CHAPTER 2: THE PRODUCTION AND DESCRIPTION OF SPEECH

2.1 SPEECH PRODUCTION

2.1.1 Introduction

This chapter provides a basis for the subsequent discussion of the speech related problems encountered as different stages of the work in this thesis. Firstly there is a general discussion of the origin and nature of the speech signal. There then follows articulatory, phonetic and mathematical descriptions of speech. Finally, voiced speech is discussed together with its relationship to the output from a laryngograph.

2.1.2 The speech signal

Speech provides human beings a means for the transmission of a complex message using sound. It is a signal that is very resistant to interference. Speech may still be intelligible even when the signal is distorted or heavily contaminated with interfering noise, although the quality of the speech will be reduced by such a process.

2.1.3 Origins of speech

The development of spoken language in humans was limited by constraints of evolution (Borden & Harris, 1980). Speech communication must be consistent with the available broadcast facilities (the speech centres in the brain and human vocal apparatus) and decoding system (the human auditory system). The organs of the body used for the production of speech, the vocal organs and respiratory apparatus, were originally evolved to permit breathing of air and the chewing and swallowing of food. However in the course of evolution, they have also been used to provide a means of communication using sound.

The use of speech as a means of communication is only possible because the code for the signal, that is the language system, is known to both speaker and listener. This

system determines the important sound contrasts and prosody.

2.1.4 The hierarchical nature of the speech signal

The hierarchical nature of the speech signal arises from its structured generation process (Borden & Harris, 1980). Some of the stages are illustrated in figure 2.1. Within the human brain, the speech centres contain information concerning the generation of speech. The phonological system used, the grammar and syntax of the language and the vocabulary are all implicitly represented. A possible description of the processes involved in speech production could be as follows: Let us suppose the top of this structure involves a cognitive level of representation where different system activity relates to different "ideas". The first step in speech generation involves a process which effectively arranges ones thoughts into the desired linguistic form and selects appropriate words and phrases to describe one's intended message. In addition these units must then be put into the correct temporal order as required by the grammar of the language. Then consideration to the different sound contrasts necessary for the given language and accent must be made. This could be thought of as corresponding to a phonemic level of processing. The message must next give rise to the signals necessary to control the muscles in the vocal apparatus. Finally, the physical behaviour of air in the vocal apparatus gives rise to an acoustic disturbance that radiates from the lips, and/or nose, carrying the message. The overall result of this coordinated activity is the radiation of sound from the speaker, a small part of which finally reaches the listener. Thus in the speech production process, there is a transformation between a linguistic to a physiological to an acoustic representation of the message. These successive layers form a hierarchically organised structure which can be used as a basis for similarly structured computer-based analysis of speech, as described in the next section.

Reception of the speech sounds in the listener results in processing with a reverse effect. There is a transformation from information in sound, to movement in the eardrum to nerve impulses in the auditory nerve and then finally activity in the higher centres in the brain.

2.1.5 Descriptions of the speech signal

There are several different ways in which one can describe speech. One may use the ideas of information theory and consider speech from the point of view of its information content (Shannon, 1968). Alternatively one may characterize speech as a signal which somehow carries the message information and look at properties of the acoustic speech waveform using parametric descriptions of the acoustic waveform (Rabiner & Schafer, 1978). In addition, one may adopt the approach of phoneticians and describe speech in terms of phonetic sound qualities which are related to the actions of the articulators in the vocal apparatus (Wells & Colson, 1971).

2.1 Descriptions of speech

2.2.1 Articulatory Levels of Description

One also can describe speech at the articulatory level, in terms of the behaviour of the anatomy of the vocal tract (Wells & Colson, 1971). The vocal apparatus, a cross-section through which is given in figure 2.2, provides a means by which nerve impulses from the brain may give rise to the acoustic speech signal. The final speech pressure waveform that is radiated at the lips and nose will depend upon the nature of the excitation and also the position of the articulators. Because the vocal tract transfer function and the excitation are both a function of time, the spectrum of speech is not stationary. By controlling the action of both the articulators and the vocal folds simultaneously, the brain may thus generate a signal in which the underlying message has been suitably coded for acoustic transmission.

The vocal apparatus is a complex sound generator. For voiced speech production, the larynx is the source of the sound and the vocal tract is a time-varying acoustic filter which modifies the laryngeal excitation depending on the position of the articulators. Voiced speech excitation is discussed in more detail in a later section. For voiceless excitation, the sound source is due to turbulent airflow at a point of construction in the vocal tract, and the location of this point is again dependent upon the position of the

articulators. Frication occurs only when the flow of air through constrictions in the vocal tract exceeds a certain critical value. Above this value, determined by the Reynolds number for air, the flow of air becomes turbulent. This turbulence gives rise to an acoustic disturbance that is noise-like in character. That is, un-correlated and with a flat spectrum.

The power needed to generate the sound largely comes from the breathing mechanism; the sources of air are often referred to as the air streams. The most common air stream due to exhaling from the lungs is known as pulmonic egressive. In addition there are oral and pharyngeal air-streams due to air movement caused by the action of the mouth and pharynx respectively. The respiratory system can be controlled by the brain so that breathing fits in to suit the speech. Mainly exhaled air is used for speaking, and expiration may last over 10 seconds in some cases.

2.2.2 The vocal tract

The vocal tract consists of two irregular tubes. There is a passage that connects the larynx to the pharynx, to the mouth and then to the outer air. In addition, when the soft-palate is lowered, there is another passage between the larynx to the nostrils to the outer air. The acoustic behaviour is the result of reflections and standing waves in these tubes and is dependent on the natural frequencies of vibration and damping within the system.

The dimensions of the vocal tract determine its resonant, or formant, frequencies. The relationship between these resonances is known as the formant structure. The vocal tract can be controlled by will to generate changes in this formant structure that are perceptibly different to a listener by the action of different articulators. Formant structure is important because it provides one means to distinguish sounds.

The articulators are the parts of the vocal tract that can be moved to alter the sounds that can be produced. The tongue can be moved up, down, backwards and forwards in order to change the effective length and cross-sectional area of the vocal tract. In

addition, the opening at the lips can be altered, the soft-palate can be opened and closed, and the jaw can be raised and lowered. The vowel systems in languages exploit all of these methods to change the formant structure.

The motion of the articulators is constrained by their anatomy and the muscles that move them. Consequently they can only move at a limited rate from one position to another. As a result of this the present location of the articulators will have some effect on their future position. These effects manifest themselves in the speech signal as assimilation effects.

2.3 PHONETIC LEVELS OF DESCRIPTION

A description of speech that is related to the articulatory descriptions is one based upon the phonetic qualities of speech (Wells & Colson, 1971; O'Connor, 1973; Ladeford, 1975). The field of phonetics is the study and description of speech sounds. It is concerned with what sounds we produce and how we produce them.

Phonetic descriptions are based on perceptible differences in the way the vocal tract of the speaker is used to produce speech sounds. Most languages, including English, can be described in terms of a set of distinctive sound units that are known as phonemes. A table of the phonemes of English, together with examples of them, is given in table 2.1.

A phonetician can write down a representation of speech sounds using a phonetic transcription, which consists of a set of symbols. At the segmental level these symbols indicate the place and manner of articulation as well as the presence or absence of voicing. The manner of articulation refers to the kind of articulation used, for example nasal, rolls, plosive, lateral, affricate. A description of the setting of the lips is also important and it is required to know their rounding, spreading and protrusion. Suprasegmental aspects of speech, such as the intonation of an utterance has a linguistic component that may be described in terms of a fall, rise, rise-fall, fall-rise, etc.

2.3.1 Phonemes

The important point about phonemes is that they are sound units that are contrastive with respect to one another and can be used to discriminate between words. A phonetician shows that two sounds (allophones) are phonemes by finding what is known as a minimal pair to demonstrate that a contrast exists between them. This is a pair of different words that are distinguished on the basis of the phoneme under investigation. The contrastiveness of a particular pair of sounds depends upon the given language and even the dialect. Consequently a given phonemic transcription system may not be suited for transcribing other languages. Phonemes can themselves be classified into vowels, diphthongs, semivowels and consonants.

2.3.2 Allophones

A phoneme has variants known as allophones. The allophones of a phoneme constitute a set of sounds that do not change the meaning of a word, are similar to each other and occur in phonetic contexts different from one another (Ladefoged, 1975).

The allophones belonging to a given phoneme may either be arranged into complementary distribution or in free variation. If two allophones are in complementary distribution, this refers to the fact that the particular allophone used is dependent on the context (that is, the neighbouring phonemes). If two allophones are in free variation, the particular allophone used is freely selected and not dependent on context. Sounds that are in complementary distribution or free variation are only said to represent the same phoneme if they are phonetically similar. That is, they must have most of their phonetic features in common and they must sound similar to native speakers of the language.

There are various effects that occur in continuous speech. Two of these are assimilation and elision. Assimilation is a phenomenon whereby a phoneme consonant changes so that is has, for example, the same place of articulation as the following consonant. This makes the production of the sounds easier, since it requires less articulator movement

than would otherwise be needed. Another related phenomenon is elision, whereby a phoneme in an utterance is missed out, again to facilitate speech production by simplifying the required articulations.

It is valuable to make some brief general statements concerning the acoustic properties of certain categories of speech sound, as an aid in understanding the problems involved in speech fundamental period estimation.

2.3.3 Consonants

Consonants constitute the sounds that are not vowels and are differentiated by place of articulation (bilabial, labiodental, alveolar, dental, velar, paleto-alveolar, post-alveolar) their manner (plosive, fricative, affricate, nasal, continuant) and whether or not they are voiced. The differentiation between vowels and consonants must be made in terms of the relationship of the sounds in a language system and cannot be done solely on the basis of acoustic characteristics.

Plosives are transient non-continuant sounds and are characterised by three distinct phases. Firstly there is an approach phase, during which the appropriate articulators move towards their target positions. Secondly there is a hold phase, where the vocal tract is blocked off by closure of the articulators. Finally there is the release phase, when the articulators separate again. After the plosive release there may be a voiceless excitation due to the release of breath, and this is known as aspiration. Therefore plosives give rise to a brief transient burst of noise, as released air flows through the constriction. Thus a plosive is characterised by a short silence typically followed by a short noise burst when the stop is released. The length of the silence depends on the tempo of the utterance. It is shorter in voiced sounds than unvoiced sounds. However, the main difference between voiced and unvoiced plosives is that in the former the vocal folds vibrate during the closure as the pressure builds up, whereas in the latter case they do not. Often a small amount of low frequency energy can still radiate through the walls of the throat during the closure in a voiced plosive.

In an affricate, there is a plosive followed by a homorganic fricative. The latter is a fricative with friction occurring at the point of release of the plosive.

Nasal consonants involve the lowering of the soft-palate and a complete closure in the oral cavity so that air can only escape via the naso-pharynx. When the nasal passage is open, the closed oral cavity serves as a resonant cavity that traps acoustic energy at its natural resonant frequencies. The effect of this is to add an anti-resonance to the transfer function of the vocal tract, and results in the removal of energy from the radiated speech at the frequency of this anti-resonance (Flanagan, 1972). Since the oral opening of the vocal tract is closed off during a nasal, nasals consequently are of lower intensity than oral consonants. Different nasal consonants are differentiated by the place at which the obstruction of the oral tract takes place.

Fricatives are consonants in which there is turbulent air flow at a narrow region in the vocal tract, giving rise to noise-like acoustic excitation at the point of the narrowing. The location of the point of the narrowing determines which fricative is produced. This noise source is filtered by the action of the resonance of the oral cavity forward of the constriction and the anti-resonance of the oral cavity behind the constriction. Due to their noise-like excitation, fricatives are characterized as having non-periodic waveforms with significant energy at high frequencies (that is above a few kHz, which is not the case for vowels). In voiced fricatives, the point of constriction in the vocal tract is the same as for their unvoiced phoneme counterparts. However, there is also voiced excitation due to vocal fold vibration.

2.3.4 Vowels

Vowels are voiced sounds that are characterized by a lack of constriction of the vocal tract (it should be noted that whispered speech can be still treated as voiced phonemically, even though there is no vocal fold vibration but turbulence at the glottis instead). It is essentially the cross-sectional area of the vocal tract that determines its resonant frequencies and consequently the vowel quality that is produced. The dependence of the cross-sectional area of the vocal tract on the location in the vocal

tract is known as the area-function of the vocal tract. For vowel sounds there are no obstructions of the vocal tract, although the area-function depends mainly on the position and attitude of the tongue, and also to a lesser extent on the position of the jaw, soft-palate and the rounding of the lips. The vertical position of the tongue is often described in terms of height of the tongue, where a CLOSE tongue position represents the highest the tongue can be raised, whereas a OPEN tongue position is the furthest down it can be placed. The horizontal position of the tongue is described as FRONT, CENTRE or BACK, depending upon whether the tongue is forward in the mouth, midway or back in the mouth.

From the production point of view, vowels are more difficult to describe than consonants because the shape of the vocal tract cannot be as easily identified.

The auditory quality of a vowel is usually described by ear with respect to a reference set of vowels, known as the cardinal vowels. The quality of these vowels is independent of language and the cardinal vowel system provides a classification scheme on the basis of perceptible difference between a given vowel and the reference set. The cardinal vowels consist of a set of vowels that provide a coverage of all the possible vowels that can be produced. Thus they constitute a sampling of vowel space along the dimensions of open to close and front to back. In addition to tongue position, vowels may have different amounts of lip rounding.

In the case of diphthongs, the vocal tract area function changes smoothly between those of the appropriate two vowels. In all other respects, a diphthong has the features of an ordinary vowel.

Semivowels are a group of phonemes that are difficult to characterize. Their acoustic properties are similar to vowels and they are generally characterized by a gliding transition of their area-function between those of the adjacent phonemes. Consequently they are strongly influenced by their context. The distinction between semivowels and vowels is made linguistically with reference to their behaviour in a syllable, and not only on acoustic grounds.

2.3.5 Intonation

The most important function of speech fundamental frequency is as the carrier of intonation. Intonation is the temporal pattern of perceived pitch and it has two different purposes. It can convey grammatical information that forms part of a language system. As such, it is mainly the relative change in intonation that is important. For example, it can be used as a means of encoding stress into an utterance, which provides a means of emphasizing certain words. In addition, it can also convey information relating to the emotional state of the speaker. The fundamental frequency contour is important for the intelligibility and naturalness of the utterance (O'Connor & Arnold, 1961). In tone languages (such as Chinese) fundamental frequency changes produce lexical meaning contrasts.

2.5 DIGITAL REPRESENTATIONS OF THE SPEECH WAVEFORM

Speech propagates through the air as an acoustic pressure waveform. For the purposes of computer speech analysis, it is necessary first to convert it in a different form and this usually takes the shape of amplitude measurements of the speech pressure at regular time intervals (Rabiner & Schafer, 1978). The conversion of the acoustic speech waveform into a digitized speech pressure waveform involves firstly converting acoustic pressure variations in the air to electrical fluctuations using a pressure microphone (It is also possible to use a velocity microphone which responds to the velocity of the air rather than the pressure, but this type of microphone is less common). The output from the microphone is then low-pass filtered and then sampled at a uniform rate by means of an analogue-to-digital (A/D) converter, which converts the amplitude measurements to a number. It is necessary to ensure that the bandwidth of the signal to be sampled is less than half the sampling frequency, otherwise aliasing will occur and this is prevented by the low-pass filter (Nyquist, 1928). If the sampled data is aliased, then it will not be possible to reconstruct the original waveform from it, because it no longer uniquely represents the original waveform. It is also important that the resolution of the A/D converter is sufficient for the application, because the process of quantization of the continuously valued input signal into a set of discrete levels introduces uncertainty

in the signal representation that can be considered as additive noise (Rabiner & Schafer, 1978).

A description of speech in terms of the sampled representation of the speech pressure waveform is a very general representation that is only concerned with preserving the wave-shape of the signal by the appropriate choice of sampling frequency and levels of quantization. Such a description involves no other a priori knowledge particular to the characteristics of speech.

2.5.1 Parametric models

Parametric models of speech are more abstract that this and are concerned with representing the signal in terms of the output from a production model (Fant, 1970; Flanagan, 1972). In a simplest case of such a model, speech production is represented as an excitation source driving a time-varying linear filter that represents the acoustic effects of the excitation spectrum, vocal tract, and radiation effects at the lips. For voiced speech, the excitation source in this model must mimic the excitation due to the repeated opening and closure of the vocal folds. For voiceless excitation, it must mimic the noise-like excitation due to turbulent airflow in the vocal tract. In more sophisticated models, the effects of the excitation spectrum, vocal tract and lip radiation can be represented separately. In both cases, the time-varying linear filter must account for the resonances of the vocal tract, which are known as the formants. For simple purposes the vocal tract can be approximately modelled as two tubes. This production model is useful for the generation of synthetic speech as well as a model for speech analysis. For synthesis of voiced speech it is the first three resonances that are most important (Holmes, 1988).

2.5.2 Acoustic variability of speech

Different speakers will have different larynx sizes, vocal tract sizes, phonetic and linguistic upbringing, speech habits, emotional states and vocal fold characteristics. All these factors affect the speech produced in different ways. Consequently there will be

a large difference in the acoustic realizations of utterances for different speakers (cross-speaker variabilities). In addition, variabilities also arise because of differences that occur in a given speaker as a function of time (occasion-to-occasion variability). An example of speech variability in short-term acoustic representations is demonstrated by the fact that the first two formants for different speakers for the same vowels overlap, as shown in figure 2.3 (Peterson & Barney, 1952).

2.2 VOICED EXCITATION

There now follows a more in-depth description of voiced speech excitation, because this area is of particular interest to speech fundamental period estimation.

The basic acoustic function of the larynx is to act as the sound source during voiced speech production. A cross-section through the larynx is shown in figure 2.4, and front and back views are shown in figure 2.5. Its action gives rise to a glottal wave which acts as a carrier for the speech message imparted by the effects of the vocal tract. In addition, the characteristics of the voice source are important because it contributes to the means by which the physical, psychological and social characteristics of the speaker can be conveyed.

2.2.1 Vocal Fold Vibration

Voiced excitation occurs when air flows between the vocal folds causing them to vibrate and the main peak of excitation results from their closure. The result of vocal fold vibration is thus a modulation of the air flow that passes into the vocal tract and constitutes a quasi-periodic acoustic excitation.

The vibration of the vocal folds that characterises voiced speech is complex. The vibrating system is three dimensional, and consequently its motion is more complicated than simple harmonic motion. It is a vibrating system that has different modes of oscillation. In normal voice, the vocal folds constitute a thick shelf across the larynx (figure 2.4) all of which moves periodically together and then apart again. In other

modes of vibration, the vocal folds can be thinned out at the edges. This results in a lighter vibrating section and consequently a higher frequency of vibration.

2.2.2 Mechanism of vocal fold vibration

The mechanisms involved in vocal fold vibration can be understood by considering the following sequence of events, which follows what is known as the myo-elastic theory of phonation (Van den Berg, 1957). Air from the lungs during exhalation is the main airstream used in phonation (known as the pulmonic airstream). The laryngeal muscles can cause the vocal folds to close, thus blocking the air passage. If this happens during exhalation there will be a build up of air pressure below the vocal folds, which will eventually force them apart. After this happens, there are two mechanisms involved in bringing them back together again. Firstly the muscle fibers and ligaments in the vocal folds are elastic, and after the vocal folds have been forced out of position, they spring back to their resting position. Secondly, as air flows through the constriction in the vocal folds, its velocity increases and consequently its pressure decreases, due to the Bernoulli effect. When the air pressure between the vocal folds drops, the external pressure tends to force the vocal folds together. There is positive feedback in this mechanism, because the closer the vocal folds get, the faster the air flow and the greater the pressure drop will be. Therefore, the vocal folds are accelerated together, resulting in a strong impulse excitation of the vocal tract as they snap shut. After this, the pressure then rapidly returns to normal atmospheric, and because of the constriction the sub-glottal pressure starts to rise again. Thus the cycle repeats itself. The overall effect is that successive puffs of air enter the vocal tract just above the larynx.

The frequency of vibration of the vocal folds depends upon the sub-glottal pressure and their resistance to movement. The resistance to movement of the vocal folds depends on their mass, length and tension. The effective length of the vocal folds can be adjusted by means of the thyro-arytenoid muscles and crico-thyroid muscles (see figure 2.5). The latter changes the angle between the thyroid and cricoid cartilages thus stretching and lengthening the vocal folds. Since all of the parameters affecting vocal fold vibration rate are controlled by the action of muscles in the larynx and air pressure

and flow, the speaker is able to alter the vibration rate at will.

2.2.3 Laryngographic descriptions of Voiced speech

A device of particular value in the analysis of voiced speech excitation is the laryngograph (Fourcin & Abberton, 1971). A description of the laryngograph and its relationship to vocal fold vibration is of particular importance here because it forms a fundamental part in the training and testing of the fundamental period estimation algorithm which is the subject of this thesis.

A laryngograph operates by measuring the conductance across the larynx at the level of the vocal folds. This is achieved by placing two electrodes across the larynx with a small alternating voltage at several MHz across them. Movement of the vocal folds causes a change in the conductance which is subsequently detected.

The output waveform from the laryngograph thus gives a measure of vocal fold activity and is temporally much simpler than the corresponding speech pressure waveform. The point of closure of the vocal folds, which gives rise to the main peak in excitation, can be easily determined from the laryngograph waveform. The manifestation of the closure of the vocal folds in the laryngograph output signal is well agreed upon (Fourcin, 1974). The point of closure is usually taken as the point of maximum gradient in the closing phase of the laryngograph signal. Agreement on the opening point is, however, less well accepted. This is because as the vocal folds open, they "peel apart" from below and the corresponding effect in the laryngograph waveform is difficult to define as a specific distinct event. Figure 2.6 shows the relationship between vocal fold vibration and the laryngograph waveform for normal modes of laryngeal activity.

2.2.4 Laryngograph signals for different voice qualities

There now follows a description of the characteristics of the laryngograph waveform for different voice qualities. According to Hollein (1972) there are three major vocal registers; modal (normal), falsetto, and vocal fry (creak).

2.2.5 Normal voice

Normal voice is characterised by regular vibration of the vocal folds, without any frication. It is used over most of the speaker's frequency range. This is typically about 90-200Hz for male and 150-310Hz for women.

With normal voice the whole body of the vocal folds vibrates, giving characteristically relatively long vocal fold closure times. The brief velocity peak of the vocal folds that occurs as they snap shut gives an excitation with significant high frequency components, which results in a well defined set of formant frequencies. The speech pressure waveform for normal voice and the corresponding output from a laryngograph are shown in figure 2.7.

2.2.6 Breathy voice

Breathy voice may be characterised by incomplete closure of the vocal folds, and by greater pulmonic airflow than in normal speech. The vocal folds vibrate but do not necessarily make contact, although lack of contact only happens during very breathy voice. The closure points as observed by means of a laryngograph are smoother, because full closure is not made. Also, the open phase is much longer than normal. This results in greater sub-glottal damping of the vocal tract, and the vocal tract resonances are therefore less well defined than with normal speech. There is also noise generated by turbulence at the glottis, which shows up in the speech pressure waveform. A more extreme case of this aspiration occurs in the case of whispered speech, when there is strong air turbulence at the glottis and the vocal folds do not meet. The speech pressure waveform for breathy voice and its corresponding output from the laryngograph is shown in figure 2.8.

2.2.7 Creaky voice

A special case of vocal fold vibration is that of creaky voice. It generally occurs at the end of utterances with falling intonation and it is characterised by laryngeal vibrations

of unusually large duration. Sometimes these are alternated with shorter duration cycles, giving a short cycle followed by a long cycle. The irregularity is perceived as a creaky voice quality. The speech pressure waveform shows clear evidence of vocal tract excitation at each closure, and since the cycle time is large, each excitation of the vocal tract has time to die down a long way before the next excitation occurs, and consequently the excitation points are well defined. There is a tendency for speakers to use creaky voice quality if they want to go down to a low pitch that is below the bottom end of their normal frequency range. The speech pressure waveform for one example of creaky voice and the corresponding output from the laryngograph is shown in figure 2.9.

2.2.8 Falsetto voice

Falsetto voice occurs when only the top edge of the vocal folds vibrates which results in damping of the vocal tract by the sub-glottal system much sooner after the excitation point than in the case of normal voice. This results in a much temporally simpler speech pressure waveform than with normal voice quality. There is a tendency for the speaker to make use of a falsetto voice to reach fundamental frequencies that are above their normal range. The speech pressure waveform for an example of falsetto voice and its corresponding output from the laryngograph is shown in figure 2.10.

2.2.9 Mixed excitation

In some cases both fricative excitation and voicing occur at the same time. This is known as mixed excitation. Because of the pulsatile nature of the air flow via the vocal tract in this condition, the frication occurs in bursts synchronously with the glottal air flow pulses. Figure 2.11 gives an example of mixed excitation in a voiced fricative.

2.2.10 Problems in using the laryngograph

There are several limitation in using electro-glottography in general to estimate the operation of the vocal folds (Colton & Conture, 1990). These range from problems in

obtaining good quality laryngograph signals with some speakers to cases where there are discrepancies between the speech and laryngograph signals.

Only a small fraction of the current from the laryngograph electrodes passes through the vocal folds. As a consequence of this, the laryngograph waveform (known as Lx) is strongly affected by gross larynx movements, blood flow through the neck and the contraction of the extrinsic laryngeal muscles. Figure 2.12 shows a large excursion in the laryngograph waveform that often occurs as a speaker prepare to phonate that has no corresponding acoustic excitation. By high pass filtering this composite signal within the laryngograph, the faster fluctuation due to vocal fold vibration can be emphasized (Colton & Conture, 1990).

2.2.11 Discrepancies between the speech signal and the laryngograph signal

There are circumstances where the laryngograph does not always give a strong indication of voicing when observation of the speech pressure waveform indicated that voicing is indeed present (Howard & Lindsay, 1988). This happens when the vocal folds vibrate without making firm contact and are "flapping about in the breeze" (Childers & Larar, 1984). This mainly occurs towards the end of unstressed voiced segments, when the vocal folds are still vibrating but no firm closure is made. Consequently there is little change in the impedance across the larynx and therefore little fluctuation on the laryngograph waveform. This phenomenon occurs more frequently in the case of female speakers than for male speakers. Figure 2.13 shows the case when there is evidence of vocal excitation in the speech pressure waveform, but little evidence for it in the laryngograph waveform. Conversely, there are occasions when there is laryngograph activity, but no speech pressure waveform, such as during a hold in a plosive. In this case the acoustic excitation occurring at the vocal folds is attenuated by the closure, and consequently there is little or no speech output. Figure 2.14 illustrates this phenomenon.

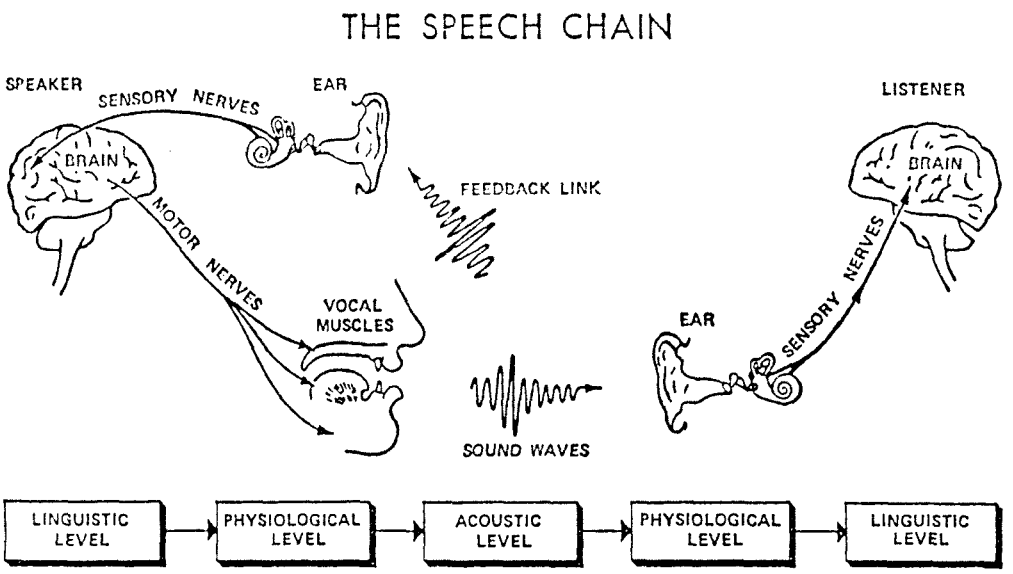

Figure 2.1 The speech chain.

This shows the stages in the generation of a message within the brain of a speaker to its transmission using sound, and then its reception in the brain of a listener (visual information, such as speaker gestures and lip moments, can also contribute to the communication process, but is not shown here). The message is shown to start as activity corresponding to a linguistic level within higher centres in the speaker's brain. Suitable nerve signals are then generated to control the vocal apparatus. This results in the broadcast of an acoustic speech wave which travels to the listener. The sound is then analysed by the ear (more particularly the cochlea) and nerve signals then convey the information to higher centres in the listener's brain, where their linguistic significance is interpreted.

(Taken from Denes & Pinson, 1973).

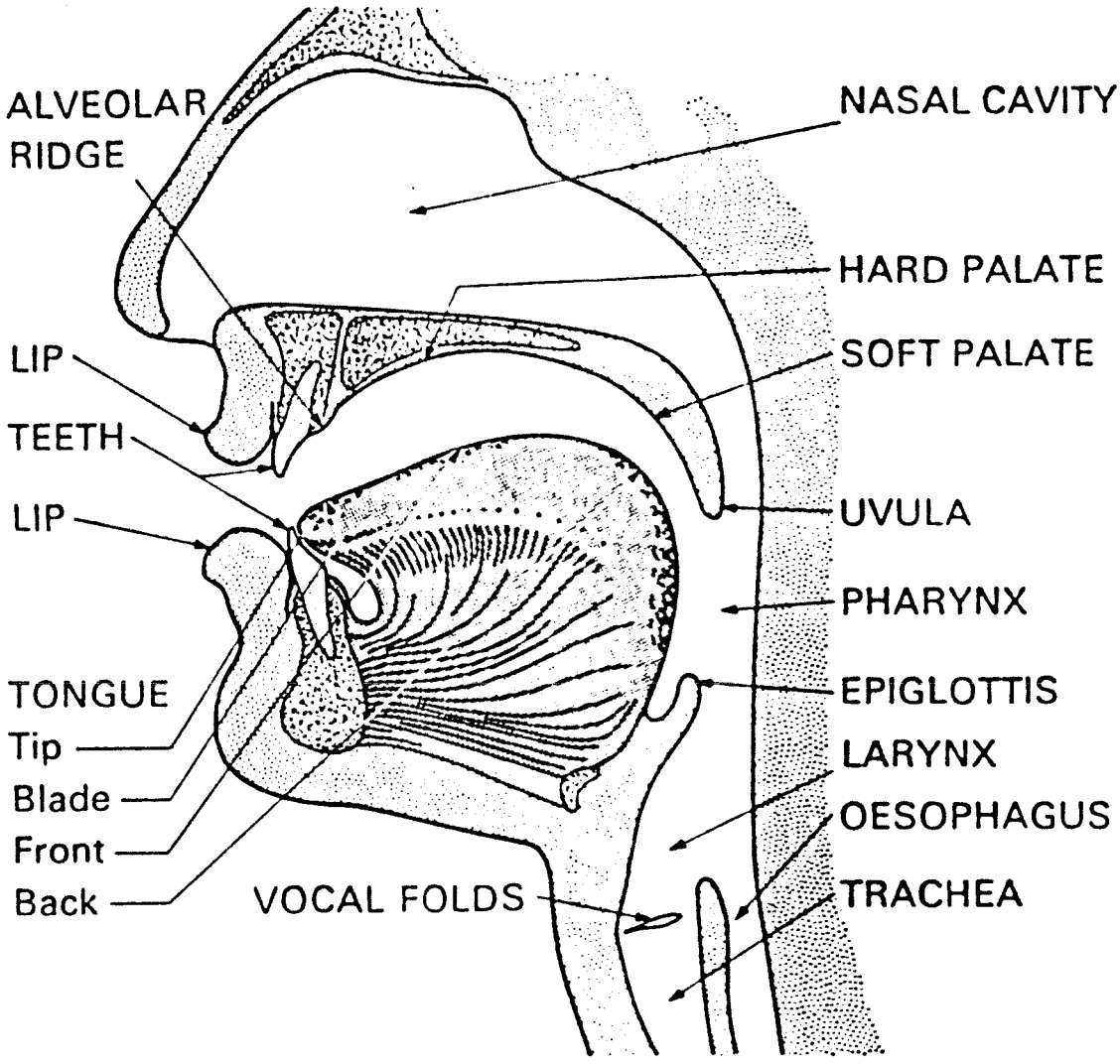

Figure 2.2 Cross-section through the human vocal tract.
The position of the articulators is shown.
(Taken from Wells & Colson, 1971).

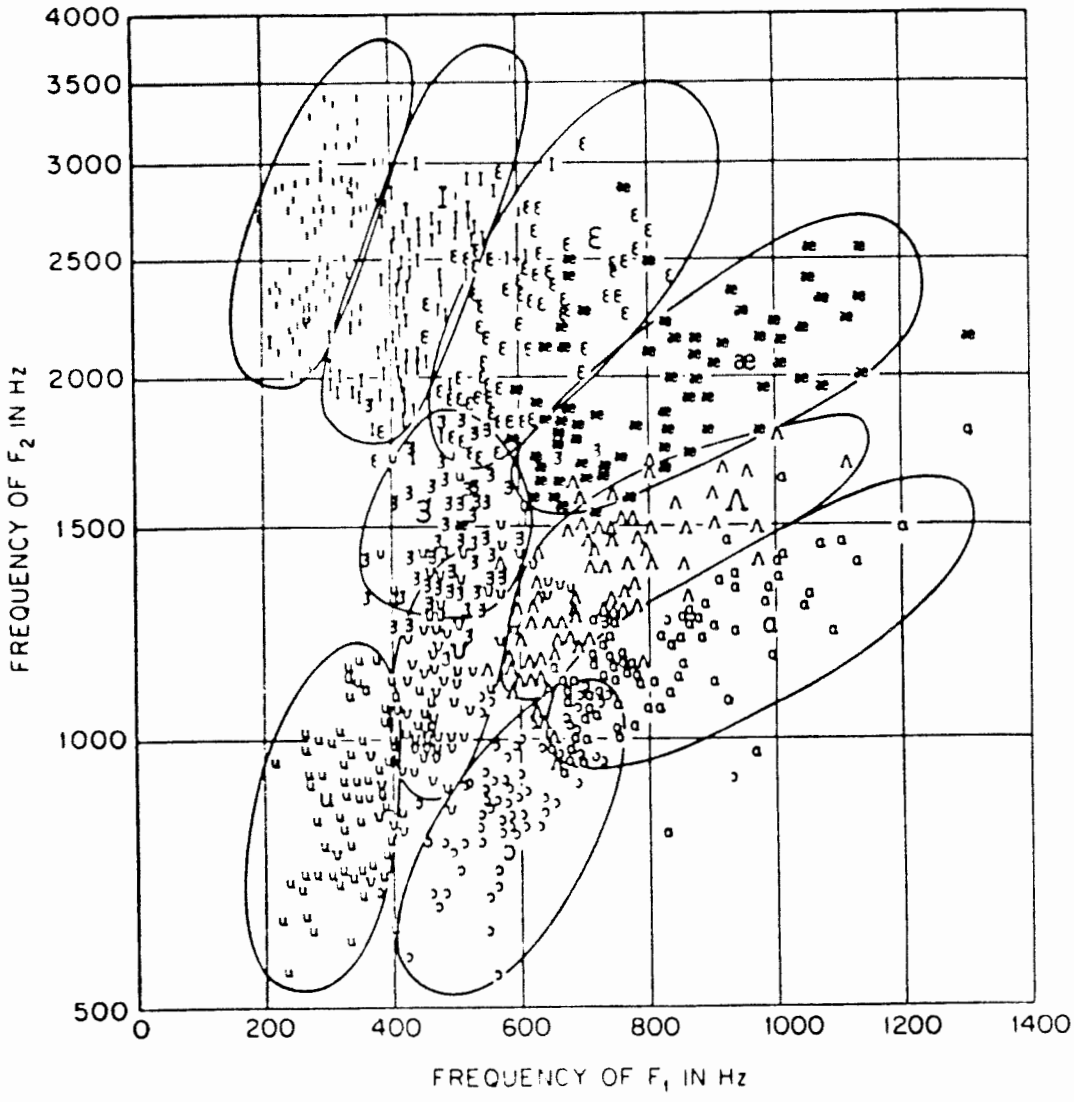

Figure 2.3 Variability of formant frequencies across speakers.

Figure shows the overlap between the first two formant frequencies of different vowels for different speakers.

(Taken from Peterson & Barney, 1952).

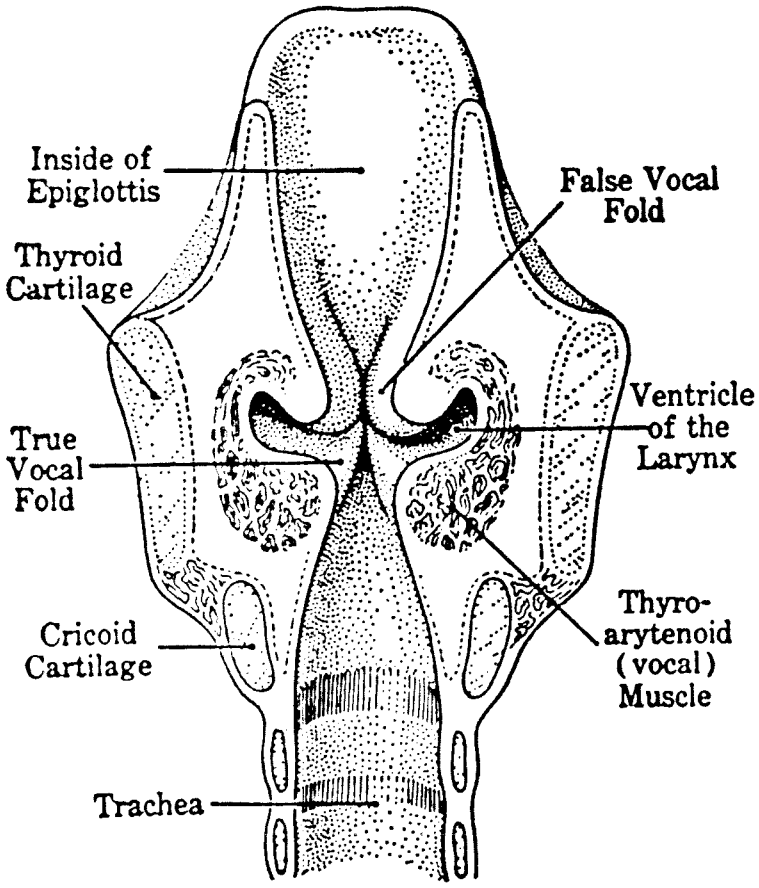

Figure 2.4 Cross-section through the larynx.
The vocal folds can be clearly seen.
(Taken from Borden & Harris, 1980).

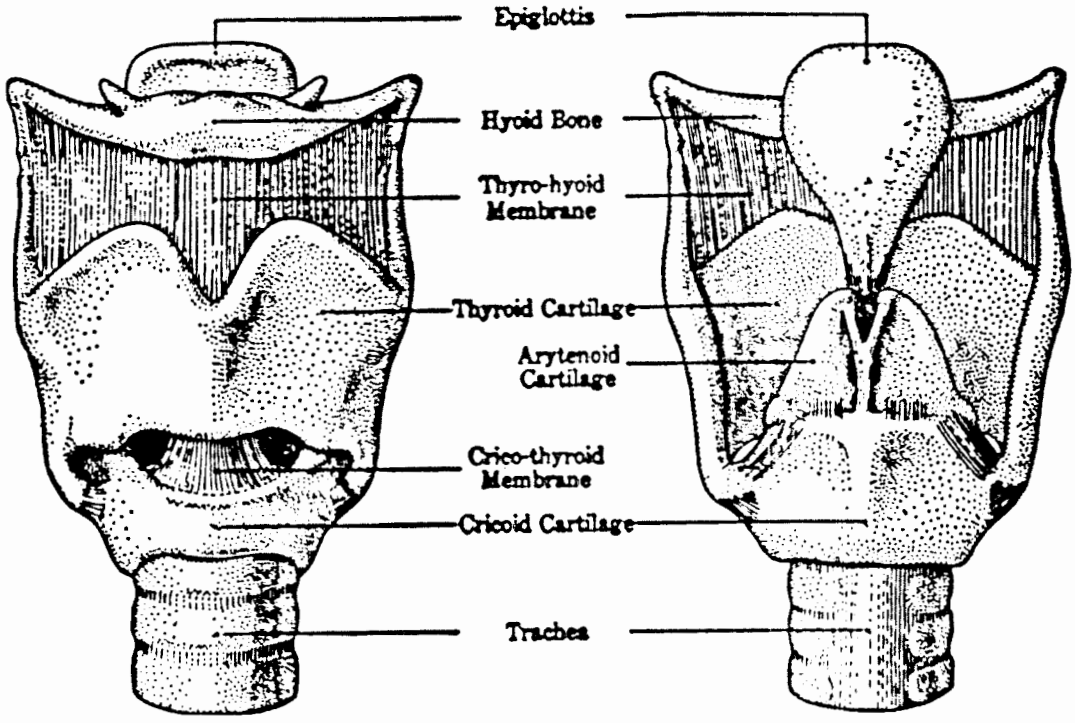

Figure 2.5 Front and rear views of the larynx.
(Taken from Borden & Harris, 1980).

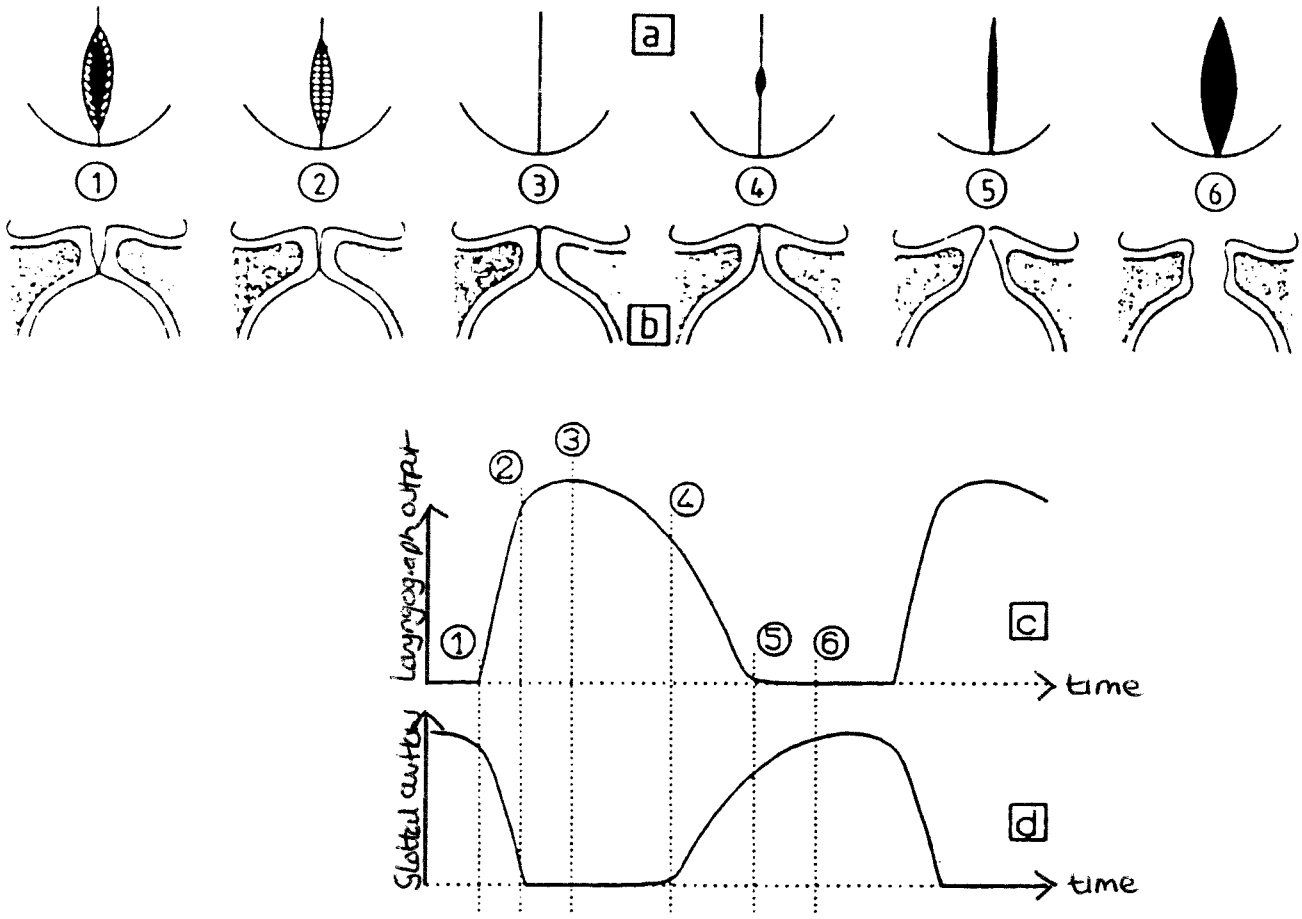

Figure 2.6 The relationship between vocal fold motion and the laryngograph waveform, during normal speech.

Six key stages in a complete period are shown. Diagrams (a) shows the view of the vocal folds from above. Diagrams (b) show a cross-section of the vocal folds. The corresponding effect in the laryngograph waveform is shown in diagrams (c). Diagram (d) shows the corresponding glottal air flow. The marked points are as follows:

(1) is the point of closure at a single point.

(2) is the instant when complete closure has been made over the length of the glottis, but not over the vertical plane.

(3) is the point of maximum closure.

(4) is the point at which opening begins.

(5) is the instant at which the entire length of the glottis is open.

(Taken from Hess, 1983; Base on Lecluse, 1977).

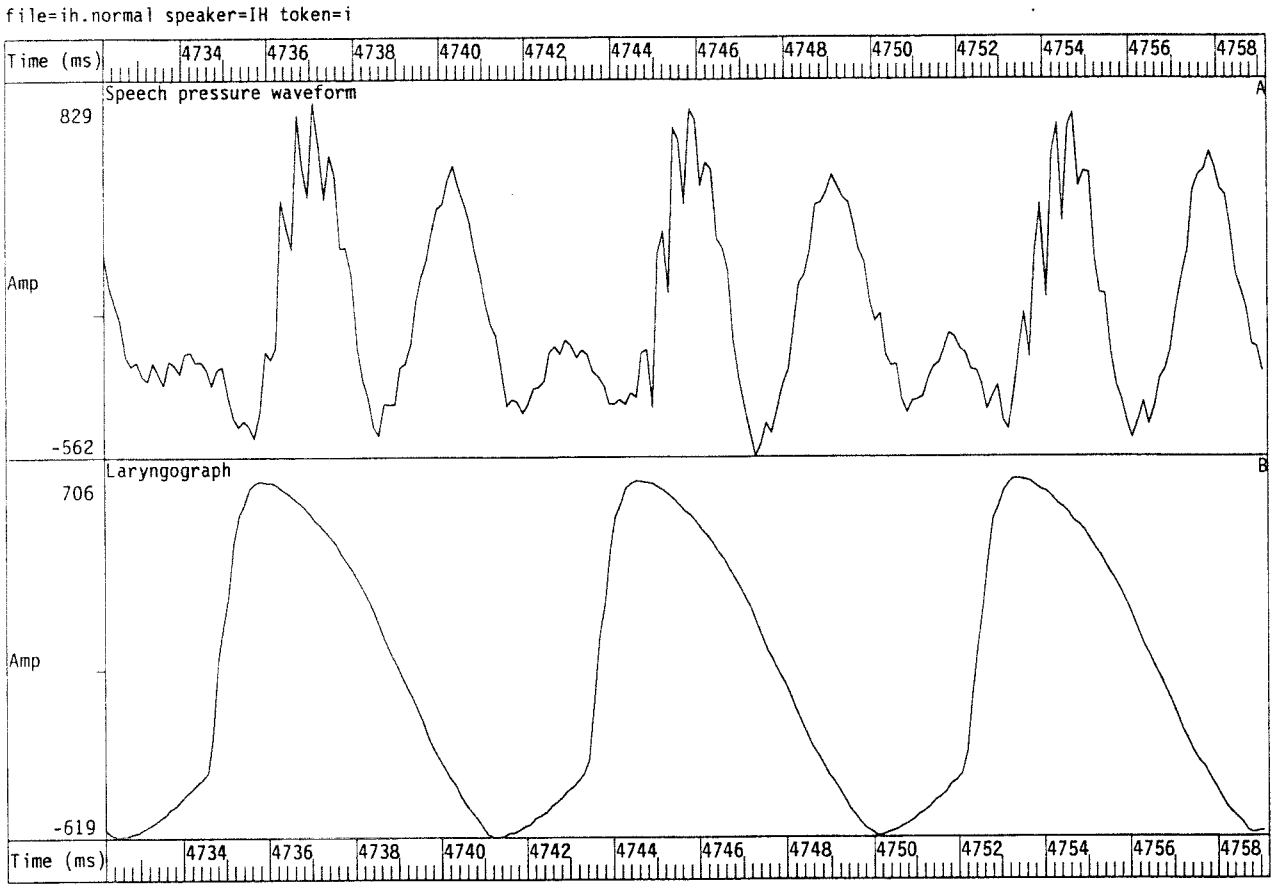

Figure 2.7 Speech pressure waveform and laryngograph waveform for an example of normal speech.

The laryngograph waveshape is similar to that shown in figure 2.6, except high-pass filtering present in the laryngograph has resulted in sloping of the horizontal sections of the waveform.

The utterance is the vowel /i/ spoken by a male.

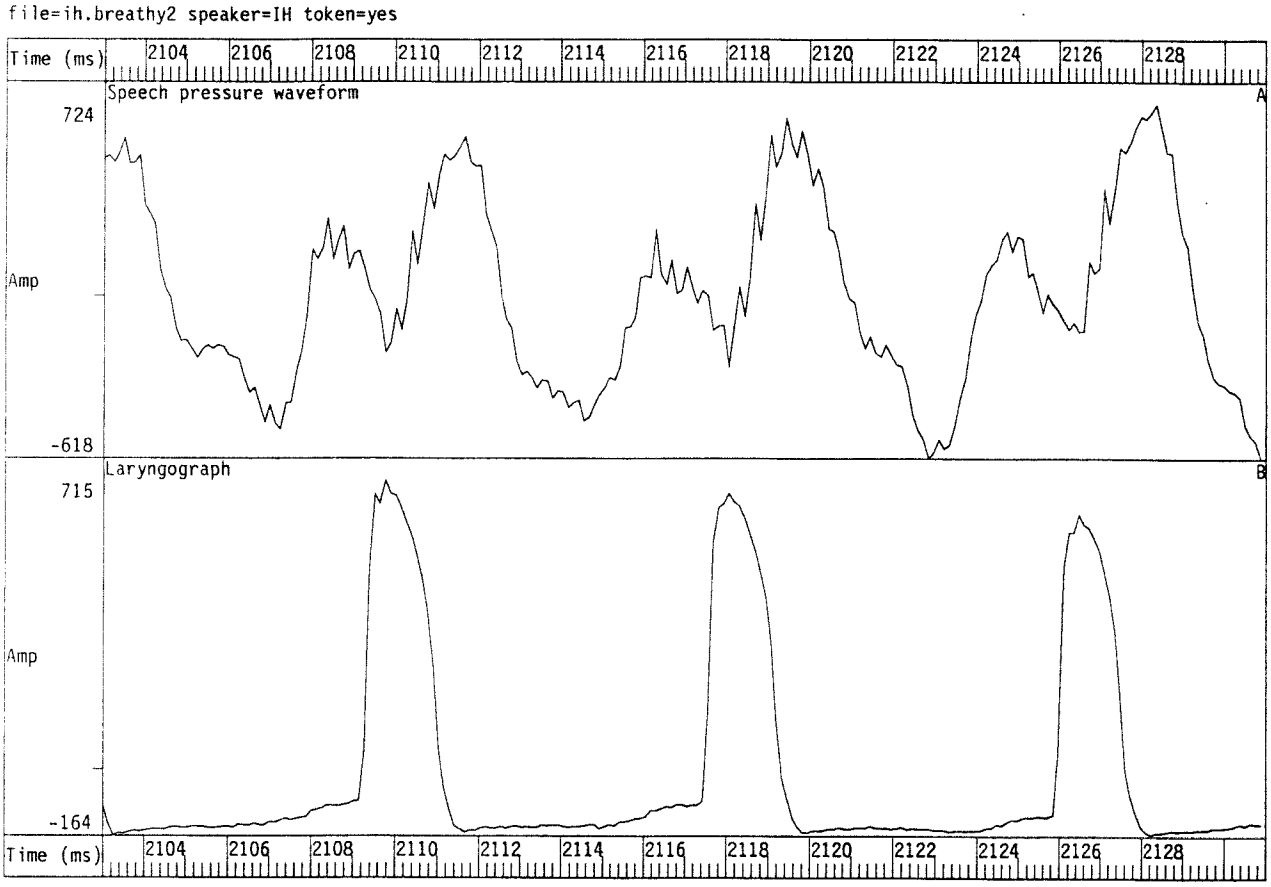

Figure 2.8 Speech pressure waveform and laryngograph waveform for an example of breathy voice quality.

It can be seen that the vocal folds maintain firm closure for a smaller proportion of the period than in the case of normal voice quality. Consequently the laryngograph waveform is positive for a smaller portion of the overall cycle. The utterance is the vowel /i/ spoken by a male.

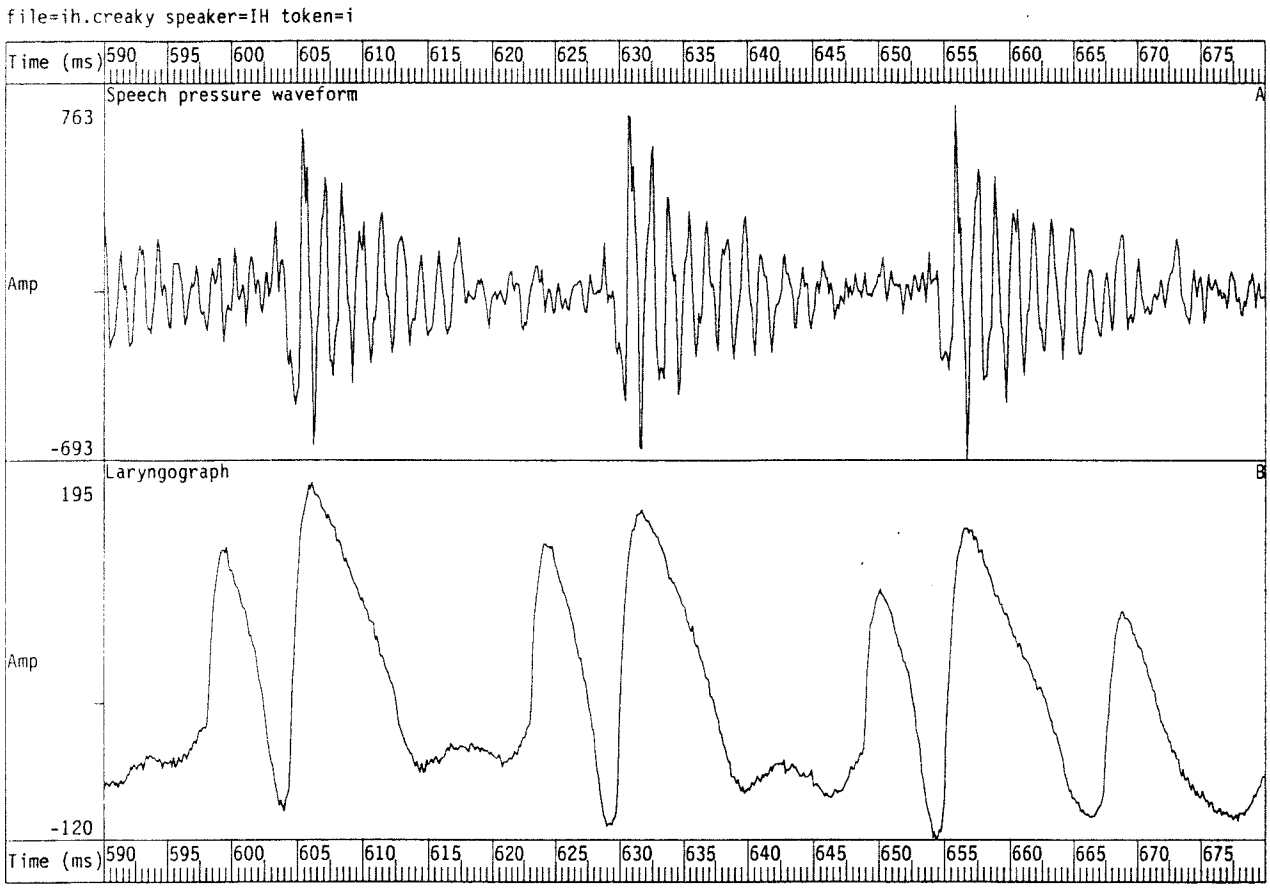

Figure 2.9 Speech pressure waveform and laryngograph waveform for an example of creaky voice quality.

In this case, the vocal fold closures occur irregularly, sometimes with a long closure followed by a shorter closure. The utterance is the vowel /i/ spoken by a male.

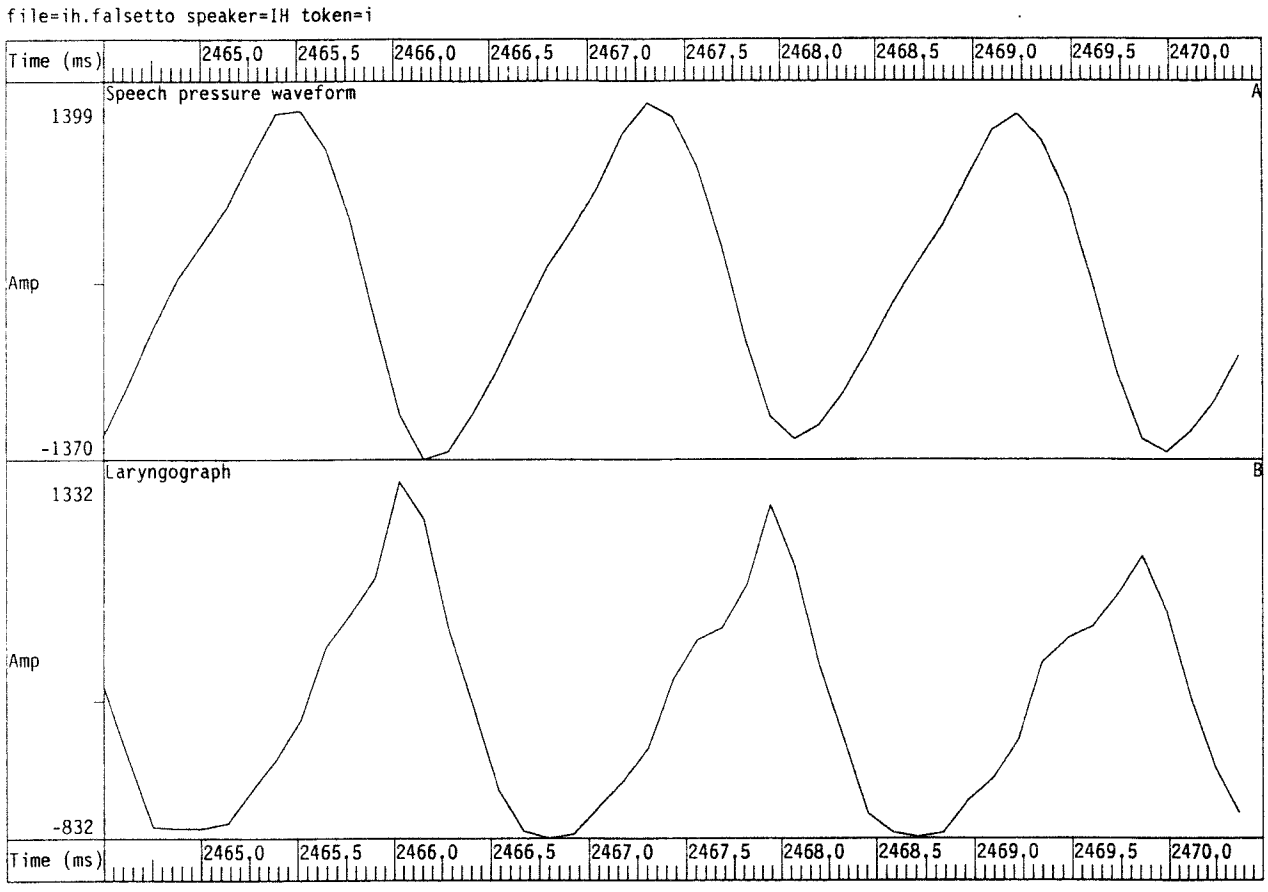

Figure 2.10 Speech pressure waveform and laryngograph waveform for an example of falsetto voice quality.

The utterance is the vowel /i/ spoken by a male.

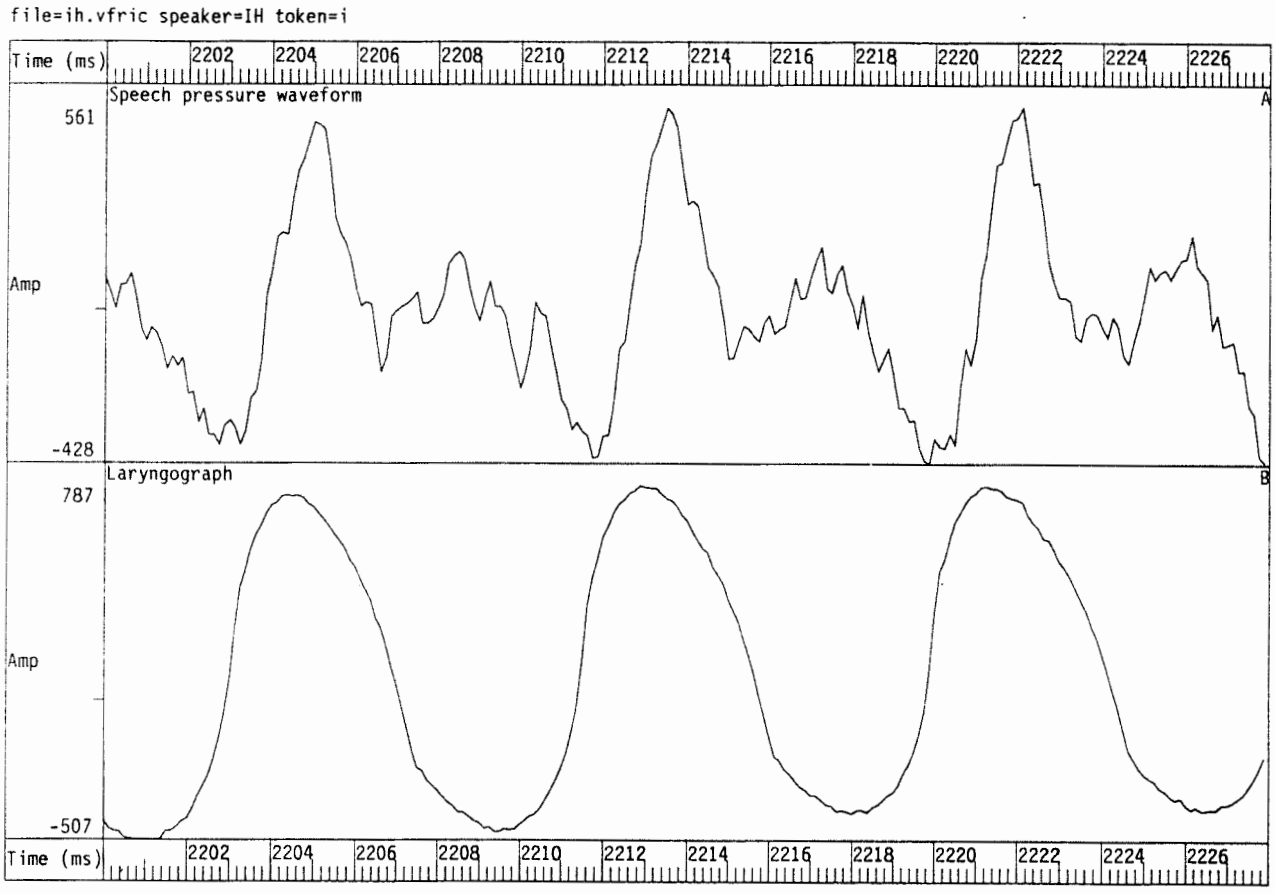

Figure 2.11 Speech pressure waveform and laryngograph waveform for a voiced fricative.

There is fricative excitation in addition to the quasi-period excitation due to vocal fold vibration. It can be seen that the frication occurs synchronously with the vocal fold vibrations. The utterance is the voiced fricative /z/ spoken by a male.

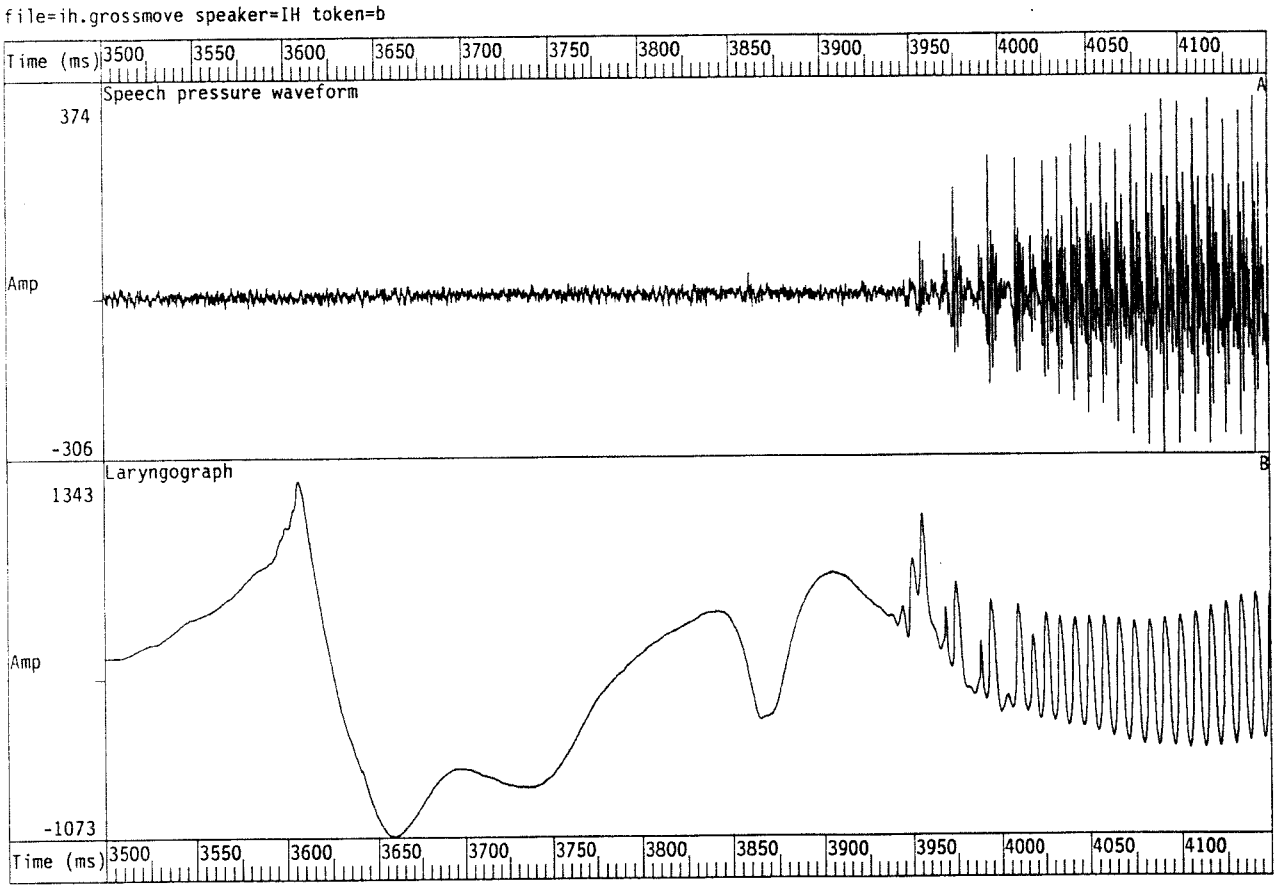

Figure 2.12 Unwanted excursion in laryngograph output waveform.

It can be seen that prior to phonation there are spurious excursions of the laryngograph waveform that have no acoustic significance. The utterance is the onset of the vowel /i/ spoken by a male.

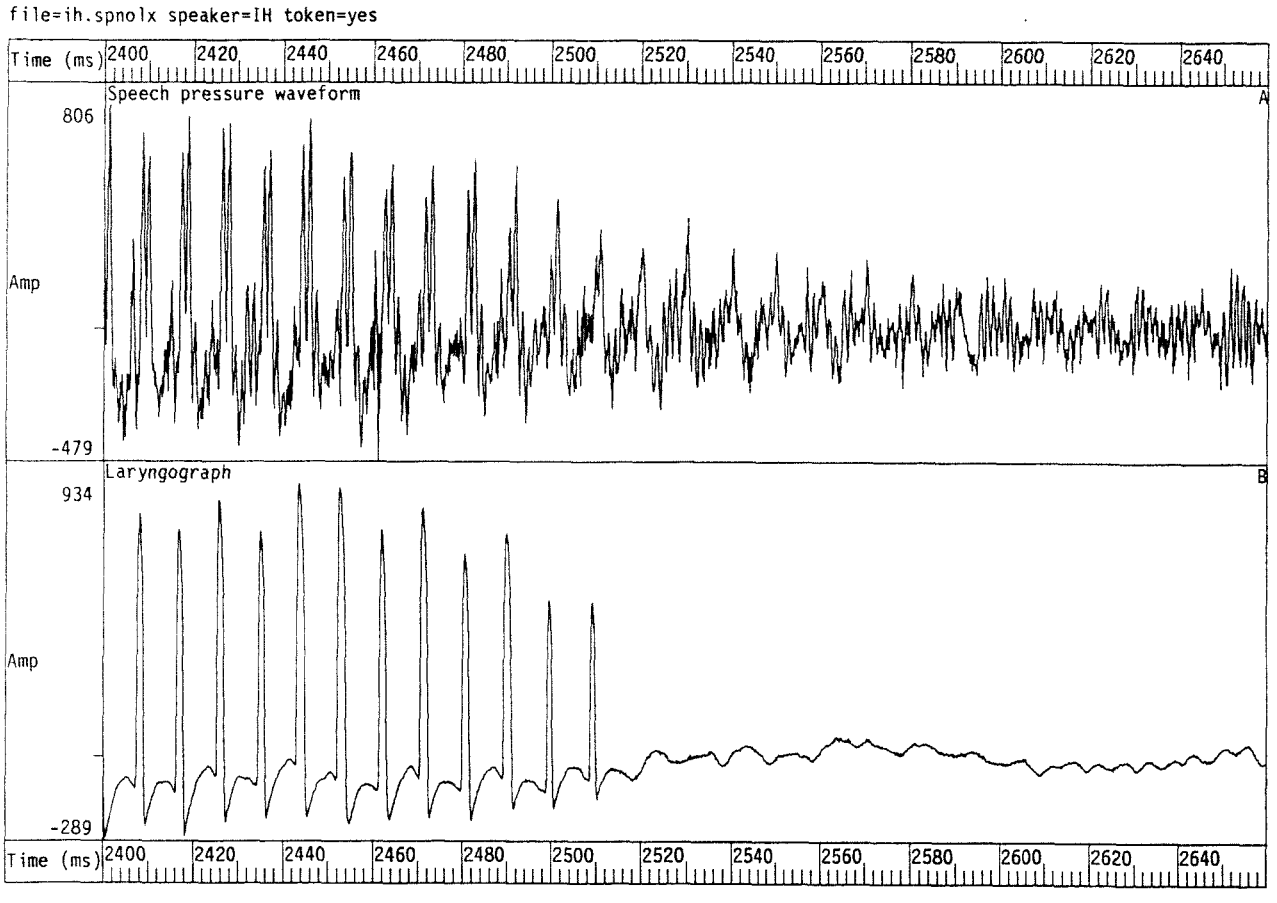

Figure 2.13 Evidence of vocal fold vibration in the speech pressure waveform, but little in the laryngograph signal.

This situation arises when firm vocal fold contact is not made, but the vocal folds are still vibrating. The section shown is the end of the utterance "yes" spoken using a breathy voice quality by a male.

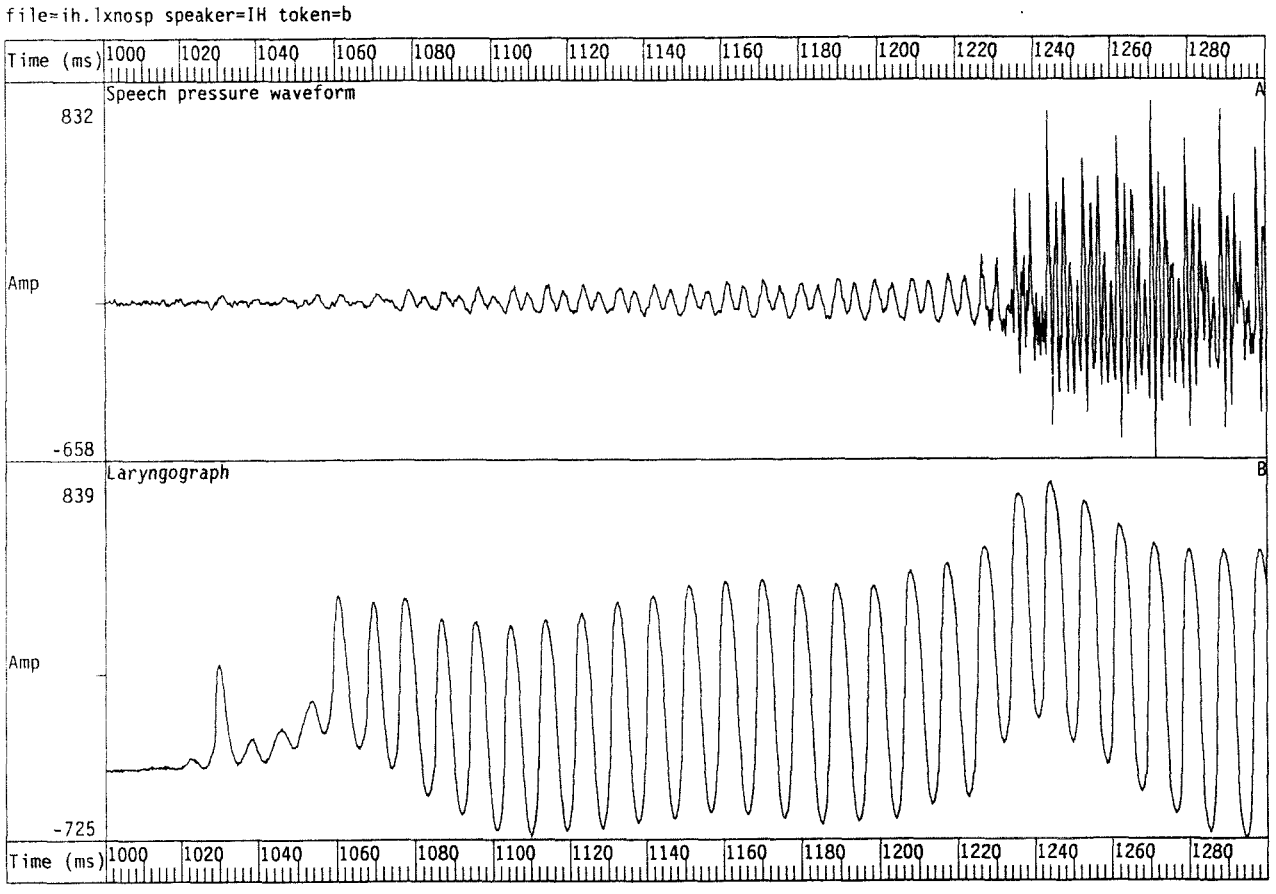

Figure 2.14 Evidence of vocal fold vibration in the laryngograph waveform, but only a small amount in the acoustic speech pressure waveform.

This occurs when there is a block in the vocal tract, such as in the case of the hold stage in a plosive, but there in still sufficient air flow through the larynx to maintain vocal fold vibration (this air flow results in an increase of air pressure behind the constriction). The section shown is the lead up to the plosive /b/, spoken by a male.