CHAPTER 4

DATA AND MEASUREMENTS

4.1 MEASUREMENTS OF LARYNGEAL FUNCTION

The larynx, just over the tongue and below the epiglottis, is so near yet so far away. We can feel it with our fingers as the "Adam's apple." Perhaps it is because it is so "near" that we feel particularly exasperated when we are frustrated in our attempts to observe the vibratory motion of the vocal folds during normal speech. Many methods have been used to study the normal and diseased larynx. Here we provide an overview of major techniques used to evaluate and measure laryngeal function. Our major categories (outlined in Table 4.1) are arbitrary, but functional. The first, a collection of methods aimed at displaying the movement of the vocal folds, is observational; these methods employ various wavelengths across the spectrum, that is, ultrasonic, infrared, visible, electron beam, and x-ray, to achieve this goal. The next category is glottography, which includes techniques that indirectly measure aspects of the glottal opening by either electrical or optical methods. The third and final category encompasses techniques that attempt to extract parameters related to laryngeal function by processing the acoustic or speech waveform, by measuring or evaluating aural parameters, or by determining the airflow and pressure at various points in the vocal system. See Childers (1977) for additional references to the literature in this area. Our coverage is neither exhaustive nor all inclusive; we do not discuss various biochemical laryngeal evaluative methods.

4.2 OBSERVATIONAL

4.2.1 Direct and Indirect Laryngoscopy

The larynx has been directly viewed through a puncture in the throat, but this is extremely rare for scientific studies. Laryngeal function is primarily observed indirectly with the aid of a laryngoscope or mirror. This was first attempted in 1807 by Bozzini, the inventor of the laryngoscope, a two-channel tube similar to a periscope (Childers, 1977, pg. 381). In Bozzini's day, one channel allowed light from a candle to enter and be reflected from a mirror onto the vocal folds. The adjacent channel was designed to provide a pathway for the reflected light to return to the viewer. However, the mechanism was not successful, probably because of inadequate illumination. By 1860, physicians world-wide were attempting to use laryngoscopy and to improve on the procedure; however, the principal factor in the success of indirect laryngoscopy in clinical practice has been the evolution of improved lighting conditions.

Indirect laryngoscopy is simple to use, inexpensive, and requires no elaborate equipment or data processing. Unfortunately, the procedure provides no permanent record useful

TABLE 4.1. Laryngeal Evaluative Methods

4.2 Observational
- 4.2.1. Direct and indirect laryngoscopy
- 4.2.2. Microscopy using light and electron beam
- 4.2.3. Still photography, stroboscophy, and cinematography: black and white, color, and infrared
- 4.2.4. Fiberoptics
- 4.2.5. Ultra high-speed cinematography
- 4.2.6. Radiography: x-ray, stroboscopic laminography, tomography, cinefluourography, and magnetic resonance imaging
- 4.2.7. Ultrasound

4.3 Glottography: optical and electrical
- 4.3.1. Photoelectric (or optical) glottography
- 4.3.2. Electroglottography, laryngography
- 4.3.3. Electromyography

4.4 Acoustic, aural, and airflow
- 4.4.1. Inverse filtering
- 4.4.2. Acoustical measurements
- 4.4.3. Spectrograms
- 4.4.4. Perturbations: jitter, shimmer
- 4.4.5. Fundamental frequency
- 4.4.6. Intensity
- 4.4.7. Vocal quality
- 4.4.8 Glottal airflow and subglottal pressure

for medical history or therapy. Furthermore, when the laryngoscope or mirror is used in a subject who can tolerate the instrument, the subject is able to phonate only a very limited range of sounds, usually sustained vowels, since the vocal folds are otherwise hidden by the tongue or epiglottis. In addition, the mirror may interfere with normal vocal activity and distort the sound being phonated. Since the vocal folds vibrate rapidly, the viewer is not able to follow their detailed motion by use of laryngoscopy alone.

4.2.2 Laryngeal Microscopy

This procedure has used both light and electron beams to study laryngeal functioning, but is not widely used, even in scientific studies.

4.2.3 Still Photography, Stroboscopy, and Cinematography

In order to overcome some of the disadvantages of indirect laryngoscopy, investigators have used photography in various forms. The developed film provides a permanent record that can be analyzed for scientific details. Photography offers the investigator a means by which he or she can "stop" the vibratory motion of the vocal folds.

The first known attempt to photograph the larynx was in 1860. This method is still used today, primarily to preserve images obtained by indirect laryngoscopy, but the method is also useful as a teaching tool, for documentation, and as a basis for comparison of pre- and post-treatment conditions for various laryngeal diseases. The record is easily obtained,

relatively inexpensive, and available to educators and clinicians alike. The technique, however, is dependent upon indirect laryngoscopy and its accompanying inherent problems. Finally, the image on each frame of the film is a blurred composite of the many vocal cord vibrations that occurred during the exposure of each successive film frame.

Stroboscopic laryngoscopy illuminates the vibrating vocal folds via the laryngeal mirror with an intermittent or flashing light providing the illusion that the rapid vibratory motion of the vocal folds has been stopped or slowed. The stroboscope was first suggested as a method for investigating the movement of the vocal folds in 1866, but was not actually tried until 1878. Despite improvements in stroboscopic laryngoscopy, its primary limitation is that the investigator sees only brief "snapshots" of consecutive cycles of the vibratory pattern; he or she never sees a complete vibratory cycle. The image on the film is a composite of the vocal folds illuminated by the strobe at consecutive phases or cycles in the vibratory pattern of the vocal folds. Because of this, the composite image may contain artifacts not easily recognized. For example, the vocal folds may vary slightly in periodicity or their opening or closing pattern may vary over successive illuminations by the strobe. The image on the film represents a composite of these variations. Thus, measurements taken from the film may be an incorrect representation of the actual activity that took place during the vibratory cycle. Since the vibratory motion of the vocal folds is usually aperiodic, stroboscopy is not a particularly effective means for analyzing vocal cord function; nevertheless, it has been, and is still, used in a clinical setting despite the fact that it possesses the general limitations of indirect laryngoscopy. This technique will probably not be used in the future since laryngeal research has progressed to a much higher level of sophistication. Current investigative procedures require considerably more quantifiable detail concerning the motion of the individual cycles of the vibratory pattern of the vocal folds. Stroboscopy cannot supply these data.

Cinematography or regular-speed motion picture photography, using cameras with a film rate of approximately 24 frames per second, was first employed to film the larynx in 1913. Motion pictures, combining the principles of stroboscopy with those of laryngeal photography, have also been taken of the larynx.

Regular-speed cinematography also provides a permanent record of the gross function and pathologies of laryngeal structures, and the films have proven useful in both diagnosis and follow-up treatment in a clinical environment. The procedure is relatively simple to use and is not particularly expensive. The primary disadvantage of this technique has already been described, that is, there is no quantifiable data concerning glottal dimensions or vibratory movements are available from the films.

The three types of film that have been used in the photography techniques are black and white, color, and infrared. Color film provides an almost exact representation of the condition of the tissue; however, the speed at which color film can be exposed is generally less than that for black and white. Infrared is not used frequently, but it does tend to enhance those areas on the vocal folds where an excessive inflammation or concentration of blood may be present, which, in turn, raises the temperature of the surrounding tissue.

4.2.4 Fiberoptics

This subcategory is not unique, but rather a result of a recent development that provides an alternate method for illuminating, viewing, and photographing laryngeal function. The primary advantage that fiberoptics offer is that the fiberoptic bundle can be inserted through the nasal passages, allowing the subject or patient to articulate normal speech. The insertion of the bundle can, however, be painful. The vibratory pattern of the vocal folds is filmed through the eye piece of the fiberoptic bundle with film rates from 24 to 64 frames per second having been achieved.

A special fiberoptic laryngoscope has also been designed for use with high-speed stroboscopic light and a high-speed camera. Fiberoptics have also been used in microlaryngoscopy and microsurgery, for supraglottal illumination during glottography in conjunction with the standard laryngeal mirror and for illumination with videotaping.

As mentioned, the unique advantage of fiberoptics in illumination and photography is that the larynx can be observed during the production of normal connected speech. The primary disadvantage appears to be that, at the present time, only 64 frames per second has been achieved in cinematography, but research is underway to improve the light source so that eventually ultra high-speed photography should be possible with fiberoptic systems.

4.2.5 Ultra High-Speed Cinematography

While regular-speed cameras operate at 24 frames per second, ultra high-speed motion picture photography is capable of filming speeds from 4000 to 10,000 (or higher) frames per second. Such rates are required to achieve a detailed slow-motion study of the vibratory pattern of the vocal folds. The first use of ultra high-speed cinematography in laryngeal research was reported in 1937 at Bell Telephone Laboratories and by Moore (1936), in a Ph D dissertation, using an apparatus similar to that illustrated in Figure 4.1. The analysis of these films launched numerous subsequent investigations. The Bell Telephone Laboratories' (Murray Hill, NJ) films achieved good photographs at a rate of 4000 frames per second.

The significance of these rates can be illustrated as follows. The ratio of the rate of exposure to the rate of projection is the factor at which the vibratory motion is slowed, for example, if the rate of exposure is 4000 frames per second and the rate of projection is 16, then the motion is slowed by $4000/16 = 250$. Thus at these rates, if the vocal folds are performing 125 vibrations per second, they then appear to make one-half vibration per second, or one vibration every 2 seconds when the film is projected onto the screen. With the exception of the work using fiberoptic bundles, ultra high-speed cinematography uses indirect laryngoscopy, typically employing a laryngeal mirror. Both black and white and color film is used.

The advantages of ultra high-speed cinematography are that the vibratory pattern of the vocal folds can be viewed in slow motion, and that a detailed frame-by-frame analyses

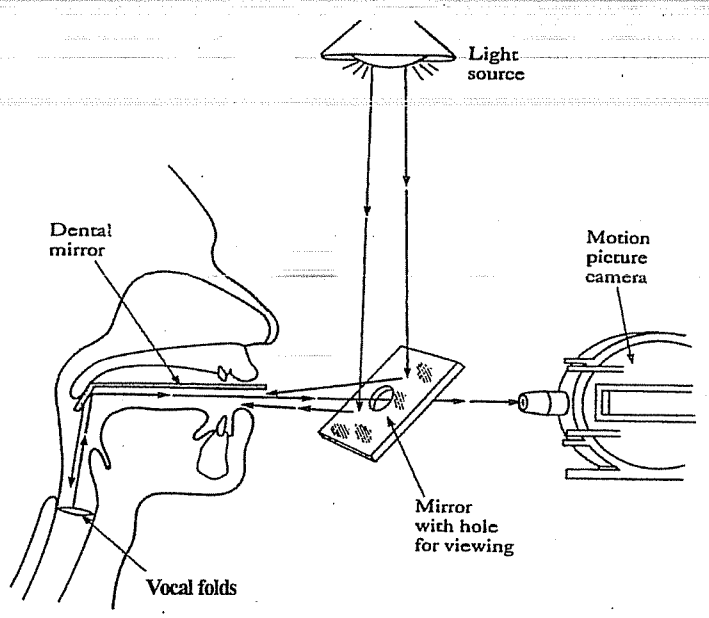

FIGURE 4.1 An outline of ultra high-speed laryngeal filming technique.

can be made. The vocal folds are typically filmed while the subject or patient is phonating a sustained vowel, but it is possible to film nonlinguistic phenomena such as a cough, laughter, or clearing of the throat.

4.2.6 Radiography

This procedure employs x-rays as an agent to expose a film plate that can be developed to yield an image of the larynx and the surrounding structure. Since the vocal folds are soft tissue, the laryngeal images are generally of poor quality, due to the low radiographic contrast. The cervical vertebrae often obscure the laryngeal image in a frontal or anteroposterior view. These techniques have found their most use recording the motion of the articulators. The outlining of the soft tissue of the tongue and velum is facilitated by gluing lead pellets to the tongue and palate. These pellets are quite visible in the films.

The first x-rays of the larynx were attempted in 1913 from a lateral view. The procedure has undergone many changes since that time. Frontal and lateral views have been taken along with laminographic or tomographic (body sectioning) procedures, as well as cineradiographic or cinefluouragraphy. Radiography will not become a major investigative tool in laryngeal research. It is expensive, offers a risk to the subject due to the cumulative radiation dosage, and relevant data are difficult to obtain. The procedure is not able to provide data about the glottal area or the vibratory frequency, two parameters presently considered essential in laryngeal research.

Magnetic resonance imaging has been used in recent years because it offers less risk than x-ray exposure and it can readily capture the motion of the articulators, particularly when the pellets mentioned above are used.

4.2.7 Ultrasound

Ultrasonic systems have been used in a manner analogous to x-ray systems to obtain images of the skeletal structure within the body. However, only limited success has been achieved in laryngeal research due to the soft tissue in the vocal folds. In the hope that greater success might be achieved in another way, investigators have used ultrasonic systems to measure the velocity of vocal cord motion during speech by detecting the doppler shift caused by the reflection of the ultrasonic wave from the surface of the vocal fold.

Glottal closure has also been measured by using a source on one side of the neck and a transducer on the other. When the folds are open, the air reflects the ultrasonic wave; when the folds are closed, the wave passes through the tissue from the source to the transducer.

The procedure is easy to apply and the patient experiences no discomfort and can speak normally; thus, the tongue and epiglottis do not interfere with the data-collection process. It has been noted that most commercially available equipment does not generate ultrasound frequencies sufficiently high enough to provide adequate resolution of vocal cord motion. It is unlikely that ultrasound will be used extensively in laryngeal research since the data obtained by this procedure are limited to information about glottal closure.

4.3 GLOTTOGRAPHY: OPTICAL AND ELECTRICAL

The techniques to be discussed in this section are indirect methods for monitoring laryngeal function. The data obtained are generally related to glottal area or closure. These procedures

have, and are being, used because they are simple to apply and overcome many of the disadvantages of the observational techniques discussed previously. However, as will be pointed out, these indirect procedures also have their limitations.

4.3.1 Optical Glottography

Perhaps the major impetus for the interest in optical or photoelectric glottography is the hypothesis that the volume–velocity of airflow at the glottis is proportional to the glottal area. Under this hypothesis, one aspect of laryngeal function may be measured using a simpler procedure than photographic techniques. For example, optical glottography attempts to determine glottal function by measuring the amount of light that passes through the glottis. The light source may either be above the vocal folds, introduced through the mouth or nasal passages, or introduced through the neck below the larynx. In the former case, the light that passes through the glottis is usually monitored by a photocell next to the neck below the larynx. In the latter case, a curved light-conducting rod or fiberoptic bundle might be introduced through the oral cavity with the photocell attached to the end of the rod or bundle outside the lips. As the vocal folds vibrate, the light is modulated and picked up by the photocell whose electrical output may be displayed on an oscilloscope or a strip chart recorder or recorded on magnetic tape.

The results of this technique have sometimes agreed with and sometimes differed from data measured by ultra high-speed films on a frame-by-frame basis. A possible cause for this is that the light can be reflected in irregular and unpredictable ways within the throat and from the mucosal surfaces of the vocal folds. Even the placement of the photocell may affect the monitored waveform.

This method cannot account for vocal fold movements away from or toward the light source, but the ease with which this method can be applied makes the procedure very attractive. It is relatively inexpensive and can be used in a clinical setting. If the artifacts can be isolated and the data substantiated through the simultaneous application of other procedures, such as ultra high-speed cinematography, then optical glottography may be increasingly used in both research and clinical environments.

4.3.2 Electroglottography

Introduced in 1958 by Fabre, this method measures the variation in electrical impedance across the neck as the vocal folds vibrate (Childers, 1977, p. 385). The impedance increases as the glottis becomes enlarged and decreases as the glottis closes. Thus, the variations in glottal area are related to the variations in impedance.

Electroglottography has been used by several investigators and has been compared to optical glottography and to ultra high-speed photography. The laryngograph is an electroglottographic instrument that depends on the amount of vocal fold contact rather than the width of the glottis. This instrument uses modified electrodes that are applied in a manner different from those previously used.

Electroglottography and laryngography do not measure the glottal area or other glottal-related parameters. The procedure is apparently limited to frequency measurements. The technique is inexpensive, comfortable, easily used, portable, and does not interfere with normal speech. Despite its limitations, it is expected that this procedure will be increasingly applied because it can provide a quick and permanent record in a clinical setting. Figure 4.2 illustrates the procedure for recording the electroglottographic signal.

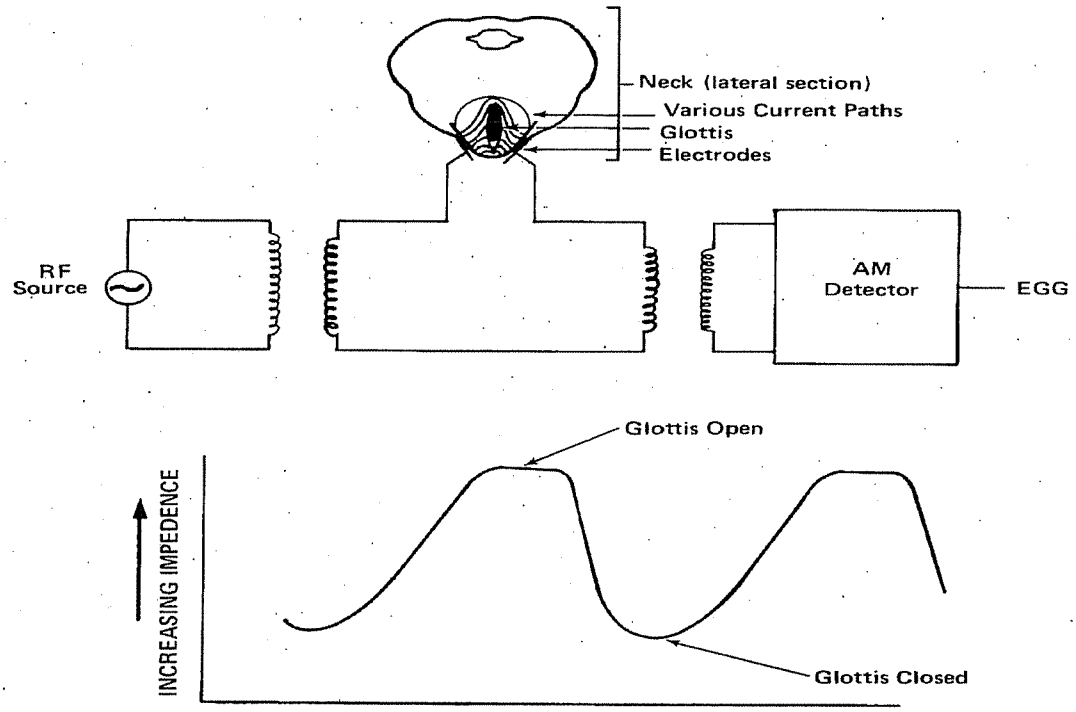

FIGURE 4.2 Electroglottography and the electroglottographic waveform.

4.3.3 Electromyography

Here, the laryngeal muscle activity is monitored. Therefore, this procedure is not a glottographic technique; however, since it is related, the method has been so categorized. The activity of the laryngeal muscles is monitored using needle electrodes inserted into the muscles.

4.4 ACOUSTIC, AURAL, AND AIRFLOW

This category is a collection of loosely related topics. The first and second subcategories cover procedures that attempt to extract parameters related to laryngeal function by processing the acoustic or speech waveform.

4.4.1 Inverse Filtering

We shall shortly discuss this topic at length as part of linear prediction. The theoretical foundation for this procedure is founded on the premise that the glottal volume–velocity of air, modified by the resonance and damping characteristics of the vocal tract, is radiated from the lips and nose as speech. Therefore, if a model of the vocal tract could be constructed, it would be possible to process the speech or acoustic signal in a reverse (or inverse) manner to obtain an estimate of the glottal volume–velocity. In 1959, Miller at Bell Telephone Laboratories (Murray Hill, NJ), was apparently the first to obtain such an estimate. This work spurred additional effort, which attempted to verify that the glottal volume–velocity waveform obtained by inverse filtering was a true representation of the airflow at the vocal folds. Thus, both optical and electroglottograms were taken and shown to be quite similar to the waveforms obtained by inverse filtering.

Rothenberg in 1973 modified the inverse filtering method to obtain an estimate of the volume–velocity of air at the mouth by using a pneumotachographic mask to measure the

pressure differential across the mask's screen. The signal Rothenberg obtained was inverse filtered to yield an estimate of the glottal volume–velocity. Prior to the introduction of Rothenberg's method, the acoustic pressure wave was inverse filtered. However, Rothenberg's procedure yielded a signal that was not as susceptible to low frequency noise, was more accurate at low frequencies (even down to zero frequency), and was easily calibrated in amplitude.

Another modification to inverse filtering was introduced by Sondhi in 1975, who pointed out that the glottal volume–velocity waveform could be measured as follows. The subject inserts a hard-walled uniform tube into his mouth; the vocal tract and tube cross-sections are identical, and the tube termination is made reflectionless. The glottal source, therefore, feeds into a uniform lossless tube that is effectively infinite in extent. A probe microphone inserted anywhere in the tube will pick up a waveform identical in shape to the glottal volume–velocity waveform, but slightly delayed. This method has not yet been used extensively by others. The tube diameter must be matched to the vocal tract cross-section of the subject. Furthermore, this procedure cannot be applied to recorded speech. The subject must be present.

The concept of inverse filtering, as introduced by Miller, is receiving considerable attention, since it can generally be applied to recorded speech. It is simple and safe to use, the subject experiences no discomfort, and the speech may be continuous.

4.4.2 Acoustic Measurements

In acoustic measurements, intensity levels are often measured with respect to the reference intensity of 10^{-12} watts/m^2 or 0.0002 dyne/cm^2, which is the threshold of hearing. This is sometimes called the zero-reference level for acoustic sound-pressure level (SPL) measurements, which is the lowest SPL we can perceive.

$$\text{intensity level} = 10\log(P_1/P_2)\,\text{dB} \qquad (4.4.2.1)$$

The threshold of feeling, the loudest level before discomfort begins is 120 dB SPL. Table 4.2 provides a summary of various SPLs.

Microphones are typically used for SPL measurements. They typically have a dynamic range of 125 dB. A carbon microphone is one that has found frequent use in telephone handsets. A simple diagram is shown in Figure 4.3. The diaphragm compresses a loose pack

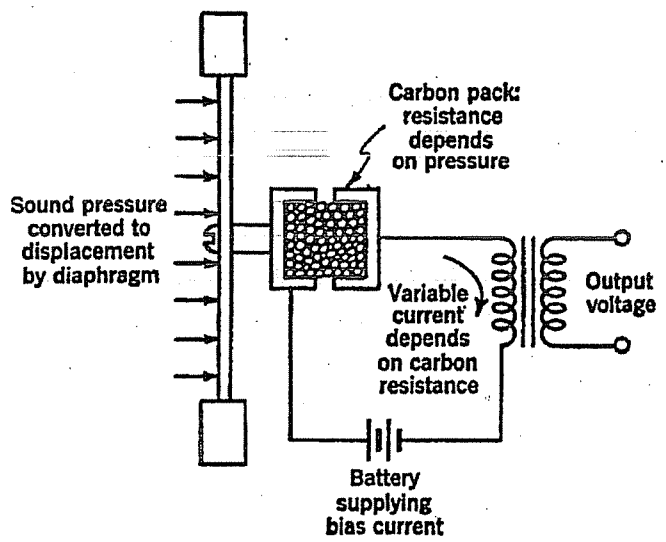

FIGURE 4.3 A carbon microphone.

TABLE 4.2. Sound-Pressure Levels for Various Conditions and Events (dB SPL)

12-inch Cannon at muzzle	220
	210
Rocket engines	200
	190
	180
	170
	160
Jet engine	150
Threshold of pain	140
	130
Threshold of feeling	120
Thunder	110
Niagara Falls	100
Subway	90
Factory	80
Busy street	70
	60
Office	50
Audience noise	40
Quiet home	30
Recording studios	20
	10
Threshold of hearing	0

of carbon granules, thereby increasing their area of contact and decreasing the electrical resistance of the pack. Thus, sound-pressure waves cause the resistance to fluctuate. If a constant current is passed through the carbon granules, the voltage drop across the pack is inversely related to the instantaneous sound pressure.

Another common microphone is a moving coil microphone, illustrated in Figure 4.4. Here a coil is moved in a magnetic field by the sound pressure to induce a voltage.

A third microphone is the condenser microphone shown in Figure 4.5. Here the sound pressure acts on a diaphragm to vary a capacitance (condenser). This type of microphone requires a current source. It is considered an excellent microphone and is often used in speech, singing, and music research studies.

4.4.3 Spectrograms

The spectrogram provides a time history of the spectrum, as shown in Figure 4.6, for the sentence "We were away a year ago," as spoken by a male. The amplitude of the spectrum is plotted in relative terms as variations in the gray level. This time history provides a time–frequency picture of the variations of the spectral resonances, called the formants. Thus, we can view changes in the bandwidths of the formants and changes in the formants themselves as the spoken words change. Present-day procedures for calculating the spectrogram use FFT techniques. MATLAB provides a function `specgram` that we use. For each data segment or frame, the amplitude of the spectrum is quantized and mapped to an assigned gray level.

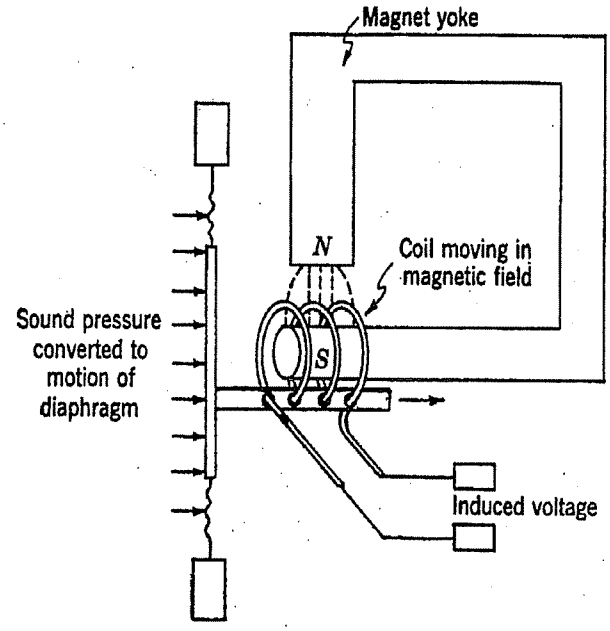

FIGURE 4.4 A moving coil microphone.

The data segment is updated and the process repeated. The spectral estimates are then plotted versus time, as shown in Figure 4.6.

The historical reason for this type of display is that some years ago, before digital techniques, the spectra and spectrogram were estimated using analog filters. For example, the spectrogram was estimated by repeatedly passing a tape recording of the signal through a bandpass filter, whose center frequency was shifted after each pass of the data. The tape recording of the data was made into a loop, which could be played over and over. The sweep of the bandpass filter was synchronized with the tape loop. The energy of the filter output for each pass of the tape loop controlled the heat of a stylus, which in turn burned a special paper. The more energy at the filter output, the darker the burn on the paper,

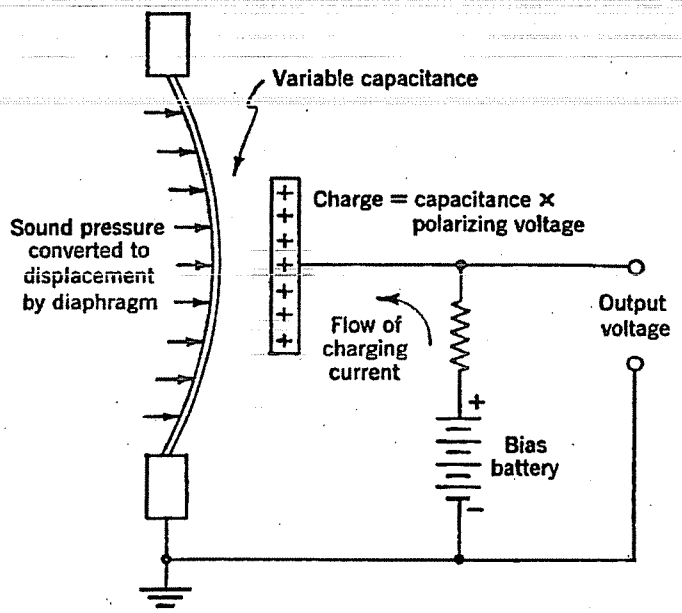

FIGURE 4.5 A condenser microphone.

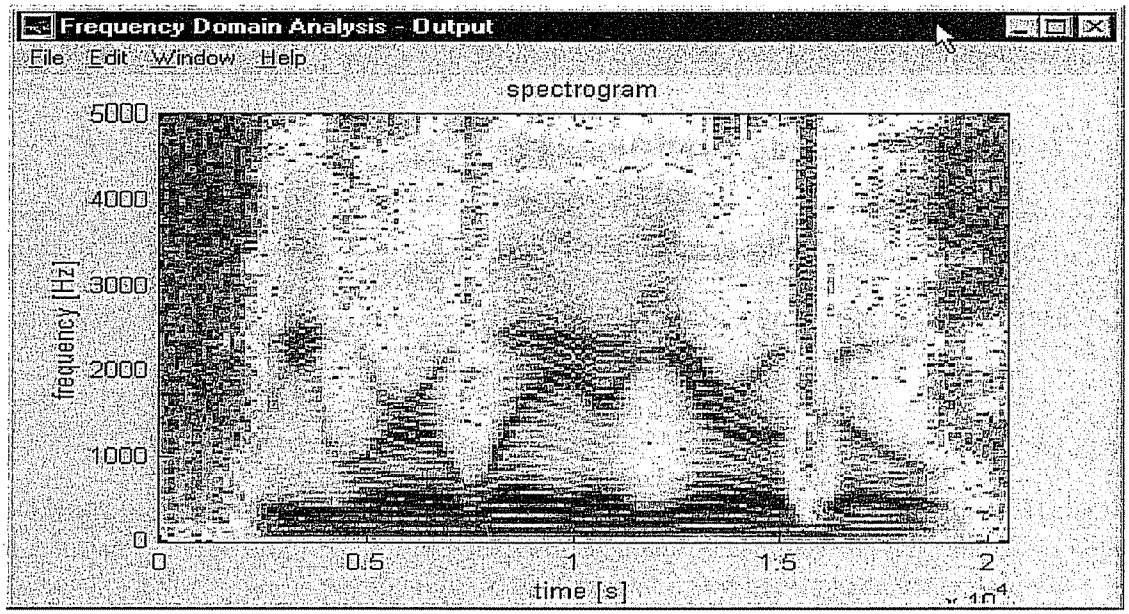

FIGURE 4.6 A spectrogram of the sentence, "We were away a year ago."

thereby creating a gray level display of the energy of the signal in the filter band at that particular center frequency. The filter center frequency then shifted up in frequency and the taped data started a new loop. Figure 4.7 illustrates this process. It took approximately 5 minutes or so to plot the spectrogram of a 1 or 2 second record of speech. But this display has proven to be useful to speech researchers and is the reason it is still used today with digital techniques. See Olive et al. (1993) and Potter et al. (1966) for spectrograms of speech data.

4.4.4 Perturbations: Jitter and Shimmer

Perturbations in the pitch period are called jitter, while similar perturbations in pitch amplitude are known as shimmer. Such perturbations occur naturally during continuous speech, but measurements from acoustic waveforms have demonstrated that the perturbations for pathologic and normal speakers differ. Speech synthesis has been used to verify that the

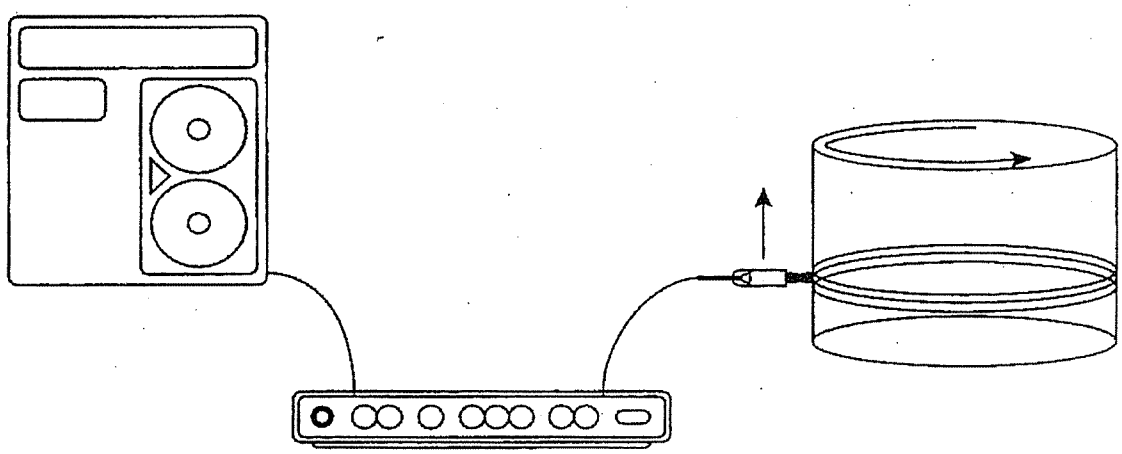

FIGURE 4.7 Illustration of making an analog spectrogram.

proper manipulation of jitter and shimmer does, in fact, produce speech that is indicative of either normal or pathologic voice quality.

4.4.5 Fundamental Frequency

The compliance, mass, length, and elasticity of the vocal folds affect the speaker's fundamental frequency. This is a parameter that can be measured quantitatively from the acoustic signal, the glottal volume–velocity, and the electroglottographic waveform. Deviations in the fundamental frequency may be indicative of a functional disorder or an existing pathology.

4.4.6 Intensity

This aural parameter, sometimes referred to as amplitude, can also be measured quantitatively or assessed by a trained listener. It, too, can serve as an indicator of a laryngeal disorder. A typical example is a weak voice, which may be due to pathology such as a paralysis of respiratory muscles causing insufficient subglottal pressure. A laryngeal paralysis can produce a similar effect.

Excessive intensity may occur simultaneously with an increase or a decrease in fundamental frequency, with the former causing a shrill voice, the latter leading to hoarseness. The clinician and speech pathologist learn to listen for deviations in both intensity and fundamental frequency as elementary signs of a laryngeal disorder.

4.4.7 Vocal Quality

This parameter is difficult to assess and is usually associated with nasality and hoarseness, but other descriptive terms such as breathy, husky, harsh, throaty, and strained are also used. No quantitative measure exists for these terms, and there may be considerable disagreement among researchers with respect to the relative importance of this variable. However, some attempts at quantification of quality have been made, for example, jitter and shimmer are now routinely used. The condition of the vocal folds as well as that of the vocal tract affects the quality of the voice.

4.4.8 Glottal Airflow and Subglottal Pressure

While many measurements in this area have been made, it is significant that the glottal volume–velocity has apparently never been measured directly. The average subglottal pressure has been measured directly with a pressure transducer introduced into the subglottal region through a tracheal puncture and can be measured by lowering a transducer through the glottis into the trachea.

4.5 MORE ON ELECTROGLOTTOGRAPHY

A mathematical model of the electroglottographic (EGG) waveform as a function of time is

$$\mathrm{EGG}(t) = \frac{k}{A(t) + C} \tag{4.5.1}$$

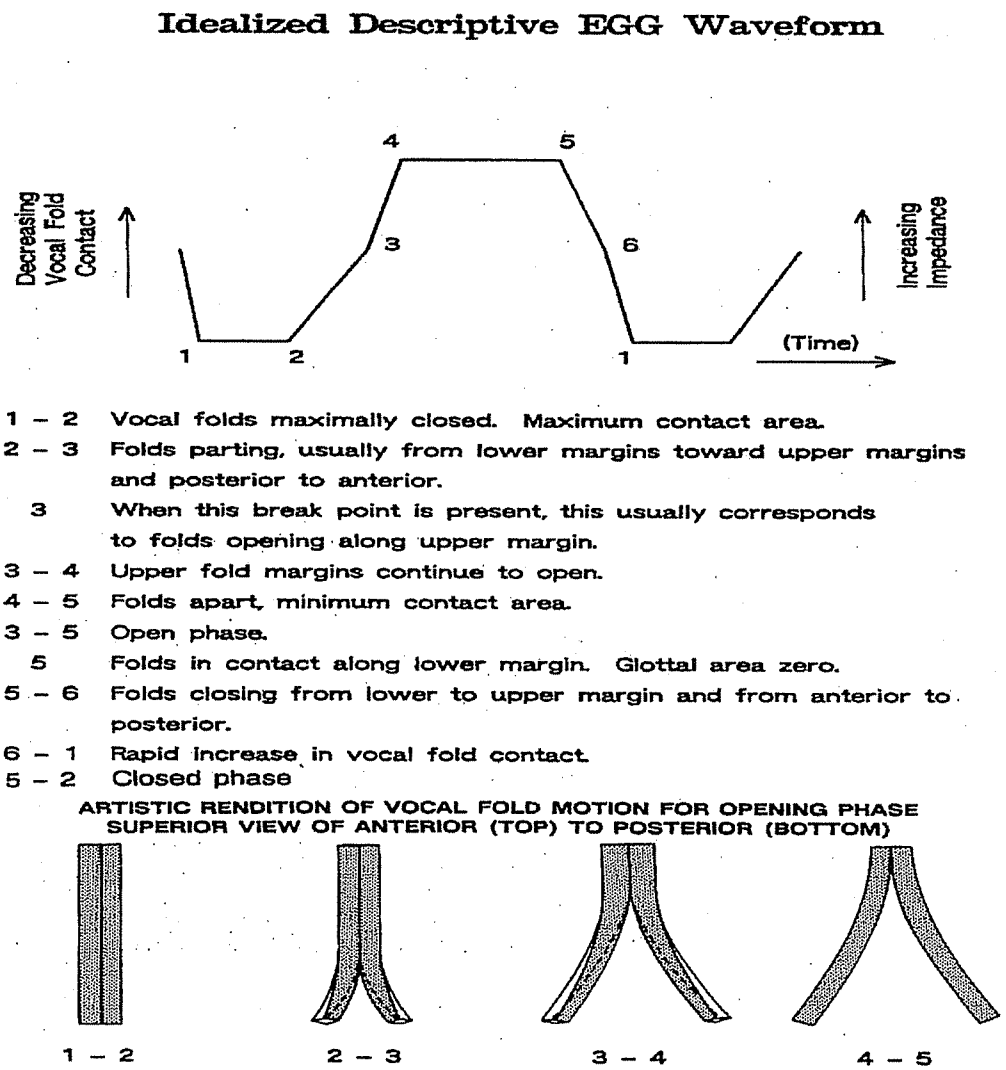

FIGURE 4.8 Idealized EGG waveform.

where A(t) is the vocal fold contact area, k is a scaling constant, and C is a constant proportional to the shunt impedance specified for the situation when $A(t) = 0$. This mathematical model, and its relationship to features of the EGG waveform depicted in Figure 4.8, has been partially verified by Childers and co-workers (1986; 1990). The vocal fold events are labeled on the EGG waveform and correspondingly on the artistic rendition of the vocal fold motion. The vocal folds are stylized and depict only the anterior one-third segment of the folds. The upper and lower vocal fold margins are out of phase. A vocal fold model has been developed using MATLAB and will be discussed in another chapter. A comparison of this EGG waveform model to actual data can be made via Figure 4.9 where, from top to bottom, the waveforms are speech, EGG, differentiated EGG, and glottal area. All waveforms are aligned. These data are for a male subject with a normal larynx with F_0 approximately 120 Hz at a low intensity.

Several vocal fold vibratory (glottal area) events are related to corresponding events (segments or peaks) of the EGG or the DEGG waveforms. For example, the instant of the opening of the glottis and the instant of the positive peak in the DEGG are coincident, as are the instant of the closure of the glottis and the instant of the negative peak in the DEGG. Similarly, the instant of the maximum glottal area and the instant of the maximum positive peak in the normalized EGG waveform are coincident. The open quotient (OQ) measured

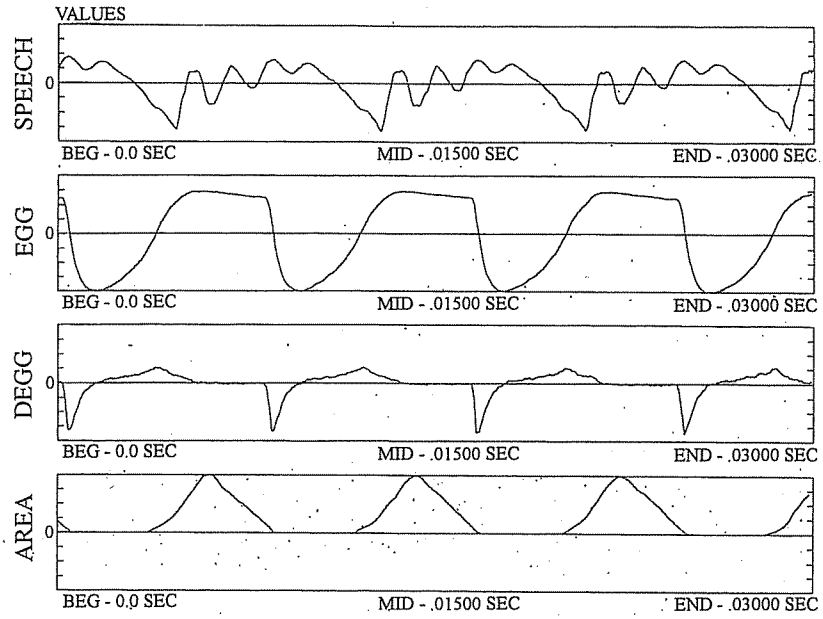

FIGURE 4.9 Data waveforms: speech, EGG, DEGG, and glottal area.

from the glottal area and the OQ measured from the EGG are nearly the same, where

$$\text{OQ} = \frac{\text{duration of glottal open phase}}{\text{duration of glottal cycle}} \tag{4.5.2}$$

Also the average perturbation measured from the glottal area and that measured from the EGG are nearly the same.

The results of such studies have shown that the EGG and DEGG waveforms are useful for extracting information about the vibratory motion of the vocal folds and that such measurements are noninvasive and easy to accomplish. A major use for such measurements is to assist in the validation of measurements made from the speech waveform only. In addition, the EGG waveform can be used as a second "channel" for processing the speech waveform for speech analysis and recognition tasks. One example of this appears in the paper in Appendix 6.

In order to facilitate the comparison of the EGG and DEGG waveforms with the speech waveform, we have developed a display program described next.

4.6 DATA-APPENDIX 4

The data described in Appendix 4 are contained on two CD-ROMs. One CD-ROM contains the Normal folders, while the other CD-ROM contains the Disorder folder, the Mimic folder, the Other folder, the Additional folder, and the Area folder. The files generally include a sychronized EGG file as well as a speech file. Be sure to read this appendix for additional details.

4.7 DISPLAY OF SPEECH, EGG, AND DEGG

The m-file called speech_egg_display.m in the display_speech_egg folder can be used to load, play, and compare speech, EGG, and DEGG ASCII data files. Change directory to

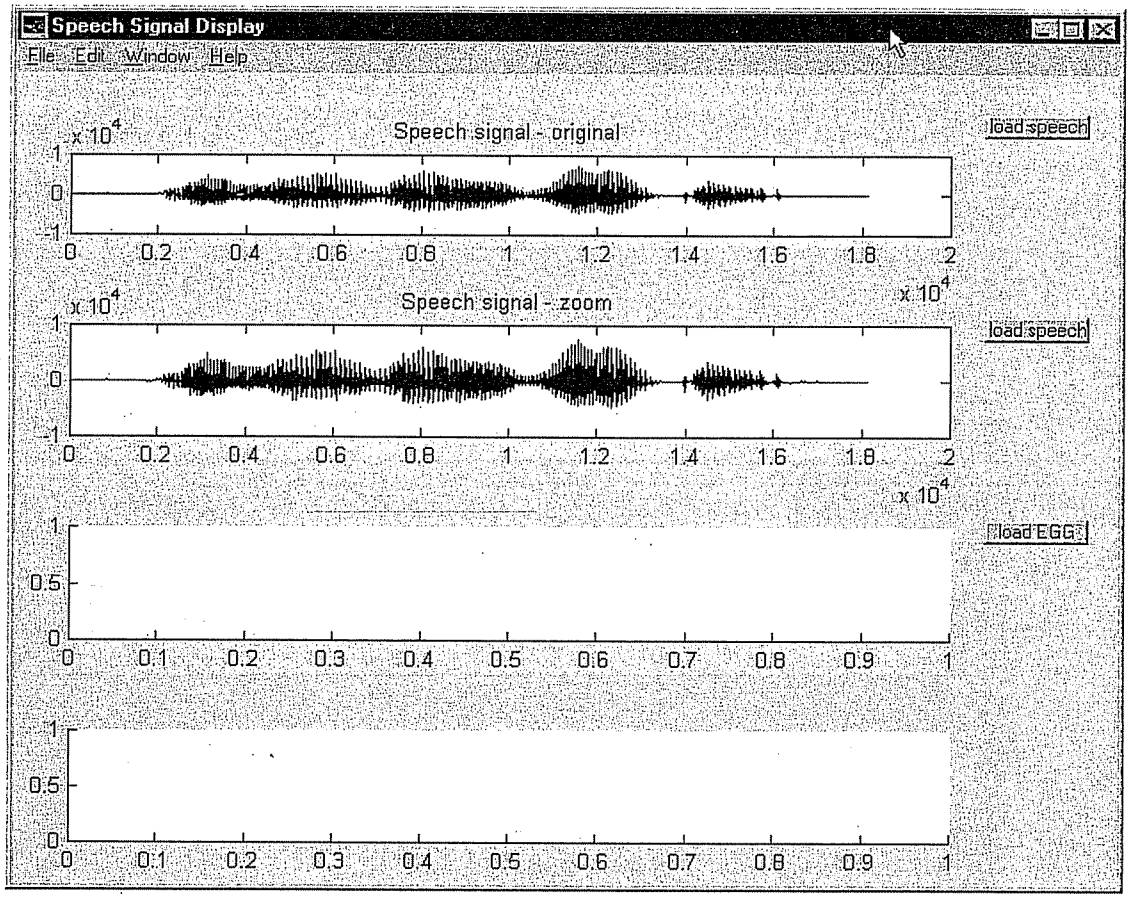

FIGURE 4.10 Display of speech and EGG data files.

the display_speech_egg folder and type speech_egg_display in the MATLAB Command window to initiate the program. The window will be similar to that shown in Figure 4.10, except that only the top load speech button will appear. No other buttons will be present. The user is to load a speech file, which will be plotted in the top panel. Repeat the load for the same speech file for the next panel. After completing the loading of the speech file twice, the results appear as shown in Figure 4.10. Next load the corresponding EGG file. There is a slight pause (30 to 60 sec) after loading the EGG file, since the program calculates the DEGG. Once this calculation is completed, the EGG and DEGG waveforms are plotted and the remaining buttons for the window are displayed. This is shown in Figure 4.11. At this point, the user can play the files. Also, the waveforms or spectrograms can be compared. The purpose for loading the speech file twice is that the top panel displays the speech file in its original form, while the lower three panels can display the data in a zoom-in manner.

In summary, to use the speech_egg_display.m file, follow these steps: first load the desired speech file in the top panel, then load the same speech file in the second panel from the top, next load the corresponding EGG file in the third panel. After the DEGG is calculated, the EGG and DEGG waveforms are plotted along with the remaining window buttons, as shown in Figure 4.11.

The user can select one of the various buttons, e.g., Play or Spectrogram, for each of the displayed signals. The Waveform button redisplays the signal if the spectrogram is plotted. Note that a zoom-in or-out button is available to zoom the waveform data display. Pressing this button changes the mouse cursor to a cross hair. Move the cross hair to the desired initial x-axis data location to zoom in. The cross hair can be place at any desired

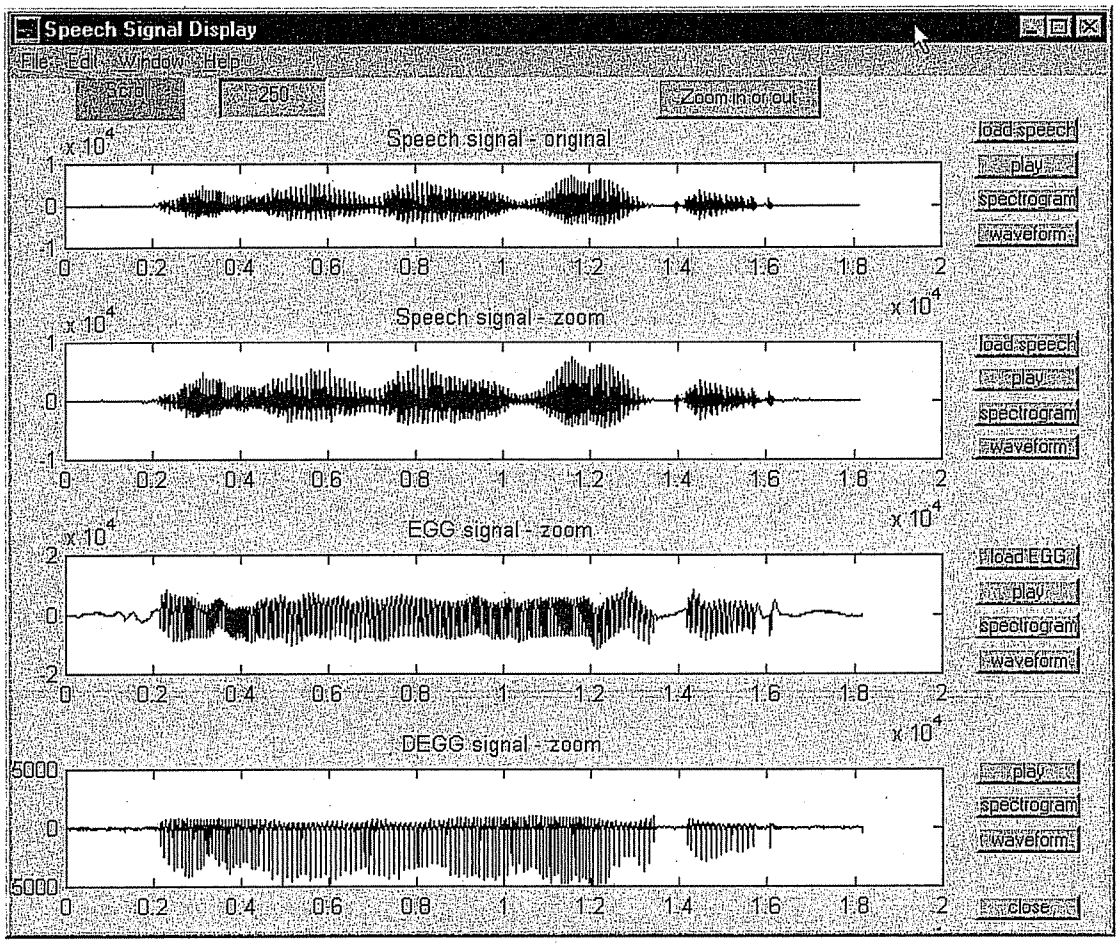

FIGURE 4.11 Display of speech, EGG, and calculated DEGG signals.

location on any of the waveforms. Press the left mouse button once. Move the cross hair to the desired end of the zoom in x-axis segment. Press the left mouse button once again. This zooms in one level on the signals in the three lower panels, as shown in Figure 4.12, which shows the zoomed-in waveforms plotted in each panel. Repeat these steps to zoom in another level. To stop the zoom-in process, press the right mouse button twice slowly while the cross hair is visible. The cross hair disappears and the standard mouse cursor reappears. To zoom out one level, press the zoom-in or-out button. The cross hair reappears. Press the left mouse button, followed by a press of the right mouse button. The data are zoomed out one level. Each repetition of this sequence zooms out another level, until the original signal level is reached. At any point in the zoom-out process, press the right mouse button twice slowly to exit the zoom-out option. Note that the scale of the zoomed-in data show the x-axis locations of the selected beginning and ending points of the desired zoomed in data segment. Thus, the user can refer to the top panel to keep track of the location of the zoomed-in data segment.

The scroll option is to be used after zooming in on the data at least one level. The scroll option becomes available automatically after zooming in. When the data are zoomed in, the scroll display is updated by adding the < and > scroll buttons to the figure, as shown in Figure 4.12. The default scroll value is 250 data points. This value can be changed by highlighting the number in the scroll window with the mouse cursor and typing in a new value. Press the left (right) mouse button on the < (>) button to scroll the data to the left

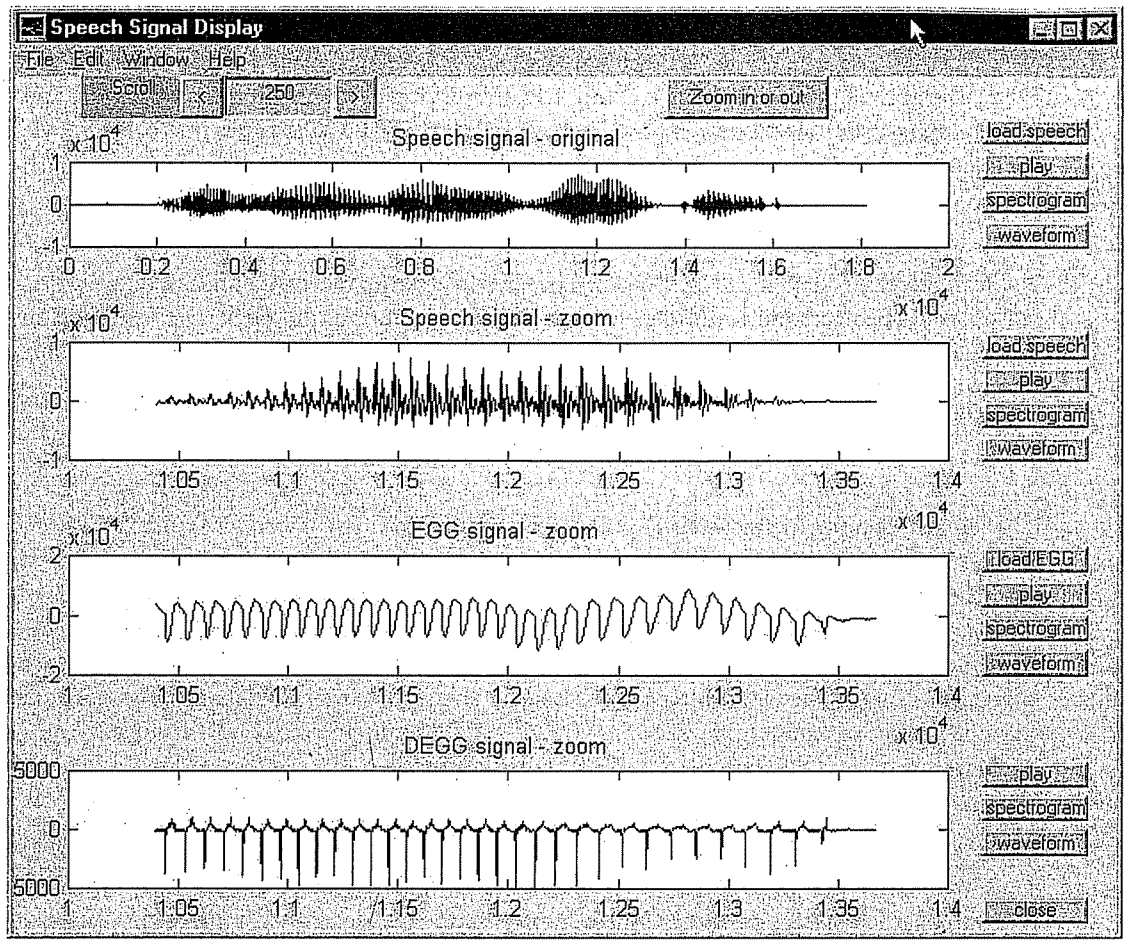

FIGURE 4.12 Display of zoomed-in speech, EGG, and DEGG signals.

(right). You can continue scrolling in either direction until the end of the data record. The scroll option can be used at any zoom-in level, as long as the zoom-in level is at least one.

The play option is designed to play the data displayed. Thus, if the data are zoomed in, the data shown in the panel are played. Similar remarks apply to the spectrogram option. The Close button closes out the Speech signal display window.

4.8 SUMMARY

The purpose of this chapter is to introduce the reader to many of the measurement procedures that are used in speech science. However, usually only the speech signal is used in most, if not all, speech-engineering applications, including speech recognition and speech synthesis. Nevertheless, some of the measurement techniques are useful to assist in the validation of algorithms that have been developed for speech only analysis. One such technique is the EGG. An extensive data set of synchronized EGG and speech waveforms is included with this text. Also, one algorithm that uses both the EGG and speech waveforms for silent and voiced/unvoiced/mixed (four-way) classification of speech is discussed in Appendix 6.

The following problems are a mixture of theoretical and experimental. Their purpose is to examine the theory behind some existing algorithms and to illustrate the development of additional algorithms for speech analysis. After this chapter, we move on to linear prediction and its use in speech analysis and synthesis.

PROBLEMS

4.1 The average magnitude difference function is useful for pitch estimation and offers less computational complexity than the autocorrelation function. Suppose that a function x(t) is periodic with period T. Then a function

$$d(n) = x(n) - x(n - k)$$

will be zero for $k = 0, \pm T, \pm 2T, \ldots$. For short segments of voiced speech one would expect that d(n) would be nearly zero, although not exactly so, at multiples of the period. The short time average magnitude of d(n) as a function of k will be small whenever k is close to the period. Thus, the short time average magnitude difference function (AMDF) is defined as

$$\text{AMDF}(k) = \sum_{m=-\infty}^{\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)|$$

When x(n) is close to periodic in the interval spanned by the window, w(n), then AMDF(k) will have sharp nulls for $k = T, 2T, \ldots$. It is not uncommon for the window to be the rectangular window. The AMDF requires fewer multiplications than the autocorrelation function. Consequently, it has been used in real-time speech processing systems. Implement this algorithm in MATLAB and test its use for pitch estimation for the sentence "We were away a year ago" file m0125s.dat. Compare this result with that obtained using the pitch algorithm available in the analysis software. Next, test the algorithm on the sentence "Should we chase those cowboys?" file m0127s.dat, and compare your results with that obtained using the pitch algorithm available in the analysis software. You may want to use the zoom option so that the results are more easily seen. Save your results for use in another problem below.

4.2 Clip the speech waveform using a center clipper, i.e., delete the speech waveform between $\pm C$. This can reduce the amount of the data that needs to be processed for the autocorrelation estimation of the pitch. The threshold in the Pitch/Jitter/Formant option can be used for this purpose; however, the waveforms are not displayed. Therefore, implement in MATLAB a center clipper for speech. Let C be a variable that the user can set. Then center clip the sentence "We were away a year ago" file m0125s.dat with $C = 70\%$ of the maximum of the data. Next calculate the short time autocorrelation function using a hamming window of length 256. Plot the original data, the center clipped data, and the short time autocorrelation function. Estimate the pitch contour from the short time autocorrelation function. Play the original data as well as the center clipped data. Discuss your impressions of this auditory evaluation. Repeat the problem with $C = 40\%$ and $C = 90\%$. Discuss your results. Save your results for use in another problem below.

4.3 The pitch algorithm in the Pitch/Jitter/Formant option uses median smoothing to eliminate sharp discontinuities in the contour due to algorithm errors. Design a median filter of length 3 and one of length 5. Apply each filter to the pitch contours estimated in Problems 4.1 and 4.2. MATLAB function, medfilt1, can be helpful for this problem as well as the median filter implementation in the pitch/jitter/formant m-files. Does a median filter give a more reasonable contour? Why?

4.4 The cepstrum can be used for pitch estimation as well as formant estimation. In this problem, estimate the pitch contour using the cepstrum (real or complex) for the sentence "We were away a year ago" (file m0125s.dat). Compare your result with the results from Problems 4.1 and 4.2 and the pitch/jitter/formant algorithm.

4.5 Implement in MATLAB a formant estimation algorithm using the Cepstrum. Analyze the sentence "We were away a year ago" (file m0125s.dat) using your implementation and compare your results with that using the pitch/jitter/formant algorithm.

4.6 Use the software described in this chapter to compare the speech, EGG, and DEGG for the sentences "We were away a year ago" and "Should we chase those cowboys?" (files m0125s.dat and m0125e.dat and m0127s.dat and m0127e.dat, respectively). For the first sentence note that the EGG and DEGG are not of much use for estimating the voiced/unvoiced segments. Why? However,

for the second sentence the EGG and DEGG are helpful. Design an algorithm for classifying the speech into voiced/unvoiced segments using both the speech and EGG (DEGG) waveforms. You do not have to implement the algorithm. Compare your algorithm with that in Appendix 6. What is a major problem with such algorithms?

4.7 If $\hat{x}(n)$ is the complex cepstrum of $x(n)$, then match the first column of possible properties of $\hat{x}(n)$ with the second column of properties to be considered for $x(n)$. For each property in the first column determine the corresponding property in the second column. In all cases, assume $x(n)$ is real. Any property in the second column can be used only once. (After a problem in Oppenheim and Schafer, 1975)

(a) $\hat{x}(n)$ real	1. $x(n) = -x(-n)$
(b) $\hat{x}(n) = -\hat{x}(-n)$	2. $x(n) = x(-n)$
(c) $\hat{x}(n) = 0,\ n < 0$	3. $x(n)$ real
	4. $x(n) = 0,\ n < 0$
	5. $\sum_{n=-\infty}^{\infty} x^2(n) = 1$
	6. $\sum_{n=-\infty}^{\infty} \lvert x(n)\rvert = \frac{1}{\sqrt{2\pi}}$

4.8 Let $x_1(n)$ and $x_2(n)$ be two sequences and $\hat{x}_1(n)$ and $\hat{x}_2(n)$ their complex cepstra. If $x_1(n) * x_2(n) = \delta(n)$, determine the relationship between $\hat{x}_1(n)$ and $\hat{x}_2(n)$.

4.9 Suppose we have two sequences $x_1(n)$ and $x_2(n)$, with transforms $X_1(z)$ and $X_2(z)$, respectively. Suppose $x_1(n)$ is minimum phase and $x_2(n)$ is maximum phase. If $|X_1(z)| = |X_2(z)|$ for z on the unit circle, determine the relationship between $x_1(n)$ and $x_2(n)$.

4.10 The purpose of this problem is to compare the speech, EGG, and DEGG waveforms and spectrograms for the same speaker, but for different vocal conditions, i.e., normal (modal) voice (file m0125s.dat), vocal fry (creaky voice) (file m0305ms.dat), breathy voice (file m0405ms.dat), hoarse voice (file m0505ms.dat), computer voice (file m0605ms.dat), and rough voice (file m0705ms.dat). The latter five files are in the Mimic folder along with the corresponding EGG files. Describe the major differences between the features of the waveforms and the spectrograms for these six different voice types. What is the major contributor to the differences between these voice types? Do the EGG and DEGG waveforms offer any features that might be helpful in voiced/unvoiced classification?

4.11 The purpose of this problem is to compare the jitter for the six voice types considered in Problem 4.10. The files to be compared are as follows: normal (modal) voice (file m0125s.dat), vocal fry (creaky voice) (file m0305ms.dat), breathy voice (file m0405ms.dat), hoarse voice (file m0505ms.dat), computer voice (file m0605ms.dat), and rough voice (file m0705ms.dat). The latter five files are in the Mimic folder along with their corresponding EGG files. Calculate and compare the Pitch/Jitter/Formant contours for these files. Do this for both the speech and EGG files. What are the jitter values for all of these voices? Compare the results for the speech analysis with that for the EGG analysis. Which voice type has the least jitter and which the largest jitter? An average jitter greater than 5% to 7% is often considered a deviant voice by clinicians. Would any of these voices be considered deviant? How do the results for the speech analysis compare with that for the EGG analysis?

4.12 The purpose of this problem is to illustrate the difficulty of identifying and labeling word boundaries for sentences with different phoneme content. The three files to be compared are m0125s.dat, m01zs.dat (in Other folder), spe9_1.dat (in Additional folder). The three sentences are "We were away a year ago" (all voiced), "That zany van is azure" (voiced fricatives), and "We saw the ten pink fish" (unvoiced plosives and fricatives). Use the Energy and ZCR option with a hamming window of length 256 and 10% overlap to make your first estimate of the word boundaries using the speech files only. Repeat this using only the EGG files. Next calculate the spectrogram for each file and label the boundaries by hand. Finally, use the Zoom in and Play options to locate the word boundaries. Compare and discuss your results. Could the EGG data be useful?

4.13 The purpose of this problem is to practice phoneme labeling for five sentences: "We were away a year ago" (all voiced) (m0125s.dat), "Early one morning a man and a woman ambled along a one mile lane" (many nasals) (m0126s,dat), "Should we chase those cowboys?" (fricatives) (m0127s.dat), "That zany van is azure" (voiced fricatives) (m01zs,dat; Other folder), and "We saw the ten pink fish" (unvoiced plosives and fricatives) (spe9_1.dat; Additional folder). Use plots of the speech and EGG waveforms and spectrograms to identify and label the phoneme boundaries. The Zoom option and the Energy and ZCR and the Pitch and Formant contours may also be helpful. Do not use the Play option. Do this problem without the help of your ears. Make a list of the major features that you found useful for identifying and labeling the phonemes. When you have finished, compare your results with the results in Appendix 5. Discuss the possibility of designing an automatic phoneme recognition system. Could the EGG data be useful?