

manner. Rather, the central-peripheral coupling is bi-directional, since feedback from the periphery affects the clock. Studies of rhythmic finger movements (Kay, Saltzman, and Kelso, 1991) suggest that mechanical perturbations do introduce shifts in the phasing of such movements. Results reported by Saltzman (1992) and by Saltzman, Löfqvist, Kay, Rubin, and Kinsella-Shaw (1992, forthcoming) indicate that this is also the case for speech, at least when the speech task consists of the repetition of a single consonant-vowel syllable.

4 Summary

The theoretical and empirical approaches to speech production that we have discussed in this chapter converge in their focus on understanding how the different parts of the vocal tract are flexibly marshaled and coordinated to produce the acoustic signal that the speaker uses to convey a message. A variety of experimental paradigms are currently being applied to the problem of coordination and control in motor systems with excess degrees of freedom. Progress in speech motor control is likely to benefit from input from other areas of movement control and in using a combined strategy of empirical studies and mathematical modeling.

NOTE

I am grateful to Vincent L. Gracco, Laura L. Koenig and Elliot Saltzman for discussions and comments on earlier versions of this manuscript. This work

was supported by Grant DC-00865 from the National Institute on Deafness and Other Communication Disorders.

14 Voice Source Variation

AILBHE NÍ CHASAIDE AND
CHRISTER GOBL

1 Introduction

This chapter deals with acoustic aspects of voiced phonation. More specifically, it focuses on the voice source which is typically defined as the airflow (or volume velocity) through the glottis, and it varies over time in a periodic way which reflects the rapid opening and shutting cycles of the vibrating vocal folds. (The source for voiceless sounds is not dealt with here.) The glottal airflow signal constitutes the input signal to the vocal tract which acts as an acoustic filter. The configuration of the supraglottal vocal organs determines the specific resonance characteristics of this filter. For a given source signal, a large number of segments may be differentiated from each other on the basis of the particular patterning of resonances and antiresonances that different supraglottal filters impart.

Figure 14.1 shows a schematic illustration of the speech production process for two vowels [u] and [i]. The source spectrum is identical in both cases: it contains all harmonic components and has a constant slope of -12 dB per octave. This means that the amplitude of the harmonics decreases monotonically with increasing frequency, so that for every doubling of frequency, the amplitude has dropped by 12 dB. We should note that this is the ideal case. The true source spectrum does not have a constant slope, and may present dips depending on the precise shape of the glottal pulse.

The filtering effect of the vocal tract (referred to as the transfer function) is rather different for the two vowels, due to the different positions adopted by the tongue and the lips. Source harmonics which fall at or near the peaks of the transfer function will be amplified by the filter. Harmonics which do not come near to the peaks will not be amplified and may be attenuated. Consequently, the output of the filter, i.e., the oral airflow, has a spectrum exhibiting peaks and dips rather than the relatively evenly falling source spectrum, and these determine the different segmental qualities of the sounds we hear (in this instance the difference between [u] and [i]). Finally, the radiated sound

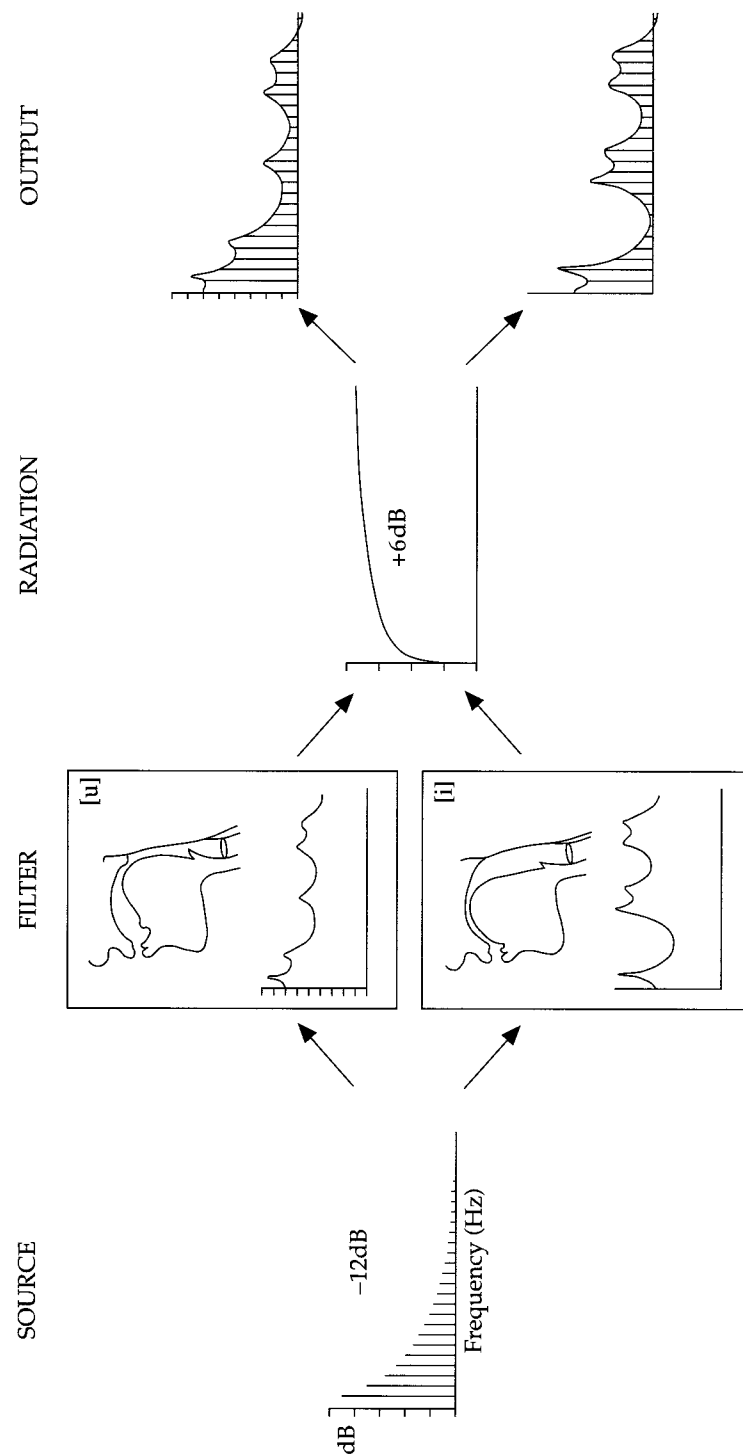


Figure 14.1 A schematic illustration of the speech production process for the vowels [u] and [i].

pressure at the lips has a spectrum that is tilted by approximately +6 dB per octave in comparison to the spectral slope of the oral airflow.

The illustration in Figure 14.1 is of an idealised source which is assumed to be constant for the two vowels. In real speech, the source varies dynamically in a way that reflects the configuration of the glottis, the degree and type of any laryngeal tension that may be present, the respiratory effort being used, and even the aerodynamic consequences of any supraglottal stricture. Gobl (1988) illustrates how the source may vary in the course of a single utterance spoken with a neutral, modal mode of phonation. The variation is even greater if the speaker chooses to switch between different modes of phonation (e.g., breathy voice, creaky voice, etc.) as is often done for paralinguistic signalling of emotion and attitude. Different speakers may also vary considerably in terms of the habitual type of phonation they use.

Over the years, much work has been carried out on the acoustics of the filter, (see Fujimura and Erickson, *ACOUSTIC PHONETICS*) which corresponds to much of the segmental differentiation of place and manner of articulation. Concerning the voice source, a good deal is known about f_0 variation, and how it varies as a function of intonation, tone and stress. Relatively little is known about other aspects of the voice source and how it varies in speech. (There are of course also many studies on the intensity variation of the speech signal. Although the amplitude of the speech output to some extent reflects the amplitude of the source, one should bear in mind that the total amplitude of the speech output is a function of both source and filter.)

In the next section, ways of analysing and measuring the voice source are discussed. This is followed in Section 3 by brief illustrations of how the source varies for a number of different voice qualities. In Section 4 we give an overview of the factors that determine voice source variation in speech and language.

2 Analysing the voice source

2.1 Obtaining glottal flow: inverse filtering

Most experimental studies of the voice source have been based on inverse filtering. This technique is effectively a reversal of the speech production process. The speech signal is passed through a filter whose transfer function is the inverse of the supraglottal transfer function. In principle this yields the voice source in its prefiltered form, as the filtering effect of the vocal tract is cancelled. Figure 14.2 illustrates this process in the frequency domain (in terms of the signal's frequency components) and in the time domain (in terms of the glottal airflow, or its derivative). Cancellation of lip radiation is not shown here for reasons that are explained below.

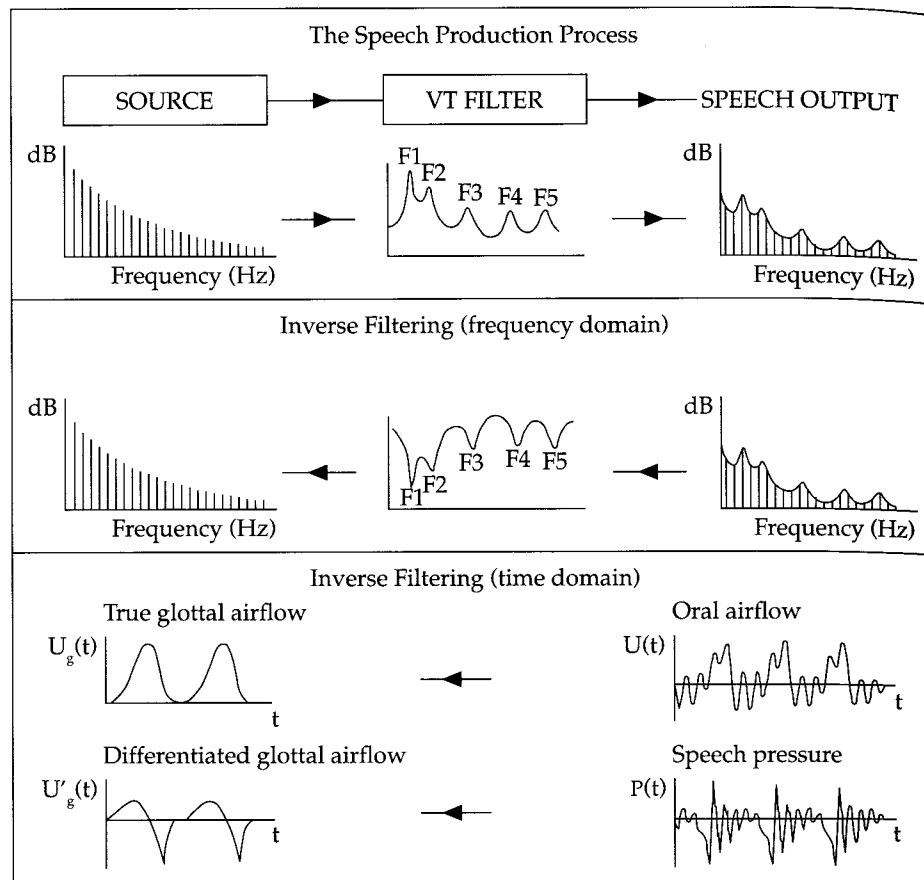


Figure 14.2 Schematic representation of inverse filtering in the frequency and time domains.

The inverse filter should contain a specification of the frequencies and bandwidths of the antiresonators (complex-conjugate zeros) required to cancel the formants (complex-conjugate poles) of the vocal tract transfer function at any given instant in time. It is important to get the number of poles right for the bandwidth determined by the sampling frequency. The average spacing between the poles is determined by the length of the vocal tract: for a typical male with a vocal tract of 17.5 cm we can expect one formant on average per 1000 Hz. The specification of the precise frequency and bandwidth is very critical for the lower formants, especially F1. Any error here will result in some distortion of the glottal pulse. Minor errors in the higher formants have little effect on the main pulse shape or its corresponding frequency spectrum (Gobl, 1988).

An all-pole function adequately describes the transfer function for many sounds such as vowels. For certain sounds such as nasals and laterals the vocal tract transfer function contains zeros as well as poles and in principle these zeros should be cancelled by the inclusion of corresponding poles in the inverse filter. As it is often difficult to estimate the zeros of the transfer function, most researchers tend, in practice, to use an all-pole model for all sounds. Although this simplifies the inverse filter specification, it does mean that sounds whose spectrum contains zeros are less accurately filtered.

To obtain the true glottal flow from the speech pressure wave, the filtering effect of the radiation at the lips needs to be cancelled as well. The radiation characteristics can be relatively accurately approximated by a first order differentiation (see, however, Fant, 1960: 44–45, for a more detailed description). The spectral consequence of the differentiation is a relative boosting of higher frequencies by 6 dB per octave. This effect can easily be cancelled by a simple integration of the signal (a real pole at zero frequency), as this is the inverse of differentiation. If the effect of the lip radiation is not cancelled, the output of the inverse filter will correspond to the differentiated glottal flow, also referred to as the glottal flow derivative. Many researchers opt to work with this signal rather than the true glottal flow. The emphasis of higher frequencies by 6 dB per octave permits a more precise modelling of the spectral slope of the source signal. It is also convenient for resynthesis purposes to lump the lip radiation with the source: one does not need first to remove it and then reintroduce it.

Inverse filtering based on the speech pressure waveform can yield detailed temporal and spectral information. However, the recording equipment and room are critical, and shortcomings in either condition can lead to disappointing results (see, for example, discussion and comments in Ladefoged, Maddieson, Jackson and Huffman, 1987). Ideally, an anechoic chamber should be used. The recording equipment must preserve the phase characteristics of the signal even at very low frequencies, which effectively means that a digital or FM recorder is needed unless the recording is done straight to computer. Analog tape recorders introduce phase distortion, and suggestions have been made by numerous authors on how this might be compensated for (Hedelin, 1986; Holmes, 1975; Hunt, 1978; Ljungqvist and Fujisaki, 1985).

Inverse filtering can also be carried out on recordings of oral airflow. In this case, a special airflow mask with a built-in differential pressure transducer must be used. Many studies have employed the circumferentially vented pneumotachograph mask designed by Martin Rothenberg (Rothenberg, 1973). When oral airflow is inverse filtered, the output is the true glottal flow. If the differentiated glottal flow is required, a real zero at zero frequency (a first order differentiator) is added to the inverse filter. The main advantage of using oral airflow recordings is that absolute values of the airflow rate can be measured, which is not possible from recordings of the speech pressure wave. This is particularly useful for measuring the "DC-leakage" during phonation where the glottal cycle lacks complete closure during the so-called closed phase. The main

disadvantage with this approach arises out of the limited frequency response of the mask. In the best case, e.g., the Rothenberg mask, the frequency response is limited to slightly over 1 kHz (see Hertegård and Gauffin, 1992; Badin, Hertegård and Karlsson, 1990). As a consequence, it does not provide for detailed spectral analysis of the source.

For a successful source analysis, it is of course essential that the estimate of the vocal tract transfer function be accurate. Many of the systems proposed for estimating the inverse filter involve fully automatic procedures, typically based on LPC analysis in one form or another. Unfortunately, they often do not yield satisfactory results for detailed source analysis, particularly where the vocal tract filter is undergoing rapid change or where the source involves a non-modal mode of phonation. At present, the most accurate source signal is obtained by using a method where the user interactively fine-tunes the formant frequencies and bandwidths of the inverse filter. Figure 14.3 is a slightly modified screen display illustrating the time and frequency domain information which guides the user in cancelling the formant peaks (in the frequency domain) and corresponding formant oscillations (in the time domain). The upper window shows the speech waveform, with a cursor marking the pulse under analysis. The second and third windows show (a) the speech waveform and (b) the inverse filter output (the differentiated glottal flow) for this pulse. The lowest window shows the corresponding spectra for (A) the speech waveform and (B) the differentiated glottal flow. The points marked as crosses in the lowest window indicate the formants, determining the complex zeros of the inverse filter. Using the mouse, each of these points can be moved in a horizontal or vertical direction to manipulate the frequency and bandwidth respectively. With each manipulation, the screen is instantaneously updated to show the new inverse filtered waveform (b) and its spectrum (B). For further details on this particular implementation, see Ní Chasaide, Gobl and Monahan (1992).

At present no automatic procedure can achieve the level of accuracy that the trained researcher can. Using the combined time and frequency information, many aspects of the source can be measured more accurately than would otherwise be possible. Yet there are also disadvantages with this method. In fine tuning the filter, it is sometimes necessary to compromise between the time and frequency information, and here it is vital that a consistent approach be adopted. This of course demands considerable skill and experience, and entails a risk that different experimenters will adopt different strategies leading to inconsistent results. Even with a trained researcher, there is some risk of circularity with this procedure. As the experimenter has certain expectations of what the glottal flow should look like, it could lead to an avoidance of unlikely-looking but valid pulse shapes. But probably the greatest problem of all is that the manual interactive method is not suited to the analysis of large amounts of data. As the analysis typically proceeds on a pulse-by-pulse basis, it is extremely time-consuming. This, and the high degree of vigilance needed, has resulted in these types of studies being limited to small amounts of carefully analysed data.

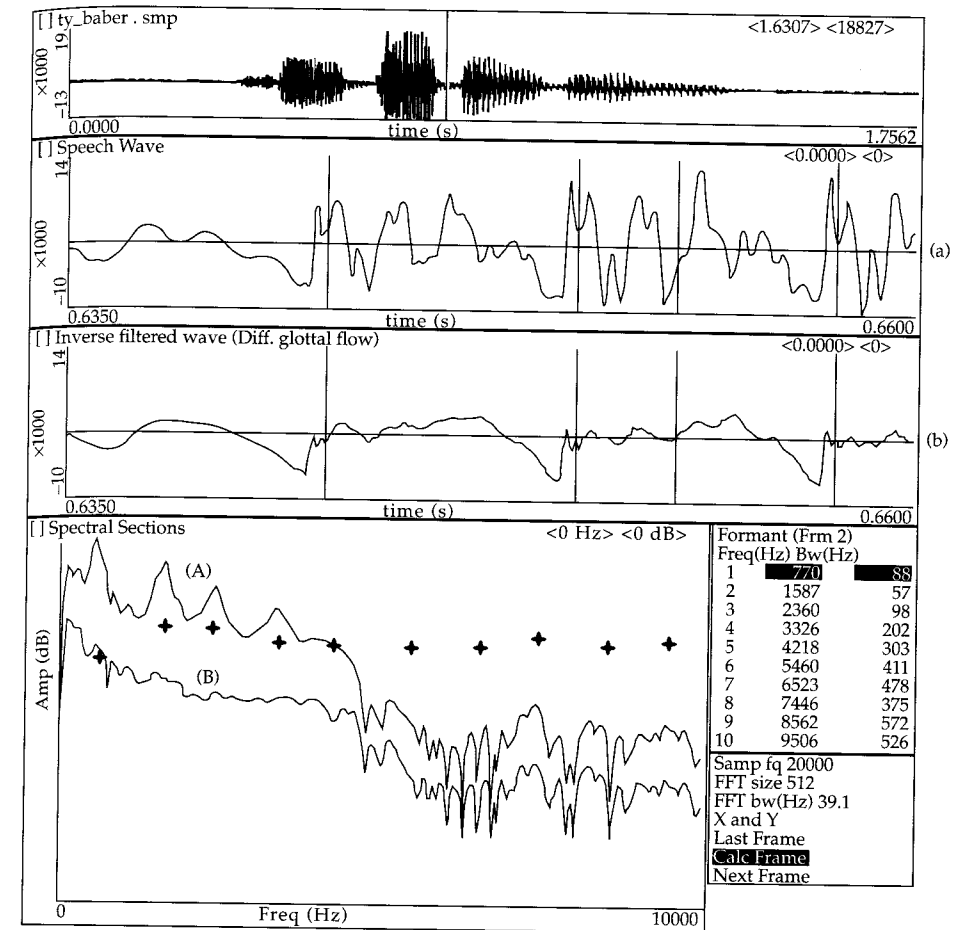


Figure 14.3 (Modified) Screen display illustrating an interactive inverse filtering method.

2.2 Voice source models

As stated above, the output signal of the inverse filter is an estimate of the glottal airflow or its derivative. Visual inspection may yield a first gross impression of some characteristics, whether the voicing is efficient, breathy, etc., but for fine comparisons precise measurements are required. These can be made directly from the source signal, as was done by Holmberg, Hillman and Perkell (1988) and by Huffman (1987). An alternative method involves matching a source model to the output of the inverse filter, and deriving the measurements from the modelled waveform.

For this approach to be successful, it is important that the model be a good

representation of the true source and that it be flexible enough to capture the important variations that may occur. Traditionally, the voice source in speech synthesis (see Carlson and Granström, *SPEECH SYNTHESIS*) was implemented as a low-pass filtered impulse train (Klatt, 1980a; Liljencrants, 1969). The only control parameters of this simple type of voice source are f_0 and the amplitude of the impulse. Its main drawback is that the spectral slope cannot be controlled. It is always perfectly regular, falling off monotonically at -12 dB per octave. Another drawback is that the phase characteristics of the filtered impulse are very different from that of the typical glottal pulse. The impulse response of the low-pass filter is time-reversed in comparison to the typical glottal waveform. This means that the main discontinuity of the waveform (corresponding to the main excitation) occurs at the rising branch rather than at the falling branch of the glottal pulse.

These drawbacks resulted in an inflexible and often unsatisfactory voice quality in speech synthesizers, and have prompted the development of more elaborate voice source models. The new voice source models all have a larger number of control parameters and more accurate representations of the glottal waveform. They are therefore more capable of capturing the frequency characteristics (e.g., the spectral slope) as well as the phase characteristics of the natural glottal waveform.

2.2.1 The LF model A model which has gained popularity in recent years, and which is used below for a number of illustrations is the LF model of differentiated glottal flow (Fant, Liljencrants and Lin, 1985). In addition to f_0 , this model has four parameters and a requirement of area balance (see below) which determine the waveshape. As can be seen in Figure 14.4, the model is made up of two segments. The first is an exponentially growing sinusoid used for modelling the differentiated flow from the time point of glottal opening t_0 to the time point of main excitation t_e . Three parameters determine the shape of this segment. These parameters are: (1) E_0 which is a scale factor, (2) $\alpha = -B\pi$ where B is the "negative bandwidth" of the exponentially growing sinusoid (i.e., the larger the α the faster the increase in amplitude) and (3) $\omega_g = 2\pi F_g$ where $F_g = 1/2t_p$ and t_p is the time of the opening branch (the time from glottal opening to maximum airflow).

The second segment is an exponential function which is used to model the differentiated flow from the time point of the main excitation, t_e , to the time point of glottal closure, t_c . This part of the glottal cycle is termed the return phase and determines the residual airflow (or dynamic leakage) after the main excitation when the vocal folds close. In the LF model, the control parameter of the return phase is TA. TA is the time constant of the exponential curve, which is determined by the projection on the time axis of the tangent at time t_e (see Figure 14.4).

The description of the LF model assumes that $t_c = t_0$, i.e., the time point of glottal closure is the same as the time point of glottal opening for the forthcoming pulse period. This implies that the model lacks a closed phase. In

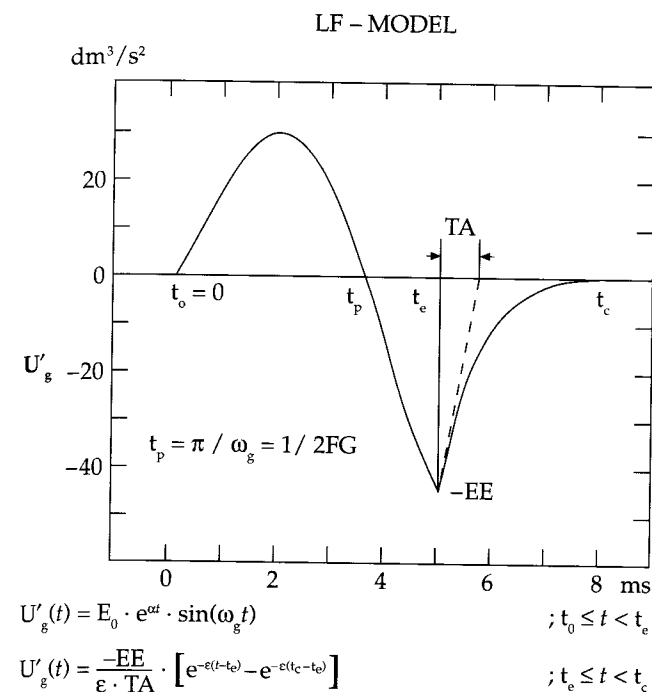


Figure 14.4 The LF model of differentiated glottal flow.

practise, for normal (small) values of TA, the exponential curve will fit closely to the zero line, providing, to all extents and purposes, a closed phase. This saves one parameter without any significant loss in flexibility. Furthermore, in order to unambiguously determine the pulse shape, the four LF parameters E_0 , α , ω_g and TA are complemented by a requirement of area balance. In other words, the positive area of the LF pulse (from t_0 to t_p), should always equal the negative area (from t_p to t_c). In terms of the true glottal flow, this means that the baseline of consecutive pulses is kept constant.

2.2.2 Other voice source models A number of other parametric voice source models have been proposed in the literature (see Ananthapadmanabha, 1984; Fant, 1979a, 1979b, 1982; Fant et al., 1985; Hedelin, 1984; Klatt and Klatt, 1990; Ljungqvist and Fujisaki, 1985; Rosenberg, 1971; Rothenberg, Carlson, Granström and Lindqvist-Gauffin, 1974). These models can be divided into two groups on the basis of whether they model the true glottal flow pulse or the differentiated one. They also differ in the number of parameters and in the functions they use to generate the glottal waveshape. Another important difference among them concerns whether or not they include a segment to model the return phase of the glottal pulse.

Fairly detailed comparisons of many of these voice source models can be

found in Ananthapadmanabha (1984) and in Ljungqvist and Fujisaki (1985). Figure 14.5 summarises some of the important features of seven different models.

2.3 Measuring the glottal signal: source model matching

As mentioned earlier, a method of extracting source measurements involves matching a voice source model to the inverse filter output, and deriving the measurements from the modelled waveform. This procedure has certain advantages over measuring parameters directly from time and amplitude points of the inverse filter output. First of all, the model matching allows us to take both time and frequency domain information into account, as the spectrum of the model can be calculated. This is particularly useful for capturing important features such as the return phase (see Section 2.4) which have important spectral consequences, but which are extremely difficult to measure accurately directly from the waveform. A further advantage is that the modelled source signal can be quickly implemented in synthesis, and in principle this should facilitate perceptual testing of the various parameters measured.

As with inverse filtering, the matching of the model can be done automatically (e.g., Chan and Brookes, 1989), but present automatic algorithms do not always yield reliable results. Again, more accurate measurements are obtained if a manual interactive approach is adopted as can be illustrated in relation to Figure 14.6, the screen display which guides the user in the matching process.

The mid panel of this figure shows the inverse filtered waveform (differentiated glottal flow) for the pulse specified in the top panel. Superimposed on this pulse, one can also see a matched LF pulse (thick line), whose contour is determined by four time points (vertical lines) and one amplitude point (horizontal line), which are manually set by the experimenter. The four time points are: (1) the time of glottal opening, t_o ; (2) the time of peak glottal flow, t_p ; (3) the time of the excitation, t_e ; (4) the time point on the basis of which the return phase is estimated, t_r (equals $t_e + TA$). The amplitude point (5) is the amplitude of the excitation, EE . The spectrum corresponding to the inverse filtered pulse is shown in the bottom panel and superimposed on it is the spectrum of the LF model pulse (thick line). The model pulse is optimised by making fine adjustments to the time and amplitude points in order to find the best overall agreement in both the time and frequency domains.

2.4 Some important voice source parameters

The LF parameters outlined in Section 2.2 determine the overall shape of the glottal pulse. For our analysis, we need to measure very specific aspects of this waveform, i.e., those aspects that are thought to be acoustically and perceptually important, and which can be more readily related to the underlying physiological events. Once the matching procedure has been satisfactorily completed,

	Model	Single flow derivative discontinuity	Provision for multiple flow derivative discontinuities	Provision for continuous flow derivative	Waveform realization
Amplitude, width and skewing of the glottal flow	(a)	yes	(yes)*	no	sinusoidal
	(b)	yes	no	no	sinusoidal
Independent control of flow derivative discontinuity	(c)	yes	no	yes	sinusoidal
Modeling of activity in the glottal closed phase	(d)	yes	no	yes	sin+polyn. exp.*sin. polynomial
	(e)	yes	no	yes	
	(f)	yes	no	yes	
	(g)	yes	yes	yes	

* Rosenberg proposed several models, some of which allow multiple discontinuities.

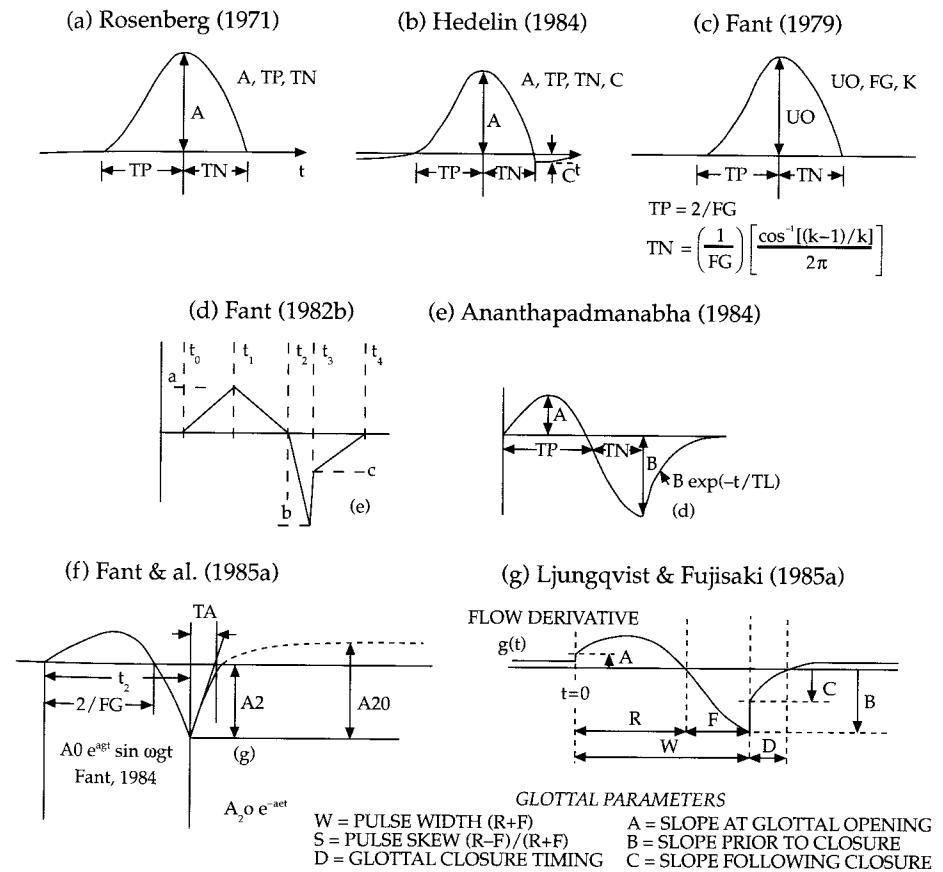


Figure 14.5 Waveforms and equations for seven proposed voice source models (after Ananthapadmanabha, 1984 and Ljungqvist and Fujisaki, 1985).

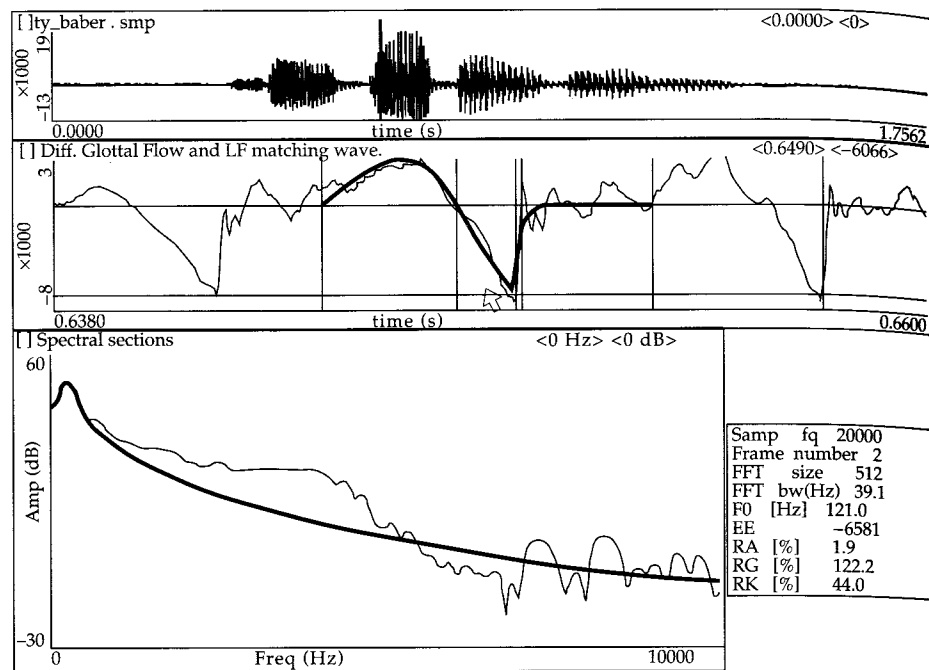


Figure 14.6 Screen display of the voice source matching method.

these source parameters can be calculated. We outline some of the most important parameters here, illustrating in Figure 14.7 how changes in the glottal waveform affect the acoustic spectrum. One must remember however that it is difficult to give a very precise specification of the spectral consequences of the individual source parameters, as they frequently interact in complex ways. The reader should also remember that the precise definition of these parameters depend on the model used. Here we define the parameters in terms of the LF model, which was used for a number of illustrations later in this chapter.

2.4.1 Fundamental frequency, f_0 The fundamental frequency = $1/T_0$, where T_0 , the fundamental period, is the time between two consecutive excitations.

2.4.2 Excitation strength, EE The excitation strength is the negative amplitude at the time-point of maximum discontinuity of the differentiated flow. It normally occurs at the maximum slope of the falling branch of the glottal pulse, which typically precedes full closure. At the production level it is determined by the speed of closure of the vocal folds and by the airflow through them. At the acoustic level it corresponds to the overall intensity of the signal. This parameter is the one that is most similar to the amplitude parameter of the traditional simple impulse source.

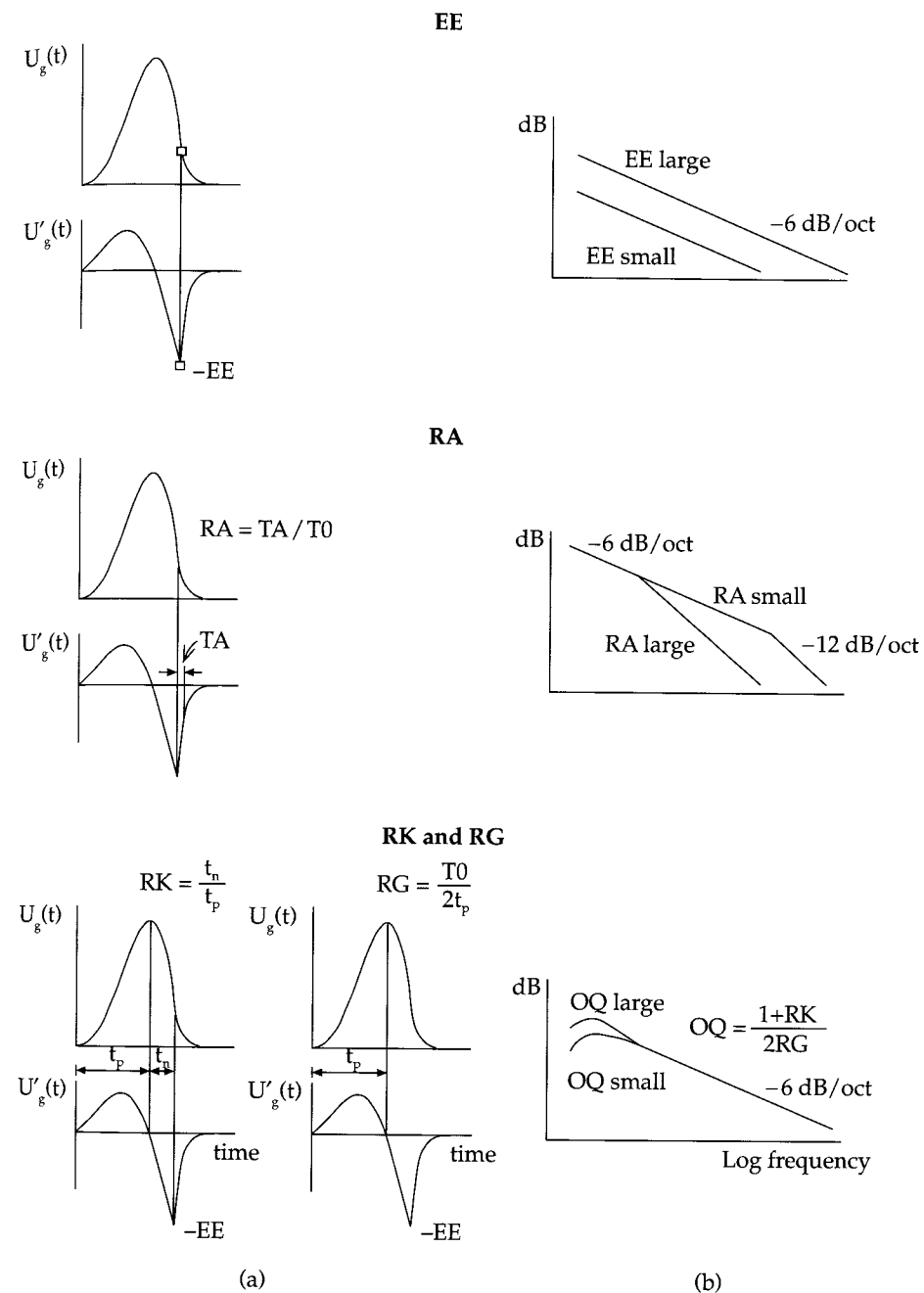


Figure 14.7 The voice source parameters EE, RA, RK and RG, in terms of the true and differentiated glottal flow, showing how changes in these parameters affect the acoustic spectrum (for explanations, see text).

2.4.3 Dynamic leakage, RA The dynamic leakage is the residual flow during the return phase, which occurs from the time of the excitation to the time of complete closure (or maximum closure if there is a DC leakage). In terms of the true glottal flow, the return phase shows up as a "rounding of the corner" of the closing branch of the pulse. In terms of the LF model, RA is equal to TA/T_0 where TA is the time constant of the exponential function modelling the return phase (see Figure 14.4). At the production level, RA relates to the sharpness of the glottal closure, that is, to whether the vocal folds make contact in an instantaneous way or in a more gradual fashion along their entire length and depth. Differences in dynamic leakage are important acoustically because they affect the spectral slope of the signal. The frequency characteristics of the exponential function of the return phase are approximately those of a first order low-pass filter. The cutoff frequency, FA, is inversely proportional to TA: $FA = 1/2\pi TA = f_0/2\pi RA$, i.e., the cutoff frequency of the filter is inversely correlated with the amount of dynamic leakage.

2.4.4 Open quotient, OQ The open quotient is a frequently used parameter. In terms of the source spectrum, it mainly controls the amplitude of the lower components. A related parameter which tends to covary with the open quotient is UP, the peak volume velocity of the glottal pulse (labelled as A, A, and U0 respectively in the first three glottal models of Figure 14.5). A large value in either of these corresponds to an increased level of the very lowest harmonics of the source spectrum.

2.4.5 Glottal frequency, FG The glottal frequency is defined as $1/2t_p$ (Fant, 1979a), i.e., it is a frequency determined by the time period of the opening branch of the glottal pulse. A more practical expression of this parameter is RG, essentially the same as FG, but normalised to f_0 , so that $RG = FG/f_0$. RG tends to vary inversely with OQ and UP. Consequently, a high RG is found with attenuated levels of the lowest end of the spectrum. The higher the RG value, the more it will approach the frequency of the second harmonic, H2, and the more it will contribute to boosting its level. A high RG and a relatively strong H2 tend to be characteristic of tense or pressed phonation. Low RG values are found where UP is high, where it contributes to boosting H1. Thus RG contributes (with OQ and UP) to the relative amplitude of H1 and H2 in the speech output, a measure frequently used in the linguistic literature (see more on this below).

2.4.6 Glottal symmetry/skew, RK Quite a lot of attention has also been given to the skewing of the pulse. In comparison to the underlying glottal area function, the glottal flow pulse is typically skewed to the right, i.e., the opening phase tends to be longer than the closing phase. This would appear to be due to the inertive loading of the vocal tract. The acoustic consequences of pulse skewing are somewhat complex. It affects mainly the lower part of the source spectrum so that a more symmetrical pulse shape has the effect of boosting the

lower harmonics. However, the degree of skewing also determines the depth of the notches (weakened or missing harmonics) in the source spectrum: the more symmetrical the pulse, the deeper the spectral dips. The locations of the notches are determined by the open quotient together with the pulse shape (cf. Flanagan, 1972: 236–242). It may be the case that the perceptual importance of the skewing has been relatively overestimated. Skewing is typically highly correlated with the excitation strength, and its perceptual contribution is easily confused with that of the excitation strength. The risk of such confusion is particularly high if a voice source model is used which lacks direct control of the excitation strength. In other words, we would suggest that the excitation strength is fundamentally a more important parameter than the skewing of the pulse.

2.4.7 Aspiration noise, AH A parameter which is often not explicitly included in voice source models is the aspiration noise. The importance of mixed excitation (periodic excitation mixed with aspiration noise) has been mentioned on several occasions (Dolansky and Tjernlund, 1968; Fujimura, 1968; Gobl, 1989; Gobl and Ní Chasaide, 1988; Hunt, 1987; Klatt, 1986a; Ladefoged and Antoñanzas-Barroso, 1985; Pandit, 1957; Rothenberg, 1974), but the noise component is difficult to estimate quantitatively. Most of the voice source models discussed above generate a perfectly harmonic spectrum, and thus do not directly include control of aspiration noise in the voiced excitation. Even if a voice source model does not explicitly incorporate a parameter for aspiration noise, a noise generator can always be used together with any source model to provide the noise component of the voiced excitation. Issues like the actual spectral content of the aspiration noise, strategies for controlling the level of aspiration noise, and the question of how to modulate the noise within a glottal period have been discussed by, for example, Klatt and Klatt (1990); Makhoul, Vishwanathan, Schwartz and Huggins (1978); Rothenberg et al. (1974).

The pulse-to-pulse stability of source parameters is also an important factor in determining voice quality. Traditionally, measures such as jitter and shimmer have been used to quantify pulse-to-pulse stability. Jitter is the random variation in f_0 and shimmer equals fluctuations of the pulse-to-pulse amplitude. Shimmer is often measured from the speech waveform amplitude, which can lead to errors as it is to some extent influenced by source-filter interaction effects. Ideally, shimmer should be measured directly from the amplitude of the glottal waveform, for example, EE. High levels of jitter and shimmer have often been found to correlate with hoarse voice. Note, however, that other source parameters are also likely to exhibit instability in certain circumstances, a fact which is probably also of perceptual importance. For examples, see the illustration of a pathological voice in Figure 14.13 and of a normal creaky voice in Figure 14.9.

Gobl (1988) has shown that many of the above mentioned source parameters tend to covary. EE is highly correlated with the negative amplitude of the speech waveform and other source parameters are often correlated with

EE. For example, the return phase typically varies inversely with EE, so that if the excitation is weaker, RA is higher. There is generally also covariation between RA and RK, so that a long return phase (and a low EE) corresponds to a more symmetrical pulse shape. Several of these tendencies have been corroborated by subsequent work (Pierrehumbert, 1989; Fant, 1994) but are not invariably present as indicated in Gobl and Ní Chasaide (1992). As our state of knowledge increases, it may become possible to predict many of the source parameters from a few of the more basic ones (hopefully the more easily measured ones) and this is an approach currently being pursued by Fant (1994). Although it is too soon to know how far the correlations he posits can be generalised, these approximations should nevertheless in the short term yield major improvements in applications such as speech synthesis.

2.5 Spectral measurements relevant to the voice source

In the preceding section, we have concentrated on time domain measurements of the glottal source, linking these to their expected spectral consequences. Frequency domain measurements can also be carried out on the output of the inverse filter. As is probably clear from the description of source parameters above, one may need to distinguish the very lowest frequencies from higher regions in any attempt to characterise and compare source spectra. The picture can be further complicated by the appearance of spectral notches, or even additional subglottal pole/zero pairs. Specific glottal pulse shapes (very symmetrical) can give rise to notches, and might be found, for example, in breathy voice. Furthermore, the more the glottis is abducted, the greater the coupling to the subglottal system and the greater the likelihood of subglottal resonances showing up in the source spectrum.

The spectral tilt is probably the most fundamental parameter one would want to measure in the source spectrum. Obtaining it is not always a simple matter as has been demonstrated by Jackson, Ladefoged, Huffman and Antoñanzas-Barroso, (1985; 1986), who explored the possibility of fitting a single regression line to source spectra. One possible method for comparing source spectra, which takes account of changing levels in different frequency regions is illustrated in Figure 14.12, and explained in Section 3.

Spectral measurements based on the speech output signal can also be very useful. For identical speech items, differing only in voice quality, average spectra (as in Figure 14.11) or even long term average spectra can help to demonstrate source differences. A measure frequently used is the comparison of the level of the first harmonic (H1) with the level of some higher frequency component. A comparison of H1 and F1 levels has been used in a number of studies (see, for example, Ní Chasaide and Gobl, 1993; Kirk, Ladefoged and Ladefoged, 1984 and Figure 14.10). Another popular measure has involved the comparison of the level of the first two harmonics (see, for example, Bickley, 1982; Fischer-Jørgensen, 1967; Maddieson and Ladefoged, 1985). A very

dominant H1 has been widely found to be highly correlated with a breathy mode of phonation whereas a relatively strong H2 can be correlated with tense or creaky voice.

Measurements based on the speech output waveform are particularly attractive to linguists working in the field, in that they do not require the level of technical facilities which the execution of inverse filtering and model matching require. However, it is important to bear in mind that although these types of measurements reflect differences in the source spectrum, they are also sensitive to other factors and can therefore not be used to infer the actual slope of the source spectrum. It is important for the experimenter to be aware of the other factors that can affect the level of different frequencies of the output spectrum, as the speech materials must be carefully chosen to take account of them. First of all, the frequencies of the formants affect their levels, and so a comparison of H1 and F1 levels would clearly not be appropriate across different vowel qualities. Formant levels are also partially determined by the degree of damping present. A high degree of damping is found where there is little or no closed phase in the glottal pulse, as for example, in breathy voice. Supraglottal factors may also affect the degree to which formants are damped. In any case, formant damping affects the levels of the output spectrum in a way that does not directly reflect the slope of the source spectrum.

All of these spectral measures are also sensitive to f_0 differences, or more precisely, to any shift in the ratio of f_0 to F1 frequencies. For example, the comparison of H1 and H2 levels may be a valid measure when F1 is high and f_0 low. However, when F1 is low or f_0 is high (or in the worst case where both of these factors pertain), the levels of H1 or H2 may be boosted depending on their proximity to the F1 peak. In such cases the relative levels of H1 and H2 are influenced by filter as well as by source factors, and so are no longer reliable indicators of the mode of phonation.

3 Some commonly occurring voice qualities

As a backdrop to Section 4 we present here a brief sketch of a few commonly occurring voice qualities. The aim is not only to show how these voice qualities may differ acoustically, but also to illustrate different kinds of measurements that are useful. The voice qualities we deal with are modal voice, breathy voice, whispery voice, creaky voice, tense voice and lax voice as described in Laver (1980). (Note that although we are concerned here only with the laryngeal aspects of voice quality, the last two mentioned may involve greater or lesser degrees of tension in the entire speech apparatus, and not purely of the phonatory system.)

The physiological descriptions here are in terms of three hypothesised parameters of muscular tension; adductive tension, medial compression and longitudinal tension (see illustration in Figure 14.8 from Laver, 1980). These

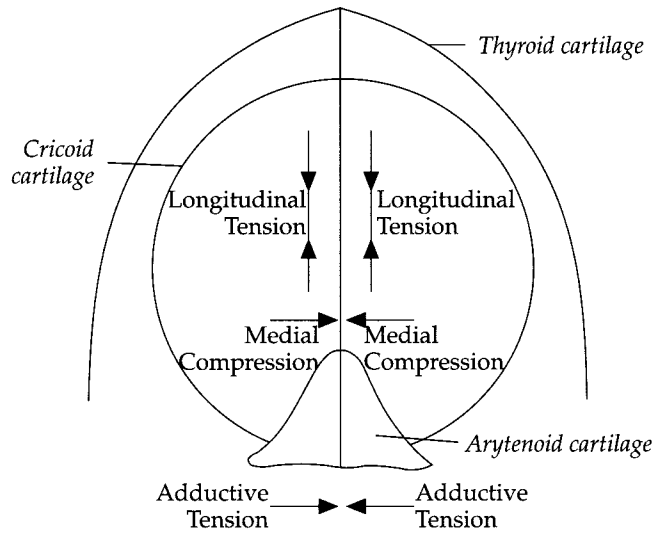


Figure 14.8 Three laryngeal parameters of muscular tension as described in Laver (1980).

determine the configuration and tension settings of the vocal folds, and interact with aerodynamic factors related to subglottal pressure and glottal airflow to yield a variety of voice qualities. For a fuller description the reader is referred to that text. (See also the descriptions of voice quality in Catford (1964) and Ladefoged (1971).) *Adductive tension* is defined as the force by which the arytenoids are drawn together, so that the cartilaginous glottis is adducted. It is controlled by the interarytenoid muscles. *Medial compression* is defined as the force by which the ligamental glottis is closed, through the approximation of the vocal processes of the arytenoids. It is primarily controlled by the lateral cricoarytenoid muscle, but the external thyroarytenoid muscle can also be involved. *Longitudinal tension* is the tension of the vocal folds, and is mediated primarily by contraction of the vocalis and of the cricothyroid muscles, whose main function is to control pitch (see also Hirose, INVESTIGATING THE PHYSIOLOGY OF LARYNGEAL STRUCTURES).

Some of the acoustic characteristics of these voice qualities are illustrated in Figures 14.9–12, and were derived using the analysis techniques outlined in Section 2. The speech materials were produced by a male phonetician, well acquainted with the Laver system for characterizing voice qualities (see above). Figure 14.9 shows source parameter values; Figures 14.10 and 14.11 show spectral measures of the speech output signal for the vocalic interval of the relatively unstressed word *strikes*. For four of the qualities, a schematic representation of the source spectra is shown in Figure 14.12, measured at the mid point of the stressed /a/ in the nonsense word *babber*. The aim here was to facilitate comparison of spectral slopes by showing the extent to which the

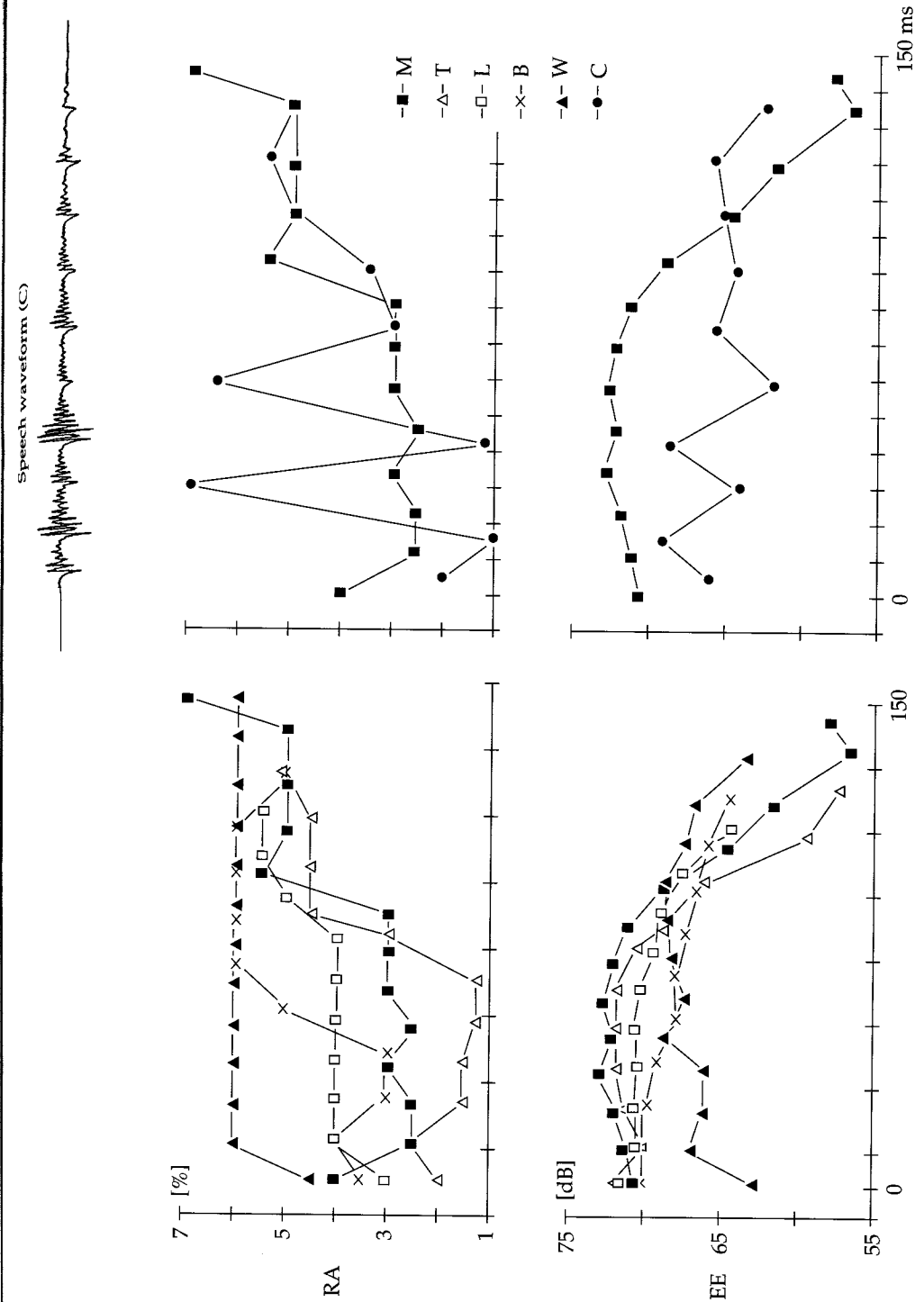


Figure 14.9 Pulse-by-pulse values for RA and EE for the voiced interval of /straiks/. Left panel shows modal (M), tense (T), lax (L), breathy (B) and whispery (W) voice. Right panel shows modal and creaky (C) voice, and additionally shows the speech waveform in the latter.

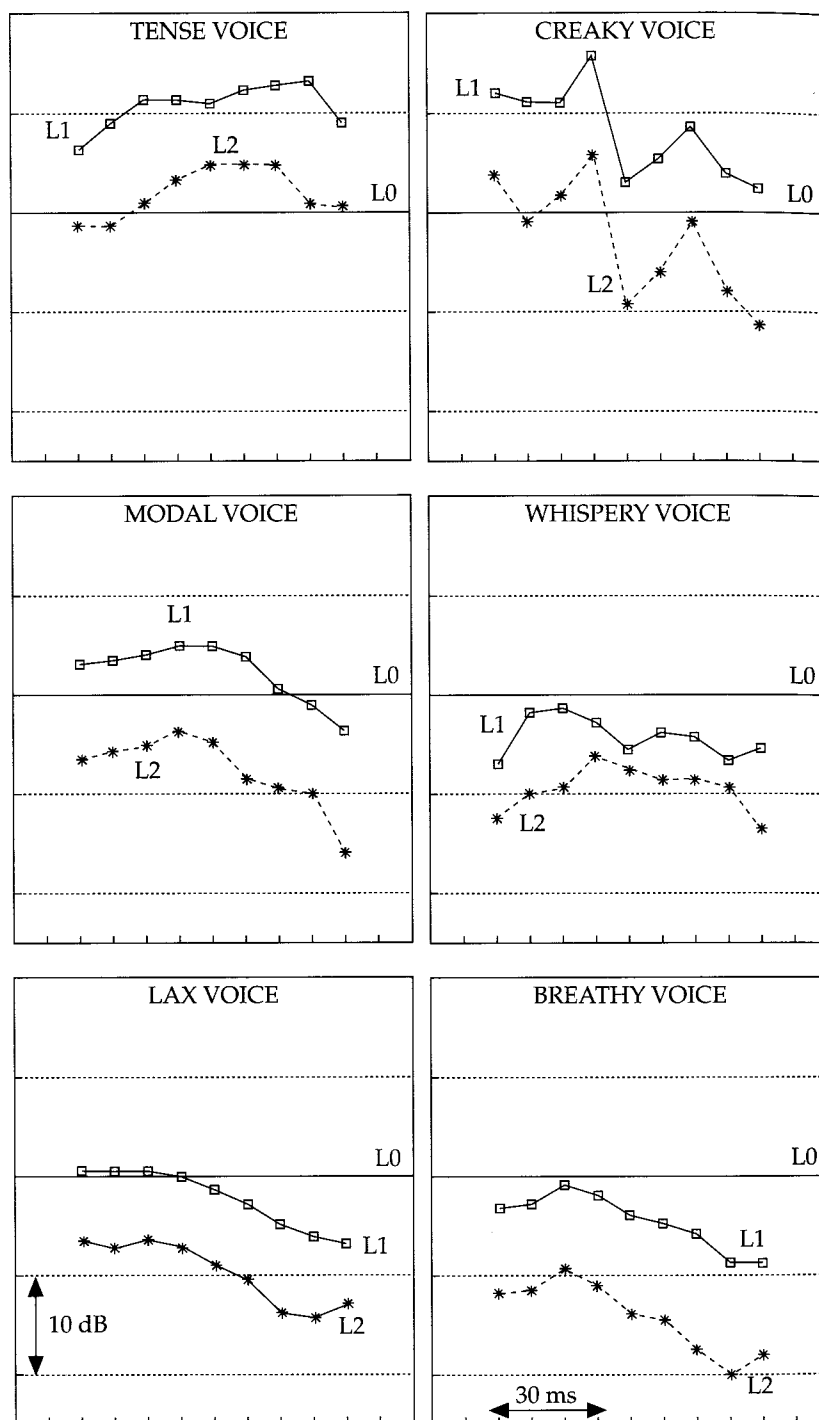


Figure 14.10 F1 and F2 levels relative to the level of H1 (L_1-L_0 and L_2-L_0) are shown for tense, modal, lax, breathy, whispery and creaky voice qualities for a 90 ms interval in /straiks/.

slope for each quality deviates from the "ideal" source (i.e., -6 dB per octave for the differentiated glottal flow). To achieve this, the spectra were "flattened" by adding 6 dB per octave relative to the amplitude of the first harmonic (L_0). The source spectrum was then divided into four frequency bands: 0-1, 1-2, 2-3, 3-4 kHz. (For the vowel in question there is one formant in each band. Harmonics above 4 kHz were not measured.) The average of the normalised (linear) amplitudes of all harmonics within a frequency band was then calculated and plotted relative to L_0 . This average represents the deviation from the "ideal" source slope, indicated by the horizontal line at 0 dB in Figure 14.12. For further details, see Gobl (1989) and Gobl and Ní Chasaide (1992).

Modal voice is the neutral mode of phonation to which other voice qualities are compared, and "which phonetic theory assumes takes place in ordinary voicing, when no specific feature is explicitly changed or added" (Laver, 1980: 95). For this quality, adductive tension, medial compression and longitudinal tension are thought to be moderate. Both the ligamental and the cartilaginous part of the glottis vibrate as a single unit. The vocal fold vibration is further described by Laver as regularly periodic and efficient, with full glottal closure and thus, without audible glottal frication noise. Some recent studies have, however, shown that incomplete glottal closure may be very common even in what is perceived as modal voice (see, for instance, Södersten, 1994) and particularly in female speech.

The slope of the source spectrum for modal voice in Figure 14.12 is somewhat greater than the "ideal" case. Nevertheless, it is in relative terms a fairly efficient mode of phonation. It is important to remember that within utterances spoken with modal (or indeed any) voice quality, there may be considerable dynamic variation of the source (see, for example, Gobl, 1988). In certain environments there may be considerable convergence of modal and breathy/whispery voice, as can be seen in the few periods preceding the voiceless consonant /k/ in Figure 14.9. This is a contextual effect which appears to affect all the voice qualities looked at, and is discussed in detail in Ní Chasaide and Gobl (1993).

Breathy voice is thought to involve minimal adductive tension, weak medial compression and low longitudinal tension. The vocal folds vibrate very inefficiently and they never come fully together. Thus, there is a considerable constant glottal leakage with some audible frication noise. The high dynamic leakage of this voice quality is evidenced by high RA values. Consequently, FA is much lower than for modal voice, particularly in the stressed vowel of *babber*, where we find a value of 500 Hz for breathy voice compared to 1,500 Hz for modal. The glottal pulse is also more symmetrical (high RK) for breathy voice, and has a high open quotient (OQ). Together, these suggest a high rate of airflow through the vocal folds, as would be expected for this voice quality, and this is indeed what our calculated UP (peak glottal flow) values show. This would yield a relative boosting of the lowest harmonic, a spectral feature which has been widely reported for this voice quality. The consequent sharp slope in the lower end of the source spectrum can be seen clearly in Figure

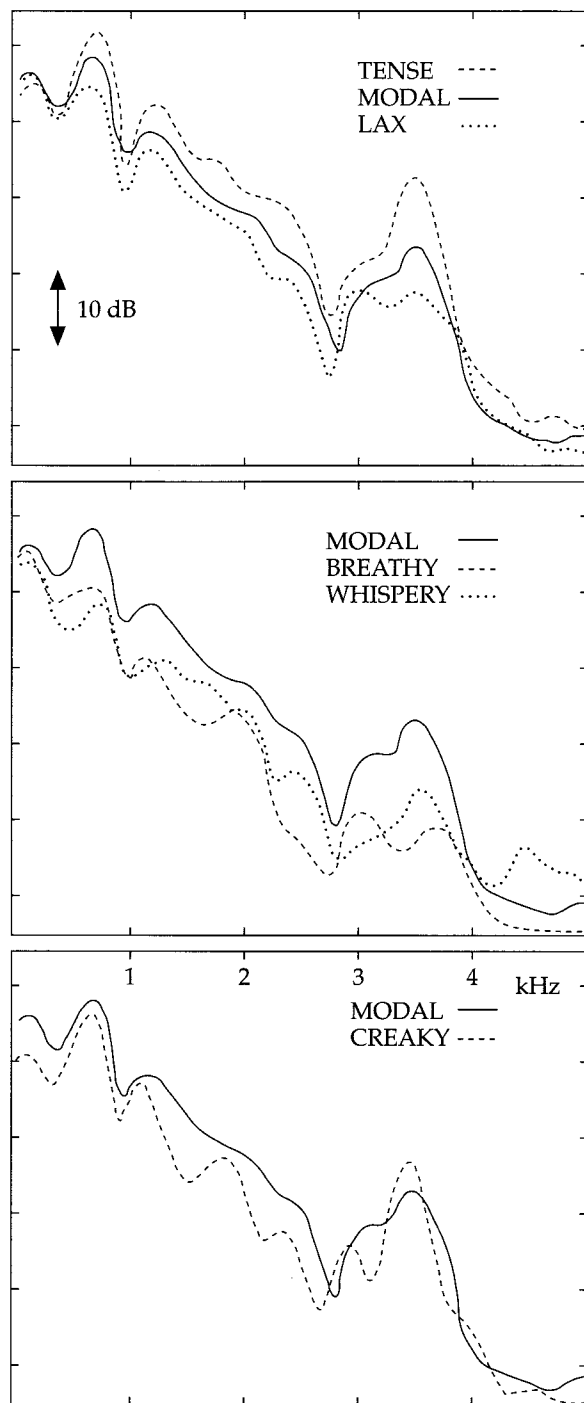


Figure 14.11 Average spectra for the voiced portion of /straiks/ shown for tense, modal, lax, breathy, whispery and creaky voice qualities.

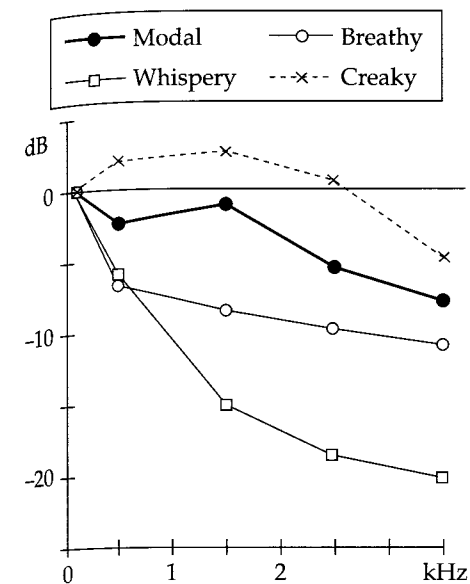


Figure 14.12 Schematic source spectra for four voice qualities in *babber* showing within four frequency bands the average deviation from a constant -12 dB/octave slope.

14.12. In the speech output signal, we see in Figure 14.10 the clear dominance of H1 for breathy voice as compared to the dominance of F1 in modal voice. This particular spectral measure captures not only the relative boosting of L_0 , but also the high degree of damping of F1, which would be expected for the breathy voice quality, where the vocal folds are abducted. These effects can also be observed in the average spectra of Figure 14.11.

Whispery voice is characterised by low adductive tension, moderate to high medial compression and moderate longitudinal tension. As a consequence, there is a triangular opening of the cartilaginous glottis, whose size varies with the degree of medial compression. In weak whisper the medial compression is moderate and the opening may include a part of the ligamental glottis as well as the cartilaginous. Whisper with increasingly higher intensity has increasingly higher medial compression and smaller glottal opening, until only the cartilaginous glottis is open. Laryngeal vibration is assumed to be confined to that portion of the ligamental glottis which is adducted, and the whispery component to the triangular opening between the arytenoids. It is very inefficient and there is a considerable degree of audible frication noise.

As pointed out by Laver (1980: 133–134), whispery and breathy voice form an auditory continuum with no clear borderline between them. In auditory terms they would be distinguished in terms of the relative dominance of the periodic and noise components: in breathy voice, the periodic component is

dominant, whereas in whispery voice the noise component would be relatively greater (see further comments on these two qualities in Section 4).

The source measurements for whispery voice are fairly similar to those of breathy voice, being in many cases more extreme deviations from modal values. Whispery voice differs mainly from breathy voice in having a lower RK and OQ, showing a more skewed glottal flow pulse, with a relatively shorter closing branch. The calculated UP values were also noticeably less than for breathy voice. As UP is highly correlated with H1 level (L_0), we find that there is less boosting of the fundamental than for breathy voice. Note that this difference does not show up in the source spectra of Figure 14.12, as these have been normalised to L_0 . In terms of the entire source spectrum, however, the fundamental component remains very dominant, as the source spectrum has an even greater slope than does breathy voice. Probably for this reason, the measurements of L_1 and L_2 relative to L_0 in the output signal (Figure 14.10) do not look very different to those of breathy voice. Of course, bandwidth differences also affect the levels of the spectral output.

Creaky voice is thought to involve high adductive tension and medial compression, but little longitudinal tension. Pitch has been observed to be extremely low, and would appear to be controlled by aerodynamic factors and not by varying the longitudinal tension, as in the other qualities. The f_0 and amplitude of consecutive glottal pulses is further known to be very irregular. Because of the high adductive tension, only the ligamental part of the vocal folds is vibrating. The folds are relatively thick and compressed, and the ventricular folds may also be somewhat adducted, so that their inferior surfaces come in contact with the superior surfaces of the true folds. This would thus create an even thicker vibrating structure. The mean airflow rate has been observed to be very low.

Although every voice quality varies dynamically in the course of an utterance, creaky voice is particularly variable. Creakiness, in the sense of irregularity of successive glottal pulses appears intermittently. It did not show up in the stressed word *babber*, but did in the relatively unstressed word *strikes* (see right hand panel in Figure 14.9 where the speech waveform for creaky voice is also shown). Here, there is an alternation of two very different types of glottal pulse. One has a reasonably high EE (which however is still lower than for modal or lax voice), a very low RA, and consequently a high FA, suggesting a fairly instantaneous glottal closure and strong higher frequencies. The other type of pulse shows rather opposite tendencies: EE is very low and RA is very high, and these should have rather different effects on the acoustic spectrum. Both types of pulse are characterised by a low OQ, a low RK and a relatively high RG (values are more extreme for the first type of pulse). In the stressed word *babber*, source values were not unlike those of the strong pulse described above, but differed mainly in that EE was considerably higher. The short open phase found generally for this voice quality correlates well with the known low airflow rate observed for creaky voice, and would have the consequence of reducing the levels of the lower harmonics relative to the rest of the spectrum.

This effect can be seen clearly in the schematic source spectrum in Figure 14.12 and in the average spectra of the speech output signal in Figure 14.11. Similarly in Figure 14.10, we find a very dominant F1 relative to the H1 level. Note that the pulse-to-pulse variation is to some extent smoothed out by the 30 ms Hamming window used in the FFT calculations for this figure. The relatively long closed phase of the glottal pulse should also contribute to narrow formant bandwidths, and this can probably also be inferred from the sharp peaks in the average spectra of Figure 14.11. The relatively high RG observed for this voice quality would also tend to boost the region of H2 relative to H1, a feature which has been noted in the literature as a characteristic of creaky voice.

Tense voice involves a higher degree of tension in the entire vocal tract as compared to the neutral setting. At the laryngeal level, the two parameters which show a particular increase in tension are adductive tension and medial compression. This would correspond to the term pressed phonation used by many authors. The increased muscular tension associated with tense voice is likely to affect the respiratory system (resulting in a raised subglottal pressure) as well as the supralaryngeal tract. Acoustically, this voice quality (measured only for *strikes*) exhibited a very low RA showing a very sharp full closure of the vocal folds. The related frequency measure FA was higher than for all the other qualities looked at in this context, and so one would expect strong higher frequencies. The glottal pulse was also more skewed for tense voice (lowest RK values), had a small open quotient (OQ) and a high RG. The picture of a highly skewed pulse with instantaneous closure and a long closed phase accords well with the physiological description of high adductive tension and medial compression. These source values suggest that the lower harmonics should be attenuated relative to higher frequencies. This effect shows up clearly in Figure 14.11, and we observe in Figure 14.10 that F1 dominates the spectrum. As with creaky voice, the high RG will affect the ratio of H1 to H2 by relatively boosting the latter. For this speaker, there is considerable similarity between tense and creaky voice parameters, if one ignores the pulse-to-pulse variability sometimes found for the latter.

Lax voice involves a lesser degree of tension in the entire vocal tract and typically tends to have opposite characteristics to tense voice. At the laryngeal level there is a reduced degree of adductive tension and medial compression. Phonation may therefore be similar to breathy voice, sounding softer and lower pitched than modal voice: however, the amount of change in these tension parameters is often less than for breathy voice. Source measurements show parameter values similar to those of breathy voice. Excepting the extremely low RG values, differences found between lax and breathy voice suggest that, of the two, lax voice is the closer to modal voice.

For all voice qualities, the reader should bear in mind that they are not fixed entities. Non-modal qualities may occur to a greater or lesser degree, i.e., may be further from or closer to modal voice. Voice qualities can also be of a compound type, as for example in whispery creaky voice (for more on this, see Laver, 1980).

As was mentioned for modal voice, there may be considerable dynamic variation even within utterances considered to have been spoken with a single voice quality. The same point holds across voice qualities: a non-modal quality may be nearer or closer to the modal depending on context. For example, our data suggest that differences were greater in the stressed than in the relatively unstressed syllable. Another example is the creakiness in creaky voice, which is intermittent and appears to be associated with particular environments. As far as we can tell at this point, it seems unlikely that a good implementation of a voice quality change in synthesis will be achieved by a single set of transformations, but will require context-sensitive rules.

4 Determinants of voice source variation

Individuals differ in the voice quality that they may habitually use. Even within the speech of the individual there are considerable dynamic changes in the voice. Some aspects of source variation are within the speaker's control, and may be linked to the linguistic content of utterances or to the speaker's paralinguistic signalling intent. Some source differences serve a sociolinguistic function insofar as social, regional or linguistic groups may tend towards frequent use of particular voice qualities. But beyond linguistic, paralinguistic and sociolinguistic influences, individuals' voices are shaped by many factors, some of which are not within their control, such as the physical properties of their vocal apparatus. This section presents an overview of some of these functions of voice quality variation.

4.1 Linguistically determined variation of the source

Variations in the voice source may be associated with segmental or suprasegmental elements of the linguistic code. The voice qualities most frequently mentioned as partaking in linguistic contrasts are modal voice, creaky voice (also called laryngealised) and breathy voice (also called murmured, or in the case of consonants, voiced aspirated). For the latter quality Laver (1980) suggests, that in terms of his classifications, whispery rather than breathy voice may be involved. However, as there is considerable variability in the realisations of breathy voiced segments, it is likely that they lie at different points on the breathy to whispery voice continuum (see below and also Section 3), and they will simply be referred to here as breathy voiced. Other more extreme voice qualities also occur, such as the very harsh "growl" described by Rose (1989) for the Zhenhai variety of Wu Chinese. This last would appear to involve the ventricular folds as well as epiglottalisation, and would sound like a pathological voice quality to an English speaker's ear. The terms *tense* and *lax* have also been used to describe contrasts based on voice quality, but as is clear from Maddieson and Ladefoged (1985), the terms can be misleading,

and likely to be used in a phonological rather than in a phonetically accurate sense. Thus, the authors speculate, *tense* might signify modal voice in one language (e.g., Wa, a Mon-Khmer language of Southwest China, where it contrasts with a *lax* quality which may be phonetically breathy voiced) but creaky voice with raised larynx in another language (such as Yi, also spoken in Southwest China, where the contrasting *lax* quality would appear to be modal voice).

4.1.1 Source variation associated with segmental contrasts The contrastive use of voice quality for vowels or consonants is fairly common in South East Asian, South African, and Native American Languages, and these have been the focus of a number of studies carried out at UCLA. Although both vowels and consonants may employ voice quality contrasts in a given language, Ladefoged (1982) points out that it is very rare to find contrasts at more than one place in a syllable. The term register is often used to describe voice quality contrasts, but is a phonological cover term, and subsumes any other phonetic features (such as vowel quality, vowel duration and small or exaggerated f_0 differences) which are often associated with such contrasts in particular languages. As a practical consequence of the facilities available (especially in field work) most investigations of linguistically contrastive source effects have tended to concentrate on spectral measurements based on the speech waveform, such as those outlined in Section 2.5. However, for the reasons mentioned there, using these kinds of measurements to characterise an essential voice quality difference can be problematic where there are concomitant differences in formant frequencies or in f_0 .

For vowels, contrasts are typically of a two way kind, e.g., the breathy voiced vs. modal voiced vowels in Gujerati (for instrumental descriptions, see Fischer-Jørgensen, 1967; Bickley, 1982). A more unusual case is the six way opposition described for !Xóǀ, a Khoisan language spoken by Bushmen in Southern Africa (Ladefoged, 1982; Traill, 1985). This language distinguishes modal and breathy voiced vowels. Each of these qualities can occur with additional creakiness, to give creaky voice and breathy creaky voice. (We should note here that the latter would be termed whispery creaky voice in Laver's system, where the combination of breathiness and creakiness is regarded as an impossible combination within his definitions.) Finally, both modal and breathy voiced vowels can occur with an additional strident quality. Strident here would appear to be fairly similar to the "growl" of Wu mentioned above (see discussion in Rose, 1989).

In the case of consonants, voice quality contrasts have been reported for stops, nasals, liquids and approximants. A modal vs. breathy voice contrast of nasals is reported for Tsonga and for other Bantu languages of Moçambique and South Africa (Traill and Jackson, 1987). Breathily voiced stops are characteristic of many Indo-Aryan languages including Nepali, Gujerati and Hindi, where they may contrast with (modal) voiced, voiceless unaspirated and voiceless aspirated stops. For a description of the contrasts of Hindi, see Dixit (1987).

Note that where consonants are described as having contrasting voice qualities, the acoustic manifestation often appears to be primarily located at the onset or offset of the vowel. The acoustic effect in these cases is attributed to laryngeal differences associated with the consonant, but affecting the initial or final portion of the vowel. Dixit (1987) is at pains to point out that although glottal abduction in the voiced aspirated stop occurs about half way through the closure, the stop interval should not itself be regarded as different from that of the normally voiced stop. In a similar vein, Traill and Jackson (1987) show for the breathy voiced nasals of Tsonga, that the acoustic effects are mostly associated with the vowel onset, and that vocal fold abduction for the breathy voiced nasal begins during the nasal consonant. The phonological domain of particular voice quality contrasts is not always clearcut, and there may be reasons for preferring to treat a particular contrast as concerning the vowel, the consonant, or the syllable. In the Jalapa de Diaz dialect of Mazatec, Mexico, contrasts involving modal voice, breathy voice and creaky voice qualities are found. Two possible analyses are suggested by Kirk, Ladefoged and Ladefoged (1984). One is to view the language as having a three way contrast at the level of the syllable. Alternatively, it can be viewed as having an opposition of modal and breathy voiced vowels, and of modal and creaky voiced consonants. For a discussion of the domain of the voice quality contrast in the Wu dialect of Chinese, see Jianfen and Maddieson (1989).

As was pointed out in Section 3, voice qualities differ from each other in a scalar rather than in a discrete way. For any given voice quality, e.g., breathy voice, it will occur to differing degrees across languages or even for different speakers of one language/dialect. Maddieson and Ladefoged (1985) point out that the ratio of breathiness to voicing for breathy voiced segments is greater in Hindi than in Yi and greater in Yi than in Tsonga. Concerning cross-speaker variation, Ladefoged hypothesises that although speakers of a single dialect may vary in the degree of breathiness they employ for a breathy voiced segment, all speakers produce the contrast by using different degrees along a continuum. Thus, for a speaker with an intrinsically breathy voice, a modal vs. breathy voiced contrast would be achieved by increasing the breathiness where relevant. It is therefore not surprising that attempts to find (from data illustrating linguistic contrasts) measures that allow classification of voice quality in absolute terms have not met with success.

Discussion so far has only been of cases where voice quality is considered to have a contrastive function, i.e., is taken to be the main phonetic feature on which a phonological contrast rests. Differences can also occur which would not be considered (phonologically) distinctive. One such case is described for Swedish in Ní Chasaide and Gobl (1993), where the voiced/voiceless nature of an intervocalic stop can greatly influence the quality of the offset of a preceding stressed vowel. This type of effect was not found for comparable French and German data. And although the source quality difference observed in Swedish is likely to contribute perceptually to the stop contrast, it would not constitute the primary cue.

Classes of consonants differing in manner of articulation would appear to have intrinsically different voice source characteristics. For a description of differences amongst voiced stops, fricatives, nasals and laterals, see Ní Chasaide, Gobl and Monahan (1993). This type of variation most likely reflects the supraglottal configuration and resulting aerodynamic conditions pertaining to the different classes of segments, and as such, one would expect it to be universal. Although it may be relatively uninteresting to the linguist concerned with the contrastive material on which phonological systems are based, we do need to know more about it, both as baseline material for descriptive analyses and for improved speech synthesis.

4.1.2 Source variation associated with suprasegmental contrasts
Suprasegmental phenomena such as intonation, tone and stress have been extensively studied, though primarily in terms of f_0 (and to a lesser degree amplitude) variation. These relatively well understood aspects were explicitly excluded from the present coverage (see Nootboom, PROSODY OF SPEECH: MELODY AND RHYTHM). However, it is worth noting that many other aspects of the glottal pulse will vary as a function of f_0 and voice level. But voice source variation plays a role in the suprasegmental systems of languages quite beyond that which is strictly dependent on f_0 and voice level.

Among tonal languages it is not unusual to find that a particular voice quality is associated with specific tones. One of the seven tones in Hmong, a Sino-Tibetan language, is described as having a breathy voice quality (Huffman, 1987). The yin and yang tones of the Wu dialect of Chinese described by Jianfen and Maddieson (1989) are also characterised by specific voice qualities. The yang tones differ from the yin in that they employ breathy phonation and begin with a lower f_0 onset. See also Rose (1989) on the rather different realisations of yin/yang tones in the Zhenhai variety of Wu.

As f_0 differences tend to be associated with different voice qualities in any case, it is hardly surprising to find register correlates of tonal contrasts and vice versa. Despite such correlations, many authors are at pains to point out that f_0 and voice quality are separately controllable, and that variation in one does not allow prediction of variation in the other. The link between voice quality and f_0 may have historical implications, and the likelihood of tonal contrasts having evolved from earlier voice quality contrasts has been discussed by Maddieson and Hess (1987), Jianfen and Maddieson (1989), and by Rose (1989). Not surprisingly, there are cases where contrasts in specific languages are open to competing analysis as involving primarily register or tone. See for example the lively debate concerning the so-called register contrast of Mon (Lee, 1983; Diffloth, 1985; Thongkum, 1987). Maddieson and Hess (1987) suggest that the six tones of Lisu should be interpreted as a four way tonal contrast, with a register contrast in two of the tones. Rose (1989) argues for an interpretation of the yin/yang difference in tones of the Zhenhai dialect of Wu as a register contrast, which interacts with a three way tonal contrast.

Variation in source parameters other than f_0 may also be relevant to the

description of intonation. Pierrehumbert (1989) carried out a pilot study on the interaction of intonational and voice source variables. Her findings show that the glottal pulse for high tones (in pitch accents) has a greater open quotient (OQ) but a higher degree of skew (lower RK) than for low tones. These results do not hold, however, across utterances produced at different voice levels: a higher voice level results in a higher f_0 but a reduced OQ. She points out that a better understanding of the interaction of pitch and voice source variables will ultimately be required for the adequate phonetic realisation of intonation in synthetic speech. Fant and Kruckenberg (1989) have demonstrated how creaky voice is used as a phrase boundary marker for speakers of Swedish. This can alternate with the insertion of pauses in read texts. In a similar vein, Laver (1980) suggests that creaky voice in conjunction with a low falling intonation is used by speakers of Received Pronunciation of English to regulate turn taking, by signalling that the speaker's contribution is completed.

There may also be correlates of stress in particular languages. Gobl (1988) has described the source characteristics of a word in focal, prefocal and postfocal position of an utterance. The dynamic range of the source excitation (EE) was considerably greater when the word was in focal position than in the other environments, being stronger for the vowel and weaker for the surrounding voiced consonants. This is effectively an enhancement of the vowel-consonant distinction in the stressed syllable.

4.2 *Paralinguistic aspects of voice source variation*

Whatever the habitually favoured voice quality of a speaker, temporary shifts are a means of signalling the speaker's mood, emotion and attitude to the listener or to the content of the message. As paralinguistic communication is a voluntary, convention-bound system of affective signalling, it can be readily used to mislead the listener. For example, speakers can adopt a tone of voice that signals interest in a topic which they find truly boring, or can feign indifference to cover a very real but controllable anger or other emotion.

A number of impressionistic observations can be found in the literature, concerning the paralinguistic significance of different voice qualities. Laver (1980) suggests that for speakers of English, breathy voice is associated with intimacy, whispery voice with confidentiality, and harsh voice with anger. Although the communicative function of certain voice qualities may tend to be universal, in many cases it is culturally determined. Thus for example, whereas sustained creaky voice is used by certain speakers of English to signal bored resignation, the same voice quality is used in Tzeltal (a Mayan language) to express commiseration or complaint (see Laver, 1980: 126).

Studies on the acoustic correlates of emotional states are summarised by Kappas, Hess, and Scherer (1991) for the following emotions: boredom-indifference, displeasure-disgust, irritation-cold anger, rage-hot anger, sadness-dejection, worry-anxiety, fear-terror, and joy-elation. These studies have tended

to concentrate on the acoustic parameters of f_0 (changes in the mean value, range, contour type, variability) and intensity (changes in the mean value). Increased mean values and increased variability of f_0 and intensity were found for many of these emotions, and it is clear that many more fine-grained types of measurements will be needed if we are to differentiate, as listeners do, the various emotions that may be acoustically signalled. As pointed out in this and other studies, research in this area faces major methodological difficulties, particularly in eliciting appropriate speech samples. For example, many studies have employed actors' portrayals of emotions and the considerable inter-subject variability found in these studies suggests that they may vary not only in their style of acting, but also in their ability to simulate emotions. Furthermore, actors' portrayals may differ from spontaneous, naturally occurring emotions. When dealing with extremes of emotions it is likely that voice quality changes are involuntary, having their origin in the physiological changes brought about by the emotional state itself. As such, they are extralinguistic and presumably universal, and do not belong to the conventional learnt system of affective signalling.

Listeners' reactions to manipulated synthetic speech may prove an additional fruitful research method in this difficult field, particularly as certain speech synthesisers now permit control of many important source parameters and not simply of f_0 and intensity. In this type of experimentation it is important to take account of the linguistic content of test utterances. The linguistic and paralinguistic functions interact in complex ways, and an inappropriate choice of linguistic elements might yield a bias in the attribution of affective colouring to voice parameters. The interaction of linguistic and paralinguistic aspects of communication has been looked at by Scherer, Ladd and Silverman (1984), and by Ladd, Silverman, Tolkmitt, Bergmann and Scherer (1985).

4.3 *Sociolinguistic function of voice source differences*

Voice quality may also have a sociolinguistic function, serving to differentiate among linguistic, regional and social groups. Supralaryngeal as well as phonatory features may be used to this end. As anyone who has taught a foreign language will attest, cross-language differences in voice quality are an important aspect of a convincing accent, but difficult to teach as they are virtually never described in the linguistic or applied linguistic literature. This can lead to cultural misperceptions, as the native speaker is likely to interpret the foreigner's voice quality in terms of his/her own paralinguistic system for affect or attitude signalling.

Within a particular language or dialect group, voice quality features may signal social subgroups. In Edinburgh English a greater incidence of creaky voice is associated with a higher social status, whereas whispery and harsh qualities are linked to a lower social status (Esling, 1978). In Norwich, working and middle class accents are differentiated on the basis of habitual phonatory

and supralaryngeal settings (Trudgill, 1974). Other social groupings may also tend towards different voice qualities. Rose (1989) suggests that the extremely harsh "growl" mode of phonation found in the Zhenhai dialect of Wu differs in terms of the sex and age of the speaker, being least harsh for women, and most harsh for old men. In investigating differences which are correlated with sex and age, it is of course important to distinguish between truly sociolinguistic markers and differences which are due to laryngeal anatomy.

4.4 Extralinguistic determinants of the voice source

There are other factors which determine the quality of the voice, many of them beyond the control of the speaker. Differences in the size, shape and muscular tone of the laryngeal structures play a major role. The voices of men, women and children reflect mostly anatomical differences, although intrinsic, anatomy-based features may be enhanced or reduced depending on the socio-cultural context. For example, women working and competing in a male environment may choose to adopt a mode of phonation more similar to that of the male. As is shown by the typically poor quality of synthesised women's and children's voices, our understanding here lags behind that of the male voice. For some descriptions of the female voice see Karlsson (1992a), Holmberg et al. (1988) and for child vs. adult male differences, see Gobl (1988).

Voice quality is also affected by the individual's physical and mental health. There has been considerable study of certain acoustic correlates (mostly f_0 and intensity) of psychiatric illnesses such as depression and schizophrenia. For a summary of studies on the vocal indicators of depression, see Scherer (1987). Pathologies of the laryngeal structures also affect vocal quality and many studies of the voice have been medically motivated. Figure 14.13 illustrates a number of source parameters (f_0 , EE, RA, RK) for a female speaker with vocal fold nodules, as compared to a normal speaker matched for age, sex and accent. The pathology appeared to be particularly associated with the initiation of voicing, where the vibratory cycle was grossly perturbed in a number of ways. f_0 was very high, the pulse excitation was weak (low EE). There tended to be a considerable degree of dynamic leakage (high RA) and the pulse shape was generally more symmetrical (high RK). Probably as important as the actual values was the unstable nature of the glottal pulse during this initial interval, evident from the considerable pulse-to-pulse fluctuations for all parameters. At a certain point (somewhere between the 10th and 20th glottal cycle) the phonatory pattern switched abruptly. Pitch dropped by about an octave to normal values and the other source parameters indicated a more normal and stable glottal pulse. For further details of this study, see Kane and Ní Chasaide (1992).

Over and above the linguistic and non-linguistic factors mentioned so far, voice quality also carries uniquely personal information and serves an important function in allowing us to identify speakers and tell them apart (see Nolan, SPEAKER RECOGNITION AND FORENSIC PHONETICS).

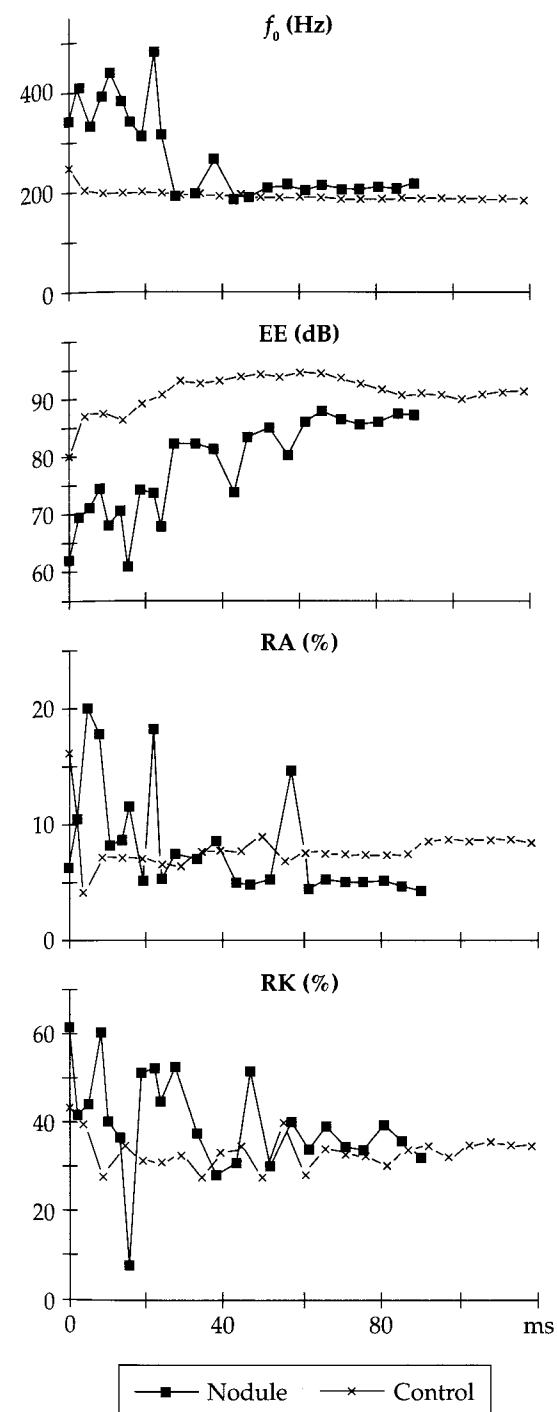


Figure 14.13 Values of f_0 , EE, RA and RK compared for a speaker with vocal fold nodules and a normal speaker, from the initiation of voicing (following stop release) in the nonsense word *baa*.

5 Possible directions of future research

The human voice has evolved as a vehicle for conveying many different types of information, and human listeners have developed the ability to detect very small, very subtle voice quality changes, and interpret their function. In spite of improvements in recent years in the techniques for describing and modeling source variation, our abilities lag far behind what the human ear can effortlessly do. At this stage of play, we can but appreciate the scale and complexity of the research that will be needed to gain a full understanding of how the voice is used in speech and language. Most of the studies to date have been very limited, either in the quantity of data analysed, or in the kinds of source measures made. Advances in the field have been particularly hampered by the lack of availability of suitable analysis tools. The manual interactive techniques outlined in Section 2, permit a fine grained analysis of the source, but because of their labour-intensive nature, are not suitable for the large scale studies that would be ideally needed for progress in this field.

The research agenda must therefore be directed not only at descriptive studies, but also at devising new techniques or enhancing current ones to automate the acquisition of accurate source data. Work aimed at automating the analysis of the source without sacrificing accuracy is currently being explored by Gobl, Monahan, Fitzpatrick and Ní Chasaide (1994). Another approach currently being pursued by Fant (1994) is to exploit the likely correlations among source parameters (explained in Section 2.4). This will of course not involve the same degree of accuracy but would provide useful first estimates and could be used for larger quantities of data. As our state of knowledge concerning correlations between source parameters advances, we may be able to infer more of the fine detail from the more gross measures.

As in other areas of speech research, acoustic analysis must be supplemented by physiological experiments, to elucidate underlying production processes. The technique for high-speed digital imaging of vocal fold vibration, described by Hirose, *INVESTIGATING THE PHYSIOLOGY OF LARYNGEAL STRUCTURES*, offers exciting prospects in this area. Research in this field should also soon show the benefits of rigorous perceptual testing, now that speech synthesisers are increasingly incorporating more sophisticated source models, permitting separate control of many important parameters.

An improved understanding of the voice source and of how it varies in speech would open the door to many applications. The most immediate application of providing a more natural and potentially variable voice in speech synthesis would greatly enhance the acceptability of synthesis-based devices. In the past, an inappropriate voice quality has often led to a rejection of these devices, even by those who would most benefit from them, e.g., the vocally handicapped (see also Carlson and Granström, *SPEECH SYNTHESIS*). One can envisage at some future date the possibility of customised voices in handicap-aids, designed to match the original voice quality of the user. When reasonably

accurate automatic analysis procedures become available, one can envisage many other applications in areas such as speaker recognition and verification. And with an increased understanding of the range and types of variation found in normal and pathological voices, such techniques might also facilitate an acoustic screening procedure for voice disorders.

NOTE

The authors would like to acknowledge that some of their work discussed in this chapter was carried out within the framework of the ESPRIT/Basic Research Actions ACCOR I, ACCOR II,

SPEECH MAPS and VOX. We are also grateful to Peter Monahan and Liam Fitzpatrick for their assistance with some of the figures presented here.