets G_1–G_5 were used, while sets G_2–G_6 were used for Group 2 speakers. Thus, five default test vowels were obtained for each vowel for each speaker. The synthesized vowels were then concatenated with the natural consonants /b/ and /d/ to form the test items T_i.

The next step was to synthesize a test item T_M for each of the reference items. The first-formant bandwidth parameter of the synthesizer (B1) was set to the value reported for the speaker in Table 3.2. The parameters that control spectral tilt (TL) and open quotient (OQ) were adjusted until the values of $H1^* - H2^*$ and $H1^* - A3^*$ at mid-vowel were close (within 0.5 dB) to the average values reported for the appropriate vowel (/ʌ/ or /ɛ/) in Tables 3.3–3.4. The parameter that controls the amplitude of the aspiration noise was adjusted until the high-frequency spectrum appeared to have a harmonic/noise content that would be given a noise rating close to the average obtained in Chapter 3. Because we set the B1 parameter to be the average value obtained for the first-formant bandwidth measured in the acoustic analysis reported in Chapter 3, we could not use B1 to adjust the amplitude of the first-formant (A1), and therefore, we did not make adjustments to match $H1^* - A1$ to the average values reported in Tables 3.3–3.4. This sometimes resulted in a difference of as much as 4 dB between $H1^* - A1$ of the natural speech and the synthesized test item T_M. For reference, the parameters used for each speaker are

[2] The F0 and formant tracking were done using software written by Dennis Klatt.

Table 5.3: *The glottal parameter sets G_M used to synthesize the test items referred to as T_M. There is one set for each speaker for each word. The parameter B1 was set to the average value measured for the vowel /æ/ in Chapter 3 (except for* **F18***'s /bʌd/ and* **F10***'s /bɛd/, which were set incorrectly). The remaining parameters were adjusted according to the average values obtained for the acoustic measures $H1^* - H2^*$, $H1^* - A3^*$, and N_S for the vowels /ʌ/ and /ɛ/ in Chapter 3.*

/bʌd/	F9	F13	F18	F2	F10	F15
Open quotient (OQ) (%)	60	63	63	57	63	64
Spectral tilt (TL) (dB)	8	3	8	19	12	17
First-formant bandwidth (B1) (Hz)	104	53	173	244	184	256
Aspiration noise (AH) (dB)	43	48	48	45	50	38

/bɛd/	F9	F13	F18	F2	F10	F15
Open quotient (OQ) (%)	64	63	65	54	65	61
Spectral tilt (TL) (dB)	10	7	9	22	22	22
First-formant bandwidth (B1) (Hz)	104	53	163	244	280	256
Aspiration noise (AH) (dB)	40	45	45	42	41	32

summarized in Table 5.3. These parameters were also held constant throughout the vowel duration, as described above. The synthesized vowels were then concatenated with the natural consonants /b/ and /d/ to form the test items T_M.

The test stimuli (the AXB sequences) were constructed by concatenating a reference word between a T_i test item and the T_M test item corresponding to the reference word. There was 600 ms of silence between the reference word and the test items. Each stimulus had a 'mirror' stimulus in which the order of the T_i and T_M test items were reversed. As a result, we had 10 stimuli for each of six speakers for each word. Two test tapes were created, one for the 'bud' test and the other for the 'bed' test. For both tests, each stimulus (AXB sequence) was repeated three times. Twenty-four (24) stimuli were repeated at the beginning of the test to allow the listeners to adjust to the test procedure—these were not included in the results. Altogether, the number of stimuli for each test was 204. The stimuli were presented in blocks of six, with each block containing one stimulus for each speaker. There were 3.5 seconds of silence between stimuli during which the listeners wrote down their answer (A or B), and 5 seconds of silence between each block of six stimuli.

5.3.3 Test administration

There were eight listeners, three male and five female, all speech researchers with experience doing listening tests. Except for one listener, the tests were completed in two sessions held on separate days. Five of the listeners did the 'bud' test first, and the other three did the 'bed' test first. The tests were administered separately for each listener, in a sound-treated room. A sample of the written instructions that they received is given in Appendix B. These written instructions were supplemented with an oral explanation.

5.3.4 Results

For each test, the number of times that a test item T_i was preferred over the test item T_M in an AXB stimulus was recorded. These counts were averaged across listeners and then converted to percentages of times that T_i was preferred over T_M for a given speaker. These percentages are presented as bar graphs in Figs. 5.5–5.6. Each graph represents one speaker. Graphs corresponding to Group 1 speakers are presented in the lefthand column, while those corresponding to Group 2 speakers are presented in the righthand column. In these graphs, a score of zero percent means that the test item T_M was always preferred over test item T_i; 50 percent means that listeners had no preference for T_M over T_i; and greater than 50 percent means that T_i was preferred over the stimulus T_M that was based on the average data for that speaker.

If the use of the acoustic measures of Chapter 3 improves the synthesis of voice quality, then, ideally, the test item T_M would always be preferred over the items T_i. However, more often than not, at least one T_i per speaker was preferred over the T_M by more than 50 percent. An explanation might be that we used the average waveform-based bandwidth measure to set the B1 parameter of the synthesizer, rather than adjusting B1 until the average H1* – A1 measure was matched. As we have seen in Section 5.2, H1* – A1 is a better correlate of perceived breathiness than is the first-formant bandwidth. An informal

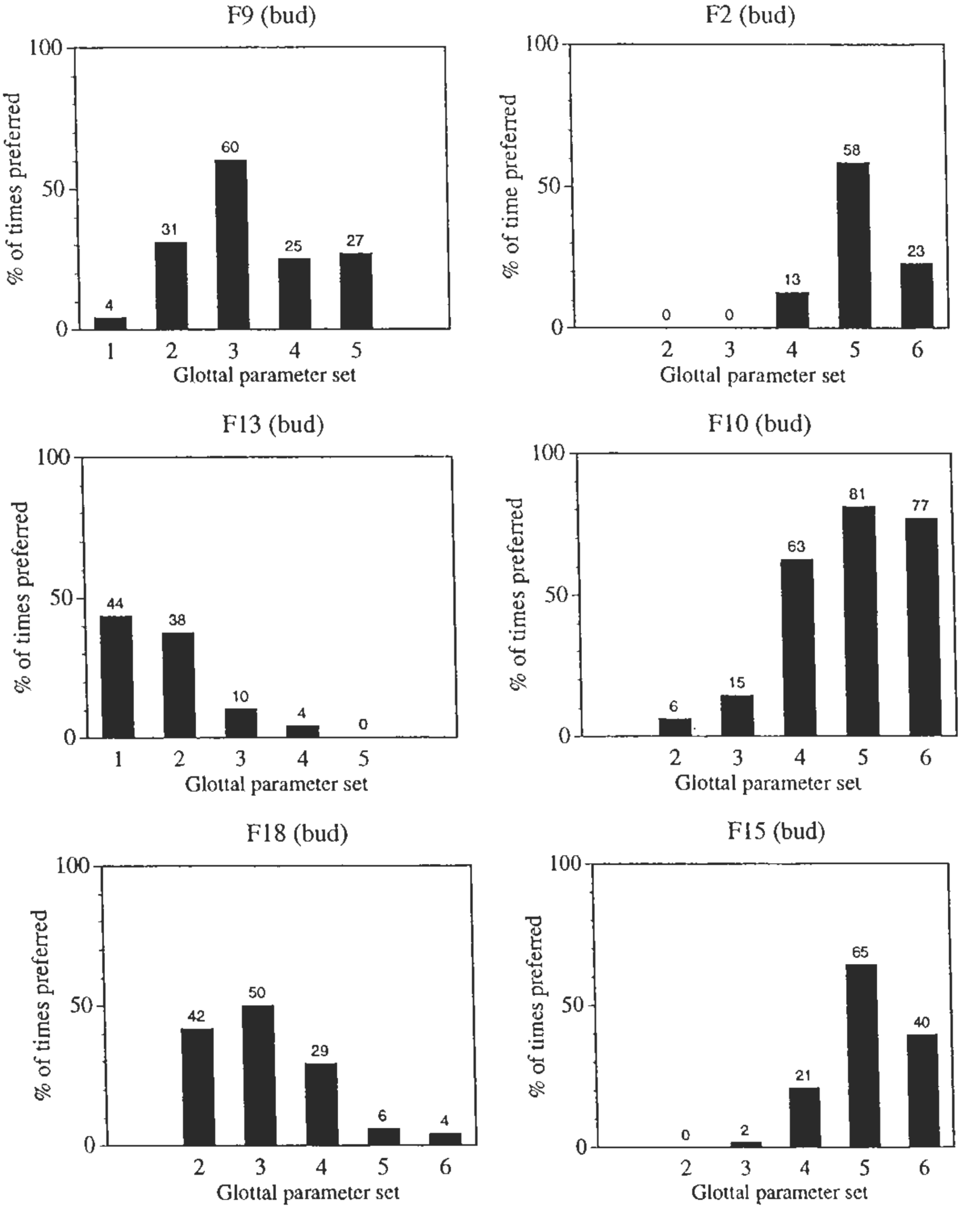
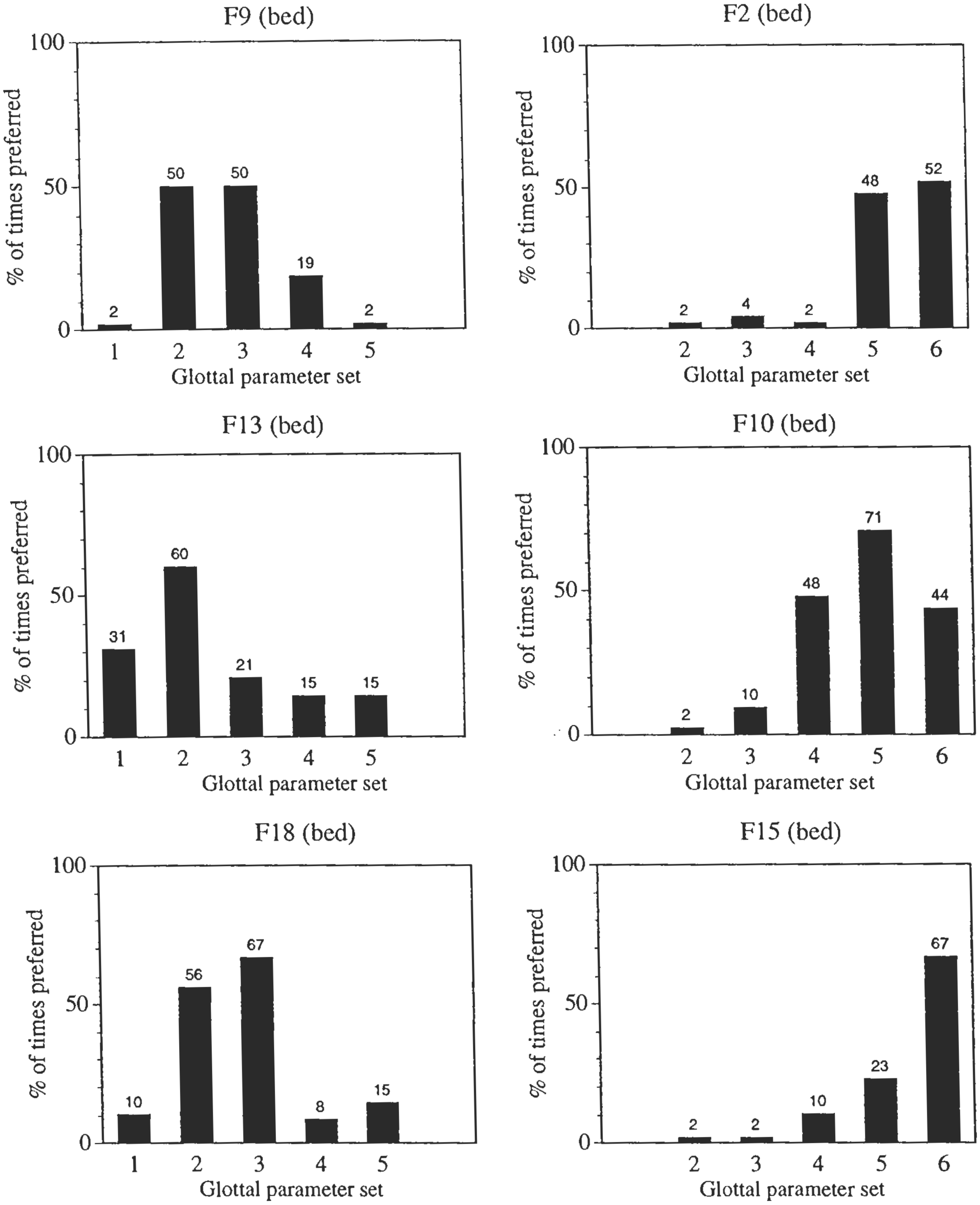


Figure 5.5: *Results of the synthesis test for the word 'bud'. Each graph represents one of the six speakers. The abscissa is the number of the glottal parameter sets* G_i*, and the ordinate is the percent of time that a test item* T_i *was chosen to be closer to the natural speech than test item* T_M *for that speaker.*


F9 (bed)

F2 (bed)

F13 (bed)

F10 (bed)

F18 (bed)

F15 (bed)

Figure 5.6: *Results of the synthesis test for the word 'bed'. Each graph represents one of the six speakers. The abscissa is the number of the glottal parameter sets* G_i, *and the ordinate is the percent of time that a test item* T_i *was chosen to be closer to the natural speech than test item* T_M *for that speaker.*

post-hoc analysis showed that when T_i is preferred by more than 50 percent, it often had a value of H1* – A1 that was closer to the value for the natural speech than did the test item T_M.

It seems clear from these graphs that a single set of glottal parameter values cannot be used to obtain good synthesis of all six speakers. The voice qualities of the speakers from Group 1 are best synthesized using the range of glottal parameters covered by the parameter sets G_1–G_3, while those of Group 2 speakers are best synthesized by the sets G_5–G_6. It is equally obvious that the glottal parameter sets appropriate for Group 1 speakers are never appropriate for Group 2 speakers, and vice versa. Even within a group, one set of parameters is not enough to describe all speakers. As an example, for the word 'bud', G_1 is best for speaker **F13**, but G_3 is best for speaker **F9**.

Some insight into which glottal parameters contribute the most to voice quality can be gained by comparing the parameter values of the most preferred set G_i to those of the set G_M that was synthesized to provide a good match to the natural speech. The third column of Table 5.4 gives the number of the glottal parameter set G_i that was most preferred for each speaker for each word. In cases where the second most preferred set G_i was within 10 percent of the most preferred set, both sets are listed. The next four columns each represent one synthesizer parameter. Listed for each parameter is the number of the set G_i that most closely matches set G_M in terms of that parameter. For example, for the word 'bud', set G_5 was judged to give the best voice quality for speaker **F2**, as shown in the third column. The value of the synthesizer parameter OQ (open quotient) for her set G_M is 57 (cf. Table 5.3), and of all the sets G_i, set G_1 (OQ= 57) has the OQ value that is closest to that of G_M, so the number 1 is recorded in the column marked OQ. As can be seen, in nine out of 12 cases the tilt parameter TL of the set G_M would predict the preferred default set G_i. The aspiration noise parameter AH predicts correctly in half of the cases, while the first-formant bandwidth B1 predicts correctly in five out of 12 cases. The open quotient parameter OQ predicts correctly in only one case. As has been suggested earlier (Klatt and Klatt, 1990), then, spectral tilt combined with aspiration noise seems to be the most important factor in determining breathiness. High-frequency noise and first-formant bandwidth also make some contribution, but open quotient does

Table 5.4: *Preferred glottal parameter sets G_i for each word and speaker (column 3), and the sets G_i that would be predicted to be preferred by the individual parameter values of set G_M (columns 4–7). Each of columns 4–7 represents one of the four glottal parameters. The numbers in these columns are the numbers of the set G_i whose value of that parameter is closest to the value of the corresponding set G_M. Numbers given in boldface are for parameters of G_M that correctly predict the preferred T_i. Note that the parameter TL (spectral tilt) correctly predicts the preferred test item T_i in almost every case. Parameters AH (high-frequency aspiration noise) and B1 (first-formant bandwidth) are moderate predictors. OQ (open quotient) is not a good predictor.*

Subject	Word	Preferred T_i	OQ	TL	B1	AH
F9	bud	3	2	**3**	2	**3**
	bed	2,3	4	**3**	**2**	4
F13	bud	1,2	3	**2**	**1**	**1**
	bed	2	3	**2**	1	**2**
F18	bud	2,3	**3**	**3**	5	1
	bed	2,3	4	**3**	4	**2**
F2	bud	5	1	**5**	6	2
	bed	5,6	1	**5**	**6**	4
F10	bud	5,6	3	3	**5**	1
	bed	5	4	**5**	6	4
F15	bud	5	4	4	6	**5**
	bed	6	2	5	**6**	**6**

not seem to be important.

5.3.5 Discussion

The results of our speech synthesis test show that when glottal parameters are varied across tokens, listeners are able to perceive the differences. Glottal parameters are not overpowered by other factors such as formant frequencies and pitch contours, but rather they contribute to the overall quality of speech. What is more, if glottal parameters are set correctly, the synthesis of voice quality for a given speaker is improved.

The values that make up the glottal parameter set G_1, given in Table 5.2, are actually the default values of the Klatt synthesizer. Yet from Figs. 5.5-5.6 we see that in only one case out of 12 (**F13**'s 'bud' token) were these parameters judged to result in a good synthesis for a speaker. Glottal parameter sets that should result in a synthesized vowel with a breathy voice quality were preferred for Group 2 speakers, while glottal parameter sets that should result in a more modal to pressed quality are preferred for Group 1 speakers. This result is in agreement with the test described in Section 5.2 showing that listeners found Group 2 speakers to be breathier than Group 1 speakers. From both of these results, we can conclude that the range of glottal parameters given in Table 5.2 is necessary to synthesize a variety of voice qualities. More specifically, this range is needed in order to synthesize speech that sounds like a particular female speaker. In fact, given that the glottal set G_6 was often preferred for three of the six speakers, this range should possibly be expanded to produce an even breathier voice quality.

We have also found evidence that some glottal characteristics have a greater influence on voice quality than others. Specifically, spectral tilt appears to strongly influence listeners' perceptions of breathy voice quality, while the difference between the amplitudes of the first two harmonics (H1* – H2*) has little effect. Aspiration noise and first-formant bandwidth have a moderate effect. This evidence is in agreement with results reported in Klatt and Klatt (1990), who found that not all glottal parameters had an equal effect on perceived breathiness.

Finally, the use of the acoustic data gathered in Chapter 3 to guide the synthesis had slightly mixed results. While test items T_M were generally considered to be good, there was usually a test item T_i that was considered to be a better match to the natural speech than T_M. We have suggested that the use of the H1* – A1 data instead of the waveform-based first-formant bandwidth measure to set the synthesizer parameter B1 might have improved the results. This hypothesis should be investigated through additional listening tests.

5.4 Summary

In this chapter we have described two listening tests through which we hoped to answer some questions about the relation between acoustic measures of vowels and perceived voice quality. Earlier, in Chapter 3, we used measures on the speech spectrum to classify speakers according to glottal configuration. One group of speakers was hypothesized to have insufficient glottal closure compared to the other group. This more open glottal configuration suggests a breathier voice quality. In Section 5.2 we found that the speakers hypothesized to have the more open glottal configuration were indeed perceived by listeners to have breathier voice quality. Three of our spectral measures of glottal parameters, spectral tilt (H1* – A3*), first-formant bandwidth (H1* – A1), and high-frequency noise (N_S), were found to be well correlated with the breathiness ratings. These results suggest that our acoustic measures may be valid for classifying speakers according to glottal configuration.

For the other test we used synthesized speech to answer several questions. We wanted to know (1) if listeners can perceive changes in glottal parameters, and if so, did these changes lead to changes in perceived voice quality; (2) if the acoustic measures developed in Chapter 3 could be used to guide the synthesis of a particular person's voice; (3) which glottal parameters have the greatest influence on perceived voice quality. The results given in Figs. 5.5–5.6 show that changes in glottal parameters resulted in differences that could be perceived, and that all glottal parameter sets G_i were not equally satisfactory for synthesizing a speaker's voice. The preferred set G_i varied from speaker to speaker, indicating that selection of synthesizer parameters that control the voice source is important for successful synthesis of a person's voice. These figures also indicate that the acoustic measures developed in Chapter 3 have the potential to guide the synthesis of a given speaker's voice, although more work is needed to improve this process. We also found that certain spectral parameters, in particular spectral tilt, are more influential in affecting voice quality than are other acoustic parameters.

Chapter 6

Summary and conclusions

6.1 Summary of findings

The aim of this thesis has been to develop a set of descriptors of the voicing source that reflect individual differences in the voice qualities of female speakers. While other researchers have worked in this area, they have primarily relied on the method of inverse filtering the speech or oral airflow signal to obtain the glottal waveform (see, for example, works by Holmberg et al. and Karlsson, listed in the bibliography). This method requires special equipment, and if the inverse filter is not chosen with care, incorrect results will occur. Also, the method is somewhat labor-intensive. A few researchers have used measurements made only on the speech spectrum to obtain descriptors of voice quality (see especially Klatt and Klatt, 1990, but also Kasuya and Ando, 1991). Such measurements have the advantage that they can be made from simple microphone recordings and have the potential to be easily automated.

We have sought to further develop this latter method of studying voicing source characteristics. We began by developing theoretical background describing how glottal characteristics may be manifested in the speech spectrum or waveform. We showed that two factors in particular have an effect on the spectrum: the presence and size of a posterior opening at the glottis, and the abruptness with which the membranous part of the vocal folds closes. An outcome of this theoretical development was the formulation of several measures to be made on the speech spectrum or waveform. These measures indicate the open quotient of the glottal waveform, the bandwidth of the first formant, the tilt of the source spectrum, and the amount of noise at mid- to high frequencies. All of these measures might be expected to increase as the glottal configuration becomes more open.

These measures were applied to the steady-state portions of vowels excised from the speech of 22 female speakers. We found substantial individual differences in several parameters, and these parameters fell into ranges predicted in the theoretical development. Several of the acoustic measures were found to have relationships predictable from theory. In particular, the measure of spectral tilt was strongly correlated with measures of aspiration noise ($p > 0.70$). Also, the spectrum-based measure of bandwidth had a good to strong correlation with the measures of spectral tilt and noise ($p > 0.65$ in most cases). These results are evidence that the acoustic measures actually reflect glottal parameters.

We used the acoustic measures to classify the speakers according to glottal configuration. Two groups were proposed, one having abrupt glottal closures and glottal chinks limited to the cartilaginous part of the vocal folds (Group 1), and the other having nonsimultaneous glottal closures, with posterior openings possibly extending beyond the vocal processes into the membranous part of the folds (Group 2). Group 1 speakers were characterized by narrower first-formant bandwidths and relatively shallow spectral tilts, while Group 2 speakers had wider bandwidths and steeper spectral tilts.

The validity of this classification was explored through physiological measures made on four of the 22 speakers. These measures included glottal waveform parameters obtained by inverse filtering of oral airflow and observation via fiberscopy of the vocal folds during phonation. Although there were some difficulties with collecting the aerodynamic data, our results showed several expected trends. First, Group 2 speakers had higher minimum flows, supporting the theory that these speakers have larger glottal openings than Group 1 speakers. Second, there was a trend for speakers with higher values of open quotient to have higher values of $H1^* - H2^*$, the difference between the amplitudes of the first two harmonics. Third, there was a tendency for speakers having greater speed quotients to have lower measures of spectral tilt as measured by the difference between the amplitudes of the first harmonic and the third-formant peak, $H1^* - A3^*$. One speaker, however, did not follow the latter two trends. Also as expected, speakers with greater H1*-A1 measures had greater AC flow to MFDR ratios.

The first-formant bandwidth as measured from the speech waveform at the beginning of a vibratory cycle should be similar to the bandwidths calculated from the measured minimum flow and subglottal pressure. However, this relation did not occur, and is probably due to sources of error for the minimum flow, including a leak between the subject's face and the Rothenberg mask (Holmberg et al., in press), and DC flow offsets that can be due to vertical movements of the vocal folds (Hertegård et al., 1992).

The results of the fiberscopy showed differences between the Group 1 and Group 2 speakers. The Group 1 speakers had either complete closures along the entire length of the vocal folds or small openings at the arytenoid cartilages. Their opening and closing movements seemed to be simultaneous along the membranous part of the folds. However, Group 2 speakers had large openings at the posterior end of the glottis that seemed to extend into the membranous part of the folds. For one speaker the closing movement appeared to be nonsimultaneous along the length of the folds.

The trends found for both types of physiological measures, then, support our hypotheses, based on the acoustic measures, about the speakers' glottal configurations.

The hypothesized difference in vocal fold configurations would also predict that Group 2 speakers have a breathier voice quality than do Group 1 speakers. We performed a listening test to explore this possibility. Four listeners were asked to rate vowels excised from the speech of our subjects for breathiness. Despite the difficulty that the listeners had making ratings on stimuli of such short duration, Group 2 speakers were perceived to be breathier than Group 1 speakers. The acoustic measures $H1^* - A1$, $H1^* - A3^*$, and N_S (spectrum-based noise rating) were found to be strongly correlated with the breathiness ratings. The latter two measures were also found to be correlated with perceived breathiness by Klatt and Klatt (1990) and Kasuya and Ando (1991).

The wide ranges of values that we observed in our acoustic measurements suggest that consideration of glottal characteristics has great importance for describing female speech. In addition to formant frequencies and fundamental frequency, these glottal attributes should be taken into account for applications such as speech synthesis and speech or speaker recognition. To test this hypothesis we ran a listening test using synthesized speech in which only the glottal parameters were varied. We found that variations in glottal parameters could be perceived, and if the parameters are set correctly, the synthesis of voice quality for a given speaker is improved. Moreover, a range of parameter values are needed to synthesize individual voices that are perceived to have varying degrees of breathiness. We also found evidence that some glottal parameters of the synthesizer have a greater influence on voice quality than others, particularly the source spectral tilt parameter.

6.2 Suggestions for future work

The use of the acoustic measures to guide synthesis of a particular speaker is promising, but more work needs to be done to refine this procedure. We especially need to investigate the use of the spectrum-based first-formant bandwidth instead of the waveform-based measure to set the synthesizer bandwidth parameter. While the measurements that we made were on steady-state vowels, we should also examine the variation of these parameters throughout an utterance. The glottal parameters are expected to vary during consonant/vowel and vowel/consonant transitions, particularly for obstruent consonants for which the glottal configuration during the consonant can influence the glottal configuration at the edges of the vowel. As we discussed in Chapter 4, this influence can extend well into the vowel (Gobl and Ní Chasaide, 1988; Ní Chasaide and Gobl, 1993), particularly for voiceless consonants. The resulting glottal variations in time may well vary from speaker to speaker, and their inclusion when synthesizing speech could influence the resulting naturalness.

The variation in glottal parameters is also influenced by prosody, and this interaction could be studied using our acoustic measures. The relation between voice quality and prosody, and its application to speech synthesis and recognition, and to speaker recognition or verification is a interesting avenue of future research.

We have limited our study of these acoustic measures to female subjects. The measures should also be applicable to male speakers, but this question remains to be explored. The measures could provide new ways to study differences between male and female speech. Another limitation of our study was that we only looked at nondisordered speech. It would be interesting to apply these acoustic measures to the speech of people with voice disorders. It is possible that these measures could facilitate the diagnosis of voice disorders.

Another aspect of these acoustic measures that remains to be studied is the variability for a given speaker on different days, or even at different times of day. We saw in Chapter 4 that such variability can be significant. In addition, Holmberg et al. (1994a) found intra-speaker variations in aerodynamic and acoustic measures across repeated recordings. In particular, they found that variation in speaking intensity is systematically related with variations in several of the measures made in Chapter 4, and thus should also influence the acoustic measures. Therefore, in future work, sound pressure level should be measured, and possibly controlled, when making such measures.

6.3 Further implications

As discussed in Holmberg et al. (in press), acoustic measures of the kind that we used in our analysis could supplement or replace aerodynamic measures, which can be difficult to collect and analyze. In particular, the use of acoustic measures could be of value in a clinical setting where equipment such as a Rothenberg mask is not available. The ability to make these measures may serve other uses in a clinical setting, such an aid to voice therapy.

The range of values that we obtained for the acoustic measures across speakers (in Chapter 3) and the ability of listeners to perceive these differences (in Chapter 5) suggest that these descriptors should be investigated as possible features for speaker, and even speech, recognition. The collection of these spectrum-based measures should be easily automated, making the use of these features in such applications feasible.

In conclusion, the acoustic measures that we have developed have great potential for improving speech-related applications.

Appendix A

Corrections to spectral measures

A.1 Correction for effect of F1 on H1 and H2 (amplitudes of the first and second harmonics)

The vocal tract transfer function, assuming an all-pole model, can be expressed as

$$T(s) = K\frac{s_1 s_1^*}{(s-s_1)(s-s_1^*)} \cdot \frac{s_2 s_2^*}{(s-s_2)(s-s_2^*)} \cdots, \tag{A.1}$$

where K is a constant, $s = \sigma + j\omega$, and $s_n = \sigma_n + j2\pi \mathrm{F}n$, $\mathrm{F}n$ being the nth formant frequency (Stevens, in preparation). The pole/zero plot for this transfer function is shown in Fig. A.1. The magnitude of the transfer function at a point s is computed by taking the reciprocal of the product of the magnitudes of the vectors from the poles to s (Stevens and Bose, 1965). These vectors are also illustrated in Fig. A.1. As can be seen from this figure, when $\sigma = 0$ and $0 \leq \omega < 2\pi\mathrm{F1}$

$$|T(s)| \approx \frac{s_1 s_1^*}{|(j\omega - s_1)(j\omega - s_1^*)|} \tag{A.2}$$

We want to evaluate $T(s)$ along the imaginary axis at F0 and 2F0. If F0 and 2F0 are sufficiently smaller than F1, we can assume $\sigma_1^2 \ll [2\pi(\mathrm{F1} - f)]^2$, that is, $\sigma_1 \approx 0$. Then Eqn. A.2 reduces to

$$|T(f)| \approx \frac{\mathrm{F1}^2}{\mathrm{F1}^2 - f^2} \tag{A.3}$$

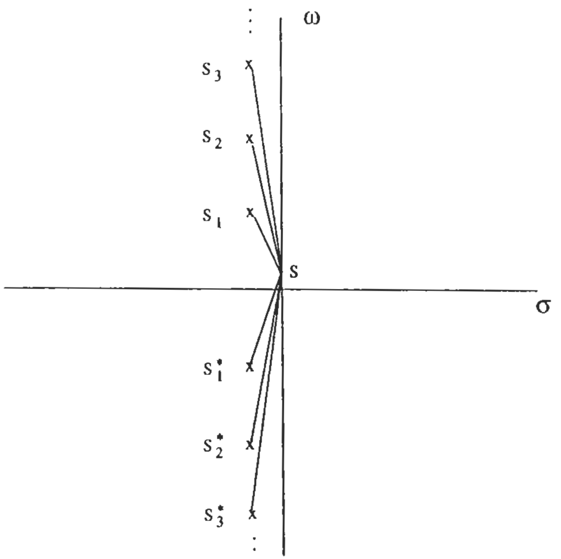


Figure A.1: *Pole/zero plot of the transfer function expressed by Eqn. A.1.*

and we evaluate Eqn. A.3 at f = F0 and f = 2F0 to obtain the corrections for H1 and H2, respectively. To apply this correction to spectra, we subtract

$$20 \log_{10} \frac{\mathrm{F1}^2}{\mathrm{F1}^2 - f^2} \tag{A.4}$$

from H1 and H2.

A.2 Correction for effect of F1 and F2 on A3 (the amplitude of F3)

Referring to Eqn. A.1 and Fig. A.1, $\sigma = 0$,

$$|T(s)| \approx \frac{s_1 s_1^*}{|(\omega - s_1)(\omega - s_1^*)|} \cdot \frac{s_2 s_2^*}{|(\omega - s_2)(\omega - s_2^*)|} \cdot \frac{s_3 s_3^*}{|(\omega - s_3)(\omega - s_3^*)|} \cdots \tag{A.5}$$

If F1 and F2 are sufficiently less than F3, we can assume $\sigma_1 = \sigma_2 \approx 0$. Then

$$|T(s = j2\pi\mathrm{F3})| \approx \frac{\mathrm{F1}^2}{(\mathrm{F1}^2 - \mathrm{F3}^2)} \cdot \frac{\mathrm{F2}^2}{(\mathrm{F2}^2 - \mathrm{F3}^2)} \cdot \frac{s_3 s_3^*}{|(j2\pi\mathrm{F3} - s_3)(j2\pi\mathrm{F3} - s_3^*)|} \cdots \tag{A.6}$$

To neutralize the magnitude of F3 across vowels, we see from Eqn. A.6 that we must factor *out* the effects of F1 and F2, and then factor *in* the effects of some neutral formant frequencies $\widetilde{\mathrm{F1}}$ and $\widetilde{\mathrm{F2}}$. That is, we multiply $|T(j2\pi\mathrm{F3})|$ by

$$\frac{(\mathrm{F1}^2 - \mathrm{F3}^2)(\mathrm{F2}^2 - \mathrm{F3}^2)}{\mathrm{F1}^2\mathrm{F2}^2} \cdot \frac{\widetilde{\mathrm{F1}}^2\widetilde{\mathrm{F2}}^2}{(\widetilde{\mathrm{F1}}^2 - \mathrm{F3}^2)(\widetilde{\mathrm{F2}}^2 - \mathrm{F3}^2)} \tag{A.7}$$

which reduces to

$$\frac{\left[1 - \left(\frac{\mathrm{F3}}{\mathrm{F1}}\right)^2\right]\left[1 - \left(\frac{\mathrm{F3}}{\mathrm{F2}}\right)^2\right]}{\left[1 - \left(\frac{\mathrm{F3}}{\widetilde{\mathrm{F1}}}\right)^2\right]\left[1 - \left(\frac{\mathrm{F3}}{\widetilde{\mathrm{F2}}}\right)^2\right]} \tag{A.8}$$

To A3, then, we add

$$20\log_{10}\left(\frac{\left[1 - \left(\frac{\mathrm{F3}}{\mathrm{F1}}\right)^2\right]\left[1 - \left(\frac{\mathrm{F3}}{\mathrm{F2}}\right)^2\right]}{\left[1 - \left(\frac{\mathrm{F3}}{\widetilde{\mathrm{F1}}}\right)^2\right]\left[1 - \left(\frac{\mathrm{F3}}{\widetilde{\mathrm{F2}}}\right)^2\right]}\right) \tag{A.9}$$

Appendix B

Instructions for synthesis test

The following are the instructions given to the listeners for the voice quality synthesis test described in Chapter 5:

In this listening test, you will be asked to make judgments on stimuli that have been synthesized to match certain voice qualities. Each stimulus is of the form AXB, where X is a reference token and A and B are the test items. The tokens are presented in the following order:

token A   reference X   token B

In this case, X is the word 'bud' as uttered by one of six women. A and B are two synthesized versions of X, the reference word. The only difference between A and B is that they were synthesized with different values of the synthesizer parameters that control voice quality. Your job is to choose which test token, A or B, is a better match to X in terms of voice quality.

The voice quality of the stimuli you will hear varies on a continuum from very pressed to very breathy. Pressed phonation tends to sound intense, sharp, and clear. Breathy phonation is less intense, somewhat noisy, and may even sound muffled. You will hear words that have been synthesized with varying degrees of breathy and pressed phonation.

When you listen to the three tokens that make up each stimulus, concentrate on the vowel in 'bud' and try to compare the voice quality of the test items A and B to that of the reference X. This can be difficult. Some subjects have found that by closing their eyes as they listened, they were able to more easily make a decision.

The difference between A and B may be quite obvious for some stimuli, for example if A has a very pressed phonation and B has a very breathy phonation. But often the difference is quite subtle, such as when you are comparing two versions of 'bud' that are both breathy, but to varying degrees. Don't worry if you don't hear a difference between A and B—it's OK to guess in such a case.

The stimuli are presented in groups of six, with 3.5 seconds between stimuli for you to write down your answer, and then 5 seconds of silence between groups. Each of the six stimuli in a group represents a different speaker.

Bibliography

[1] Ananthapadmanabha, T. V. (1984). "Acoustic analysis of voice source dynamics," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 2–3, 1–24.

[2] Ananthapadmanabha, T. V. (1993). "Modeling the return phase of the derivative of glottal flow," *Speech Communication Group Working Papers*, 9, Cambridge: Massachusetts Institute of Technology, 28–34.

[3] van den Berg, J., Zantema, J. T., and Doornenbal, P. (1957). "On the air resistance and the Bernoulli effect of the human larynx," *Journal of the Acoustical Society of America*, 29, 626–631.

[4] Bickley, C. A. and Stevens, K. N. (1986). "Effects of a vocal-tract constriction on the glottal source: Experimental and modelling studies," *Journal of Phonetics*, 14, 373–382.

[5] Bickley, C. A. and Stevens, K. N. (1987). "Effects of a vocal-tract constriction on the glottal source: Data from voiced consonants," in T. Baer, C. Sasaki, and K. Harris (eds.), *Laryngeal Function in Phonation and Respiration*, Boston: Little, Brown and Co.

[6] Biever, D. M. and Bless, D. M. (1989). "Vibratory characteristics of the vocal folds in young adult and geriatric women," *Journal of Voice*, 3, 120–131.

[7] Bose, A. G. and Stevens, K. N. (1965). *Introductory network theory,* New York: Harper & Row.

[8] Cranen, B. and Boves, L. (1985). "Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production," *Journal of the Acoustical Society of America*, 77, 1543–1551.

[9] Cranen, B. and Boves, L. (1988). "On the measurement of glottal flow," *Journal of the Acoustical Society of America*, 84, 888–900.

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

[11] Fant, G. (1962). "Formant bandwidth data," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 1, 1–3.

[12] Fant, G. (1972). "Vocal tract wall effects, losses, and resonance bandwidths," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 2–3, 28–52.

[13] Fant, G. (1979). "Glottal source and excitation analysis," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 1, 85–107.

[14] Fant, G. (1982). "The voice source—acoustic modelling," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 4, 28–48.

[15] Fant, G. (1993). "Some problems in voice source analysis," *Speech Communication*, 13, 7–22.

[16] Fant, G., and Ananthapadmanabha, T. V. (1982). "Truncation and superposition," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 2–3, 1–17.

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

[18] Fant, G., Liljencrants, J., and Lin, Q. (1985). "A four-parameter model of glottal flow," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 4, 1–13.

[19] Fant, G., and Lin, Q. (1988). "Frequency domain interpretation and derivation of glottal flow parameters," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 2–3, 1–21.

[20] Fujimura, O. and Lindqvist, J. (1971). "Sweep-tone measurements of vocal-tract characteristics," *Journal of the Acoustical Society of America*, 49, 541–558.

[21] Gelfer, M. P. (1993). "A multidimensional scaling study of voice quality in females," *Phonetica*, 50, 15–27.

[22] Gobl, C. (1989). "A preliminary study of acoustic voice quality correlates," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 4, 9–22.

[23] Gobl, C. and Ní Chasaide, A. (1988). "The effects of adjacent voiced/voiceless consonants on the vowel voice source: A cross language study," *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report*, 2–3, 23–59.

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

[25] Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, 37, 769–778.

[26] Hirano, M., Kiyokawa, K., and Kurita, S. (1988). "Laryngeal muscles and glottic shaping," in O. Fujimura (ed.), *Vocal fold physiology: voice production, mechanisms and functions*, New York: Raven Press.

[27] 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," *Journal of the Acoustical Society of America*, 84, 511–529, plus "Erratum," *JASA*, 85, 1787.

[28] Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1989). "Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch," *Journal of Voice*, 4, 294–305.

[29] Holmberg, E. B., Hillman, R. E., Perkell, J. S., and Gress, C. (1994a). "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.

[30] Holmberg, E. B., Perkell, J. S., Hillman, R. E., and Gress, C. (1994b). "Individual variation in measures of voice," *Phonetica*, 51, 30–37.

[31] Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P., and Goldman, S. L. (in press). "Comparisons among aerodynamic, electroglottographic, and acoustic spectrum measures of female voice" to appear in *Journal of Speech and Hearing Research.*

[32] House, A. S., and Stevens, K. N. (1958). "Estimation of formant band widths from measurements of transient response of the vocal tract," *Journal of Speech and Hearing Research,* 1, 309–315.

[33] Huffman, M. K. (1987). "Measures of phonation type in Hmong," *Journal of the Acoustical Society of America*, 495–504.

[34] Isshiki, N. (1964). "Regulatory mechanism of voice intensity variation," *Journal of Speech and Hearing Research,* 7, 17–29.

[35] Karlsson, I. (1986). "Glottal wave forms for normal female speakers," *Journal of Phonetics*, 14, 415–419.

[36] Karlsson, I. (1988). "Glottal waveform parameters for different speaker types," *Proceedings of Speech '88*, 7 FASE Symposium, Edinburgh, 225–231.

[37] Karlsson, I. (1990). "Voice source dynamics for female speakers," in *Proceedings of the International Conference on Spoken Language Processing 1990*, Kobe, Japan, 69–72.

[38] Karlsson, I. (1991a). "Dynamic voice quality variations in natural female speech," *Speech Communication*, 10, 481–490.

[39] Karlsson, I. (1991b). "Female voices in speech synthesis," *Journal of Phonetics*, 19, 111–120.

[40] Karlsson, I. (1992a). *Analysis and synthesis of different voices with emphasis on female speech,* Unpublished doctoral dissertation, Royal Institute of Technology, Stockholm.

[41] Karlsson, I. (1992b). "Modelling voice variations in female speech synthesis," *Speech Communication*, 11, 1–5.

[42] Kasuya, H. and Ando, Y. (1991). "Acoustic analysis, synthesis, and perception of breathy voice," in J. Gauffin and B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, San Diego: Singular.

[43] Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S. (1986). "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *Journal of the Acoustical Society of America*, 80, 1329–1334.

[44] Kiritani, S., Imagawa, H., and Hirose, H. (1990). "Vocal cord vibration and voice source characteristics: Observations by a high-speed digital image recording," in *Proceedings of the International Conference on Spoken Language Processing 1990*, Kobe, Japan, 61–64.

[45] Klatt, D., and 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.

[46] Klingholz, F. (1987). "The measurement of the signal-to-noise ratio (SNR) in continuous speech," *Speech Communication,* 6, 15–26.

[47] de Krom, G. (1993). "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, 36, 254–266.

[48] Ladefoged, P. (1962). "Subglottal activity during speech," in *Proceedings of the Fourth International Congress of Phonetic Sciences*, The Hague: Mouton, 73–91.

[49] Ladefoged, P. and Antoñanzas-Barroso, N. (1985). "Computer measures of breathy voice quality," *Working Papers in Phonetics*, 61, University of California at Los Angeles, 79–86.

[50] Linville, S. (1992). "Glottal gap configurations in two age groups of women," *Journal of Speech and Hearing Research*, 35, 1209–1215.

[51] Mori, K., Blaugrund, S. M., and Yu, J. D. (1994). "The turbulent noise ratio: an estimation of noise power of the breathy voice using PARCOR analysis," *Laryngoscope,* 104, 153–158.

[52] Ní Chasaide, A. and Gobl, C. (1993). "Contextual variation of the vowel voice source as a function of adjacent consonants," *Language and Speech*, 36, 303–330.

[53] Peppard, R. C., Bless, D. M., and Milenkovic, P. (1988). "Comparison of young adult singers and nonsingers with vocal nodules," *Journal of Voice*, 2, 250–260.

[54] Perkell, J. S., Holmberg, E. B., and 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.

[55] Perkell, J. S., Hillman, R. E., and Holmberg, E. B. (1994). "Group differences in measures of voice production and revised values of maximum airflow declination rate," *Journal of the Acoustical Society of America*, 96, 695–698.

[56] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Englewood Cliffs, N.J.: Prentice-Hall.

[57] Rammage, L. A., Peppard, R. C., and Bless, D. M. (1992). "Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: A retrospective study," *Journal of Voice*, 6, 64–78.

[58] Rothenberg, M. R. (1973). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *Journal of the Acoustical Society of America*, 72, 633–634.

[59] Shadle, C. (1985). "The acoustics of fricative consonants," *RLE Technical Report 506*, Cambridge: Massachusetts Institute of Technology.

[60] Södersten, M. and Hammarberg, B. (1993). "Effects of voice training in normal-speaking women: Videostroboscopic, perceptual, and acoustic characteristics," *Scandinavian Journal of Logopedics & Phoniatrics*, 18, 33–42.

[61] Södersten, M. and Lindestad, P-Å. (1990). "Glottal closure and perceived breathiness during phonation in normally speaking subjects," *Journal of Speech and Hearing Research*, 33, 601–611.

[62] Södersten, M., Lindestad, P-Å. and Hammarberg, B. (1991). "Vocal fold closure, perceived breathiness, and acoustic characteristics in normal adult speakers," in J. Gauffin and B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, San Diego: Singular.

[63] Stevens, K. N. (1971). "Airflow and turbulence noise for fricative and stop consonants," *Journal of the Acoustical Society of America*, 50, 1180–1192.

[64] Stevens, K. N. (1993). "Models for the production and acoustics of stop consonants," *Speech Communication*, 13, 367–375.

[65] Stevens, K. N. (1994). "Scientific substrates of speech production," in F. Minifie (ed.), *Introduction to Communication Sciences and Disorders*, San Diego: Singular.

[66] Stevens, K. N., and Hanson, H. M. (1995). "Classification of glottal vibration from acoustic measurements," in O. Fujimura and M. Hirano (eds.), *Vocal Fold Physiology: Voice Quality Control,* San Diego: Singular.

[67] Teager, H. M. (1980). "Some observations on oral air flow during phonation,"*IEEE Transactions on Acoustic, Speech, and Signal Processing*, ASSP-28, 599–601.

[68] Teager, H. M. (1983a). "The effects of separated air flow on vocalizations," in D. M. Bless and J. H. Abbs (eds.), *Vocal Fold Physiology*, San Diego, College-Hill.

[69] Teager, H. M. (1983b). "Active fluid dynamic voice production models, or is there a unicorn in the garden," in I. R. Titze and R. C. Scherer (eds.), *Vocal Fold Physiology*, Denver: Denver Center for the Performing Arts.

[70] Titze, I. R. (1989a). "Physiologic and acoustic differences between male and female voices," *Journal of the Acoustical Society of America*, 85, 1699–1707.

[71] Titze, I. R. (1989b). "A four-parameter model of the glottis and vocal fold contact area," *Speech Communication*, 8, 191–201.

[72] Yumoto, E., Gould, W. J., and Baer, T. (1982). "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Journal of the Acoustical Society of America*, 71, 1544–1550.

[73] Zemlin, W. R. (1988). *Speech and hearing science: Anatomy and physiology,* Englewood Cliffs, N.J.: Prentice Hall.