CHAPTER 15

Voice and Emotional States

Gudrun Klasmeyer and Walter F. Sendlmeier

In natural communication situations, emotional arousal is a quite complex phenomenon. Scherer (1986) proposed a theoretical model of vocal affect expression that is generally accepted. In this model emotion is seen as a process, not as a steady state of the organism. The process includes such components as physiological arousal and expression and feeling as a response to an evaluation of significant events in the environment. The organism's information processing subsystem continously scans external and internal stimulus inputs and performs a series of stimulus evaluation checks that result in the organism's emotional state. This emotional state can be labeled with emotional terms provided by natural languages and referred to as "discrete emotions." In spite of these discrete labels, the emotional arousal, itself, is gradual, can change abruptly, and can also be a mixture of different discrete emotions.

In a vocalization process, the physiological arousal due to an emotion appears to be an involuntary force. A vocalization can further be influenced by voluntary effects in the service of affect control or self-presentation.

Systematic studies of acoustic voice qualities in different emotional states of a speaker have to deal with some problems. Field recordings of emotional utterances usually lack the good quality essential for the analysis of acoustic voice quality parameters; whereas under laboratory conditions, it is quite difficult to sucessfully induce different emotions. A solution to this problem is to record emotional speech produced by actors under laboratory conditions in combination with a perception experiment with naive listeners to evaluate the naturalness and recognizability of the emotions. Drama actors tend to portray emotions in a way that the emotional content can be recognized from a distance, which results in an improper use of the voice source, although the emotional content might be recognizable from the intonation pattern and the speaking rate. Therefore, the actors have to use a special technique: They imagine an emotional situation until they really "feel" the emotional arousal, before the utterances are recorded. The naive listeners have to be instructed to evaluate utterances that sound like "theater clichés" of the emotions as unnatural.

The acoustic analysis of voices is not an easy task, because there is a lot of parallel information in fluent speech. Besides semantic information, spoken language contains evidential information conveyed by signs in speech that act as attributive markers (Laver 1994). These are used by a listener as the basis for attributing personal characteristics to the speaker. The attributes of the speaker fall into three groups:

- Physical markers that indicate characteristics such as sex, age, physique, and state of health;
- Social markers that indicate characteristics such as regional affiliation, social and educational status, occupation, and social role;
- Psychological markers that indicate characteristics of personality and affective state or mood.

The correlation between these attributive markers and acoustic parameters is rather complicated. Laver states that physical characteristics lie in a speaker's voice quality, with social markers including such features as accent and choice of vocabulary. Psychological markers are often taken to reside in a speaker's tone of voice (Laver, 1994). But, given the human vocal apparatus, it is very unlikely that tone of voice should change under emotional arousal without affecting the speaker's general voice quality or type of phonation. It is more realistic to assume that all acoustic parameters that can be used to characterize a speaker's personal voice quality might be influenced by the speaker's mood or affective state.

A listener's attribution of speaker characteristics is usually based on the general impression of the utterance. Attributions based on single acoustic parameters are ambiguous. A high "jitter" value, for example, could be a physical marker for a speaker's personal voice quality or state of health, but it could also be a psychological marker for fear.

This chapter proposes a general approach toward the objective measurement of different voice qualities in fluent speech. Initially, conclusions about typical signal characteristics of the transglottal airflow are drawn from the production strategies of varying types of phonation. The signal characteristics are discussed with regard to the question of how voice quality and type of phonation can be measured in the acoustic speech signal. In natural communication situations, voice quality is often a combination of different types of phonation. This is one reason why several acoustic parameters are required for an adequate description of voice quality. To deal with the fact that some voice qualities can hardly be regarded as quasistationary signals and, hence, are difficult to analyze using common algorithms for stationary signals, some of the introduced parameters are not orthogonal, but represent different approaches to measuring similar phenomena in the frequency and time domains. As mentioned, voice quality can be used by listeners as a basis on which to attribute speaker characteristics.

1. ANATOMY AND CONTROL MECHANISMS OF THE VOICE SOURCE

In human speech production, the breathing organs produce an airflow that is not audible by itself. In the laryngeal system, this airflow is modified into an audible sound wave. The sound wave propagates through the supraglottal cavities, where it is subject to further modifications by the articulators before it is radiated from the mouth as an acoustic speech signal. The major results of all laryngeal control mechanisms are changes in the opening section of the glottis as well as the tension and oscillation characteristics of the vocal folds. Three forces are regarded as being most important for the control of voice quality and type of phonation. These forces are (1) the longitudinal tension of the vocal folds, which is achieved by muscular tension in the musculus vocalis and the cricothyroid muscles; (2) the adductive tension of the ligamental and cartilaginous glottis, which is achieved by tension of the interarytenoid muscles; and (3) the medial compression that closes the ligamental glottis as a secondary effect of a force on the arytenoid cartilages, which is caused by tension of the lateral cricoarytenoid muscles in collaboration with tension in the lateral part of the thyroarytenoid muscles. These forces of longitudinal tension, medial compression and adductive tension are illustrated in Figure 15–1, which shows a cross-section of the human larynx.

It is obvious that emotional arousal can have an influence on the control mechanisms and therefore change the audible voice quality.

In phonetic practice the categories breathy phonation, whispery phonation, modal voice, creak (which is also called glottal fry or vocal fry), and falsetto proposed by Laver (in Chapter 4) are used to label voices. In the next section, acoustic characteristics of these types of phonation are discussed with regard to the question of how the voice quality can be measured objectively.

2. SIGNAL CHARACTERISTICS OF ACOUSTIC SPEECH SIGNALS

The distinction between voiced and voiceless sound segments has already been adequately addressed and solved in the research literature. This chapter focuses on the objective description of voiced sounds, such as breathy voice, whispery voice, modal voice, creaky, and falsetto in terms of acoustic parameters. The listed types of phonation can be determined in the acoustic speech signals by analysis of the spectral distribution of harmonic and noise components in the frequency domain and by parameters derived from the time signal. Among these parameters in the time domain, vocal effort, voicing irregularities, and the speed of changes in the fundamental frequency are important criteria by which voice quality in acoustic speech signals can be characterized.

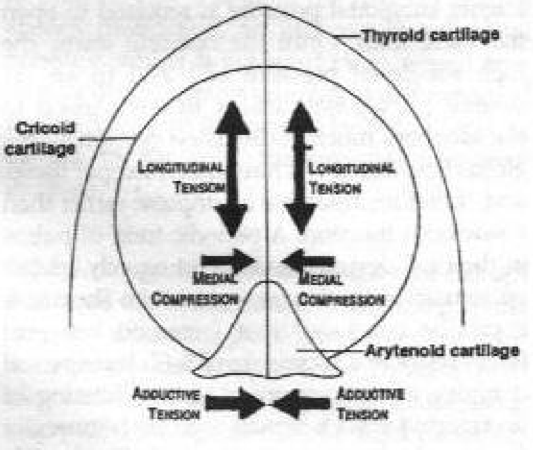


Figure 15–1. Schematic illustration of longitudinal tension, medial compression, and adductive tension.

2.1. Influence of the Supraglottal Cavities

As stated, the source signal that is the object of investigation is modified by the articulators in the supraglottal cavities before it is radiated from the mouth. The modification consists of additional turbulent noise, which has its origin in constrictions of the vocal tract as well as resonances and antiresonances. To investigate the source signal, the modifications caused by supraglottal components and mouth radiation have to be compensated for in the acoustic speech signal. A promising method of extracting the source signal from the acoustic speech signal is to compensate for mouth radiation by an integration filter. Resonances of the vocal tract can be compensated for by inverse filtering with an optimized filter. The filter coefficients are calculated using all-pole linear predictive coding (LPC) during the closed glottis interval of the preemphasized acoustic speech signal (Wong, Markel, & Gray 1979). If antiresonances are also to be compensated for, a pole-zero LPC is required, which is more difficult to handle, because it can lead to unstable filters (Makhoul, 1975). Turbulent noise caused by vocal tract constrictions causes the most difficult problem, because it can only barely be distinguished from turbulent noise originating at the glottis. Therefore the analysis of source characteristics must be restricted to the segments of the acoustic speech signal that are produced with minimal vocal tract constriction. The vowel /a/ is the phoneme spoken with the most open vocal tract and, therefore, seems to be most appropriate for investigations of the voice source in acoustic speech signals.

2.2. Parametric Description of the Source Signal

2.2.1. Modal Voice

In modal voicing, the forces of all laryngeal control mechanisms have to be offset by subglottal pressure to open the vocal folds, which are closed as long as subglottal pressure is absent or small. Therefore, glottal opening is not abrupt but gradual. The transglottal airflow rises from zero to peak flow. At maximum flow, a local pressure drop is created, which allows the vocal folds to return to a closed position. All laryngeal forces support this closing movement, which is much faster than the speed of glottal opening. Both the glottal opening function and the transglottal airflow function re-semble a sawtooth function, with a gradually increasing flank followed by an abrupt but finite time drop and a zero phase. In undistorted speech sounds with modal phonation, the transglottal airflow can be calculated by inverse filtering techniques. The spectral characteristic of an asymetric sawtooth-like triangular function consists of equally spaced spectral peaks at the fundamental frequency and at all higher harmonics, with a spectral damping of about 6 dB per octave toward higher frequencies in normal voices. Obviously, the sawtooth analogy is a very crude approximation of the transglottal flow, because natural signals do not have any sharp edges. However, it serves to illustrate the characteristics of the source signal.

2.2.1.1. Decreased Laryngeal Forces If the muscular forces helping maintain a closed glottis position were decreased, the result would be a steeper opening phase, and the glottal closure would be less abrupt. This skewing of the glottal pulse has been shown to result in decreased loudness and is, therefore, not an effective voicing manner (Sundberg, 1994). In the extreme, the sawtooth shape of the source signal changes into a symmetric triangle, or rather a sinus shape, because there are no sharp edges in physiological signals. In the frequency domain, this signal can be characterized by an increased spectral attenuation of higher harmonics, with a pure sine wave only having one single harmonic at the fundamental frequency. Inverse filtering of an acoustic speech signal produced with minimal laryngeal forces is not possible, because the closed glottis interval is unsuitable for calculation of the filter coefficients for the inverse filter. First, there is no impulse at the beginning of the closed phase, which would be necessary to interpret the acoustic speech signal as an impulse response of the supraglottal cavities during the closed glottis interval, and, second, the closed phase is often too short. But inverse filtering is not necessary, because the spectral damping of the source signal is very high. Higher harmonics that lie in the spectral region of formants have very little energy; therefore the source signal is hardly modified by the supraglottal cavities before it is radiated from the mouth.

2.2.1.2. Increased Laryngeal Forces If the muscular tension at the larynx is increased, the opening of the vocal folds is more difficult. Higher subglottal pressure is required to open the vocal folds. Once the opening starts, the high subglottal pressure will lead to an increased transglottal airflow. In comparison to the sawtooth function discussed previously, the glottal flow signal will have much steeper flanks and, therefore, resemble an impulse rather than a sawtooth function. A periodic train of pulses in the time domain leads to an equally spaced pulse train in the frequency domain. Therefore, it can be concluded that increased laryngeal forces result in a source signal with less spectral damping of its harmonics. Inverse filtering of undistorted speech signals with high muscular forces at the larynx and high vocal effort is possible, as long as the closed glottis interval is long enough to calculate the filter coefficients. It should be emphasized that this is only valid for modal voicing, in which the vocal folds vibrate completely. Falsetto, which is marked by extreme tension of the vocal folds and only partial vibration only, needs to be discussed separately.

2.2.1.3. Increased Fundamental Frequency and Increased Vocal Effort Modal voice is characterized by very little turbulent noise and by regular, periodic components. The fundamental frequency can be varied from lower to higher values. Vocal effort can change from soft to loud. An increase of the fundamental frequency, as well as an increase of vocal effort, leads to steeper glottal pulses and, therefore, to less spectral damping of the harmonics. It might make sense to split the category modal voice into two subcategories: one in which both fundamental frequency and vocal effort are low or moderate and a second category in which both are high, because these voice qualities are perceptibly quite different. Extremely high vocal effort would be shouted voice. Physiologically, the increase of vocal effort can be achieved by an increase of subglottal pressure, which also causes an increase of the fundamental frequency. This secondary effect could theoretically be offset by a relaxation of laryngeal forces, especially by a relaxation of the longitudinal tension of the vocal folds, but in practice this is seldom the case. In human speech production, an increase of subglottal pressure, which is achieved by increased tension of the breathing muscles, is accompanied somewhat by an increase of laryngeal tension, adding to the total fundamental frequency increase.

2.2.1.4. Irregular Glottal Cycles and Abrupt Changes of Fundamental Frequency In the previous schematic discussion, it proved relatively easy to differentiate between voice qualities with high or low laryngeal forces using inverse filtered signals or fast fourier transform (FFT) spectrograms. Problems arise if either the glottal cycles are irregular or the fundamental frequency changes abruptly, as is the case in fluent speech with marked prosody. In both cases, spectral lines at the harmonics will be smeared in the spectrum, which makes the determination of harmonic components more difficult. A parameter designed to deal with this problem is introduced in section 3.2. The Hilbert envelope of separate frequency bands is used to determine the composition of harmonic and noisy components within these bands.

2.2.2. Falsetto

In falsetto, only the thin edges of the vocal folds vibrate, causing the glottal pulses to take on a sinus-like shape rather than an impulse-like shape. Therefore, the spectral damping of higher harmonics can be strong. Inverse filtering is usually impossible because the closed glottis interval is too short. An unambiguous sign of falsetto is the extremely high fundamental frequency.

2.2.3. Creak (Glottal Fry, Vocal Fry)

Creak is characterized by irregular durations of glottal cycles or by a fundamental frequency of about half a speaker's normal fundamental frequency. In this way, creak can easily be determined in the acoustic time signal.

2.2.4. Whispery Phonation

So far, only the spectral distribution of harmonic signal components has been discussed in detail. Turbulent signal components appear if the glottis is neither completely closed nor widely open. The amplitude of the turbulent components can be so dominant that harmonic components are masked in the acoustic signal, at least in some frequency regions. Whispery phonation is characterized by relatively loud turbulent noise, which predominates over the harmonic components. Analysis of spectral damping characteristics of the harmonic components, as well as inverse filtering, is therefore difficult or impossible. But the strong turbulent noise can easily be detected in the acoustic time signal.

2.2.5. Breathy Phonation

Breathy voices are characterized by soft to no turbulent noise, because the glottis is fairly wide open at all times. The difference between minimal and peak transglottal flow is relatively small, which means that the alternating signal components that could be heard as tonal sound quality are also quite soft, although the absolute air flow is registered at comparably high values. The sound production is ineffective. As stated, laryngeal forces are minimal in breathy phonation, which leads to high spectral damping of higher harmonics. Analysis of spectral peaks would result in high bandwidth of the formants. Fundamental frequency is rather low. Typical signal characteristics for phonatory settings are shown in Figure 15–2.

3. ACOUSTIC PARAMETERS

In this section a brief definition is given of the derived acoustic parameters that can be used to determine the type of phonation in fluent speech signals. The parameters were selected according to the previously mentioned typical signal characteristics of the transglottal airflow. Only the combination of several acoustic parameters guarantees a reliable characterization of voice quality.

3.1. Glottal Pulses via Synchronous Inverse Filtering

Inverse filtering techniques to calculate glottal pulses from the acoustic speech signal can only be used in undistorted signals without any frequency-dependent damping or phase shifting. Theoretically, the process of glottal closure resembles an impulse. So in the subsequent closed-glottis interval, the acoustic speech signal can be interpreted as an impulse response of the vocal tract, because the subglottal volume is decoupled from the upper tract during that time. The filter coefficients for inverse filtering are calculated during the closed-glottis interval. The actual point of glottal closure can be determined by inverse filtering, with filter coefficients derived from a time interval that tends to be three to four times longer than one period duration of the acoustic signal. In general, it is more difficult to determine the exact point where the opening begins. In the present study, an 18th order covariance LPC and rectangular data windows were used. The filter coefficients for inverse filtering are calculated during the closed-glottis interval of the middle period of each realization of the German phoneme /a/. Because of different durations of the closed glottis interval in different glottis cycles, the length of the data window has to be adapted to allow for reasonable spectral shaping of the inverse filter. One hundred ms of the speech signal are filtered. Only the middle period within which the LPC coefficients were calculated was examined more closely. An example for an inverse filtered time signal from an utterance spoken with modal voice is given in Figure 15–3. (Transglottal airflow is presented in a negative direction, because this is often the case in electroglottogram displays.) Also discussed in Klasmeyer and Sendlmeier (1995), the glottal cycles may deviate considerably from the normal sawtooth shape. The inverse filtered signal should, therefore, be interpreted as a glottal pulse signal with great care, because some of the conditions for this theory may not be met. For example, pulses filtered from lax speech show hardly any obvious closed-glottis interval and (or because) the closure is not abrupt.

3.2. Spectral Distribution of Energy in Separate Frequency Bands and Detection of Harmonic Components Versus Turbulent Noise Within Separate Frequency Bands

The advantage of analysis techniques in the frequency domain is the ability to interpret

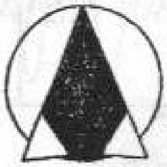
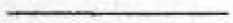
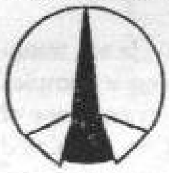
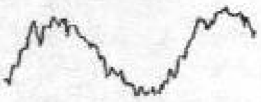
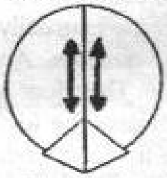
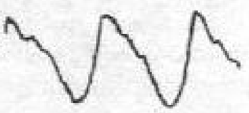
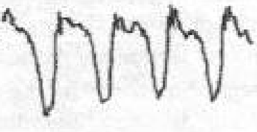
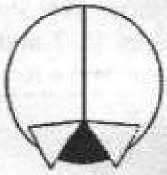
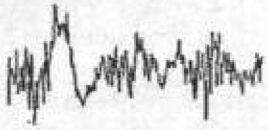
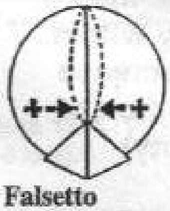
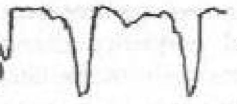
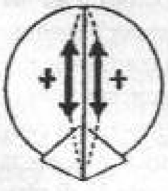
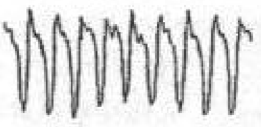
Breathing 	- maximum opening of glottis - all laryngeal forces minimal	*Pulse shape* 	- usually silent
Breathy phonation 	- extreme relaxation of muscle tension - vocal folds vibrate partially and do not close completely	*Pulse shape* 	- low F0 - flat spectral peaks - strong spectral damping - soft turbulent noise
Modal Voice I 	- regularly periodic - glottis closes completely - moderate laryngeal forces - no audible friction noise	*Pulse shape* 	- moderate F0 - normal spectral damping - very little turbulent signal components
Modal Voice II 	- passive longitudinal tension (increases F0) - strong subglottal pressure (also increases F0)	*Pulse shape* 	- relatively high F0 - less spectral damping - very little turbulent signal components
Whispery Phonation 	- moderate to high medial compression - weak adductive tension - cartilagious glottis does not close	*Pulse shape* 	- little energy in F0 and harmonics - relatively loud turbulent noise
Creaky Phonation 	- weak longitudinal tension - high medial compression and adductive tension (arytenoid cartilages closed)	*Pulse shape* 	- irregular duration of glottal cycles - steep glottal pulses with broad spectral energy
Falsetto 	- high values for all laryngeal forces - vibratory movement limited to thin glottal edges of vocal folds	*Pulse shape* 	- high F0 - little or high spectral damping - possibly loud turbulent noise

Figure 15–2. Typical signal characteristics for phonatory settings.

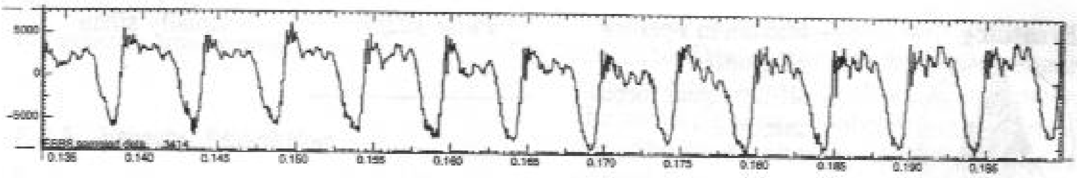


Figure 15–3. Inverse filtered time signal.

slightly phase-distorted signals. Although the inverse filtering technique described in the previous section requires signals without any phase distortion, which can only be recorded in anechoic rooms with special measuring microphones, the spectral parameters are more resilient to phase distortion. Frequency-dependent amplitude distortions will cause mistakes in measuring the time domain, as well as in the frequency domain.

To estimate spectral energy distribution, a quasistationary vowel segment without formant movements is chosen (phoneme /a/) using a wideband spectrogram as shown in Figure 15–4. Four different frequency bands are filtered from the segment: the very low band, which only contains the fundamental frequency; the low-band (0–1 kHz); the mid band (2.5–3.5 kHz); and the high band (4–5 kHz). The energy within the frequency bands is calculated and divided by the entire segment's energy.

A normal FFT of a quasistationary segment of the acoustic speech signal is appropriate for deciding whether the band contains harmonics of the fundamental frequency or turbulent noise only, if there are no voicing irregularities or steep fundamental frequency changes. Otherwise an unsynchronously inverse filtered time signal as shown in Figure 15–5 (residual signal) can be calculated.

The edges of this residual signal are multiplied with a Hanning window in the time domain before it is transformed to the frequency domain using an FFT algorithm. In the frequency domain, separate bands are calculated using spectral Hanning windows at the low band (0–1 kHz), the mid band (2.5–3.5 kHz), and the high band (4–5 kHz). All negative frequencies are set to zero. The analytical signals with frequencies above 0 Hz are transformed to the time domain using a complex inverse FFT algorithm. The Hilbert envelope of the separate frequency bands is calculated from the complex time signal by transforming the real and predicted parts into absolute values. The low-band Hilbert envelope is usually primarily harmonic for all voice qualities. An example is shown in Figure 15–6. Therefore a normalized crosscorrelation between the low and mid band and between the low and high band was calculated. Theoretically a high correlation would indicate harmonic mid-band and high-band components, whereas a low correlation would indicate noisy higher bands. Actually correlation was seldom high, even in those signals that had obvious harmonic components in higher bands. Therefore the decision whether a separate band was mainly harmonic or mainly turbulent was made directly from the Hilbert envelope. Figures 15–7 and 15–8 show examples of a harmonic and a turbulent high-band Hilbert envelope.

3.3. Average Gradient of the Fundamental Frequency Within Separate Increase or Decrease Segments (as an Estimation of Laryngeal Muscle Activity)

The fundamental frequency contour can be seen as an acoustic correlate of speech melody and tone of voice. However, the contour is not discussed in detail here. The average gradient of separate increase or decrease segments is only used to estimate the laryngeal muscle activity as one correlate of voice quality. Figure 15–9 provides an explanation: In the algorithm, local

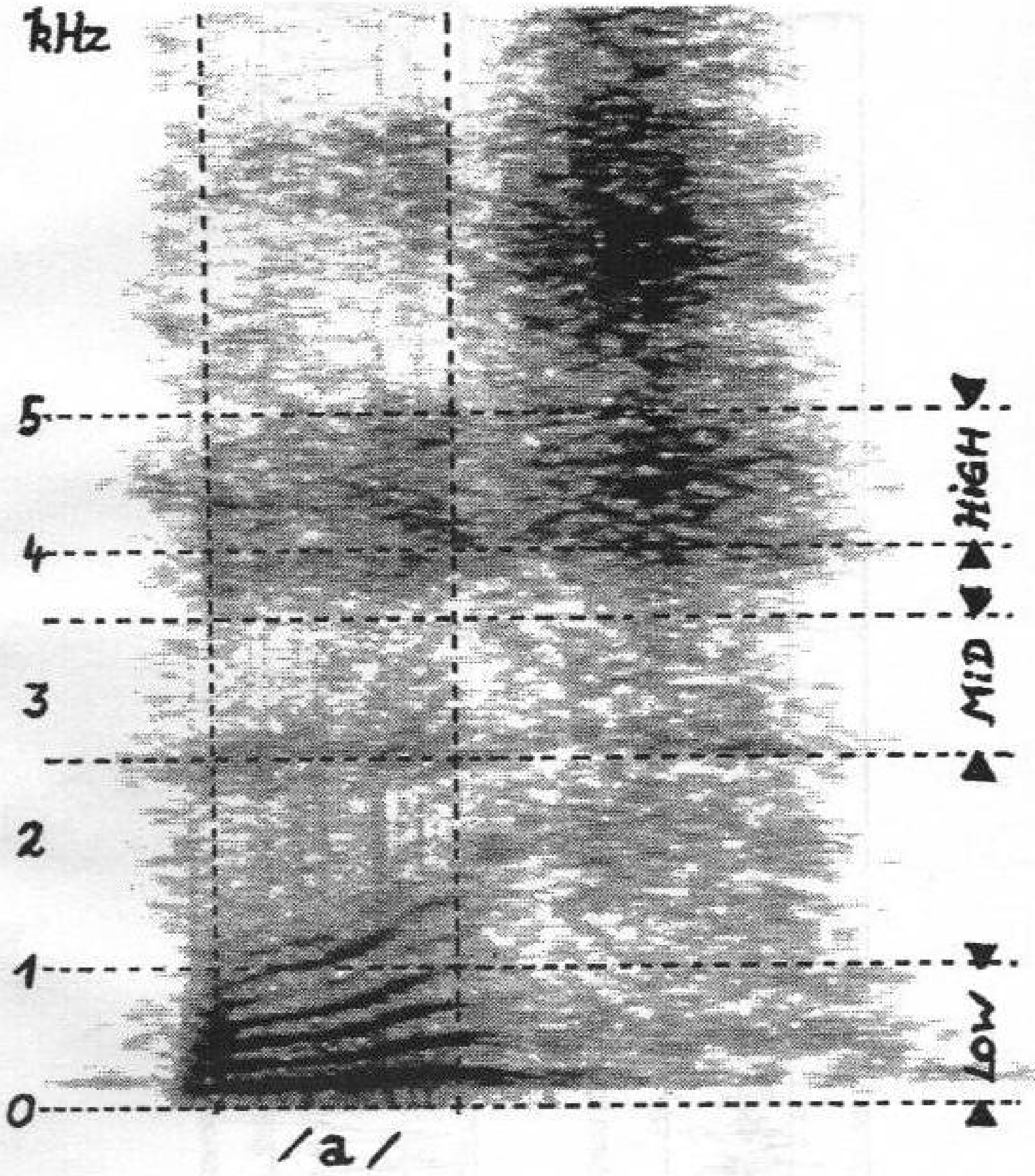


Figure 15–4. Wideband spectrogram, low-band (0–1 kHz), mid-band (2.5–3.5 kHz), and high-band (4–5 kHz).

maxima in the fundamental frequency contour are found. The average gradient is calculated in hertz/s by plotting a straight line through the contour. Rapid changes of the fundamental frequency are achieved by changes in laryngeal forces, especially by the longitudinal tension of the vocal folds.

3.4. Energy Difference of Vowels and Adjacent Fricatives as an Estimation of Vocal Effort

Steep pulses in the source signal are usually a sign of high vocal effort. An additional estimation of vocal effort can be made from an

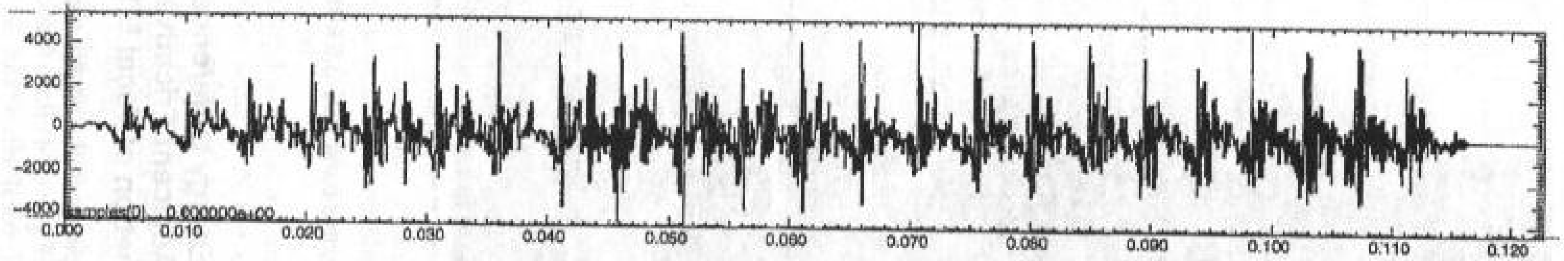


Figure 15–5. Unsynchronously filtered time signal (residual signal).

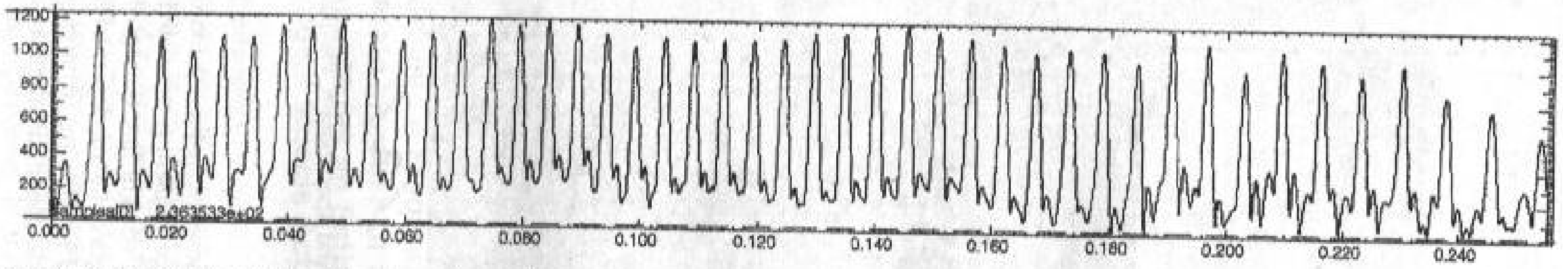


Figure 15–6. Low-band Hilbert envelope (harmonic band).

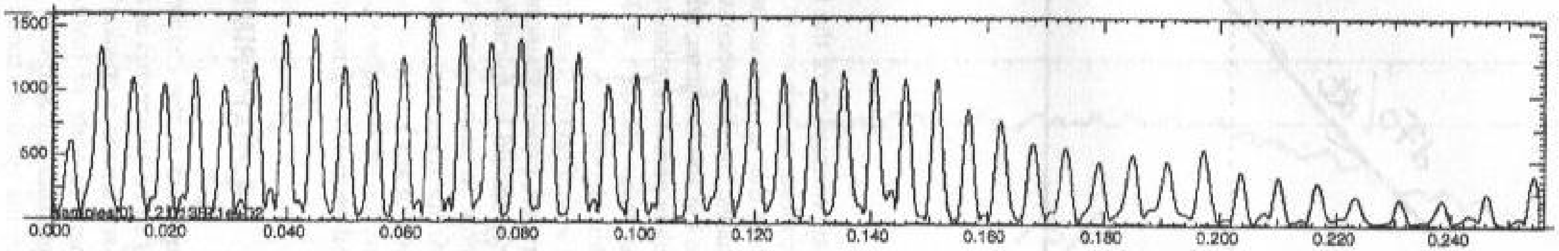


Figure 15–7. High-band Hilbert envelope (harmonic band).

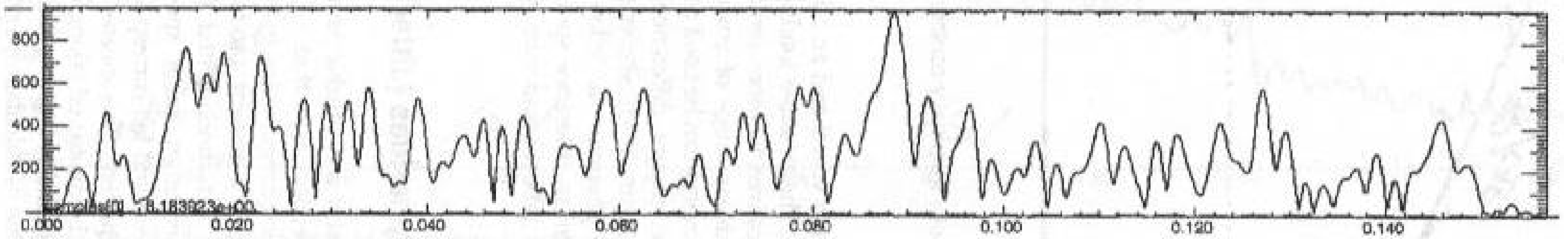


Figure 15–8. High-band Hilbert envelope (turbulent band).

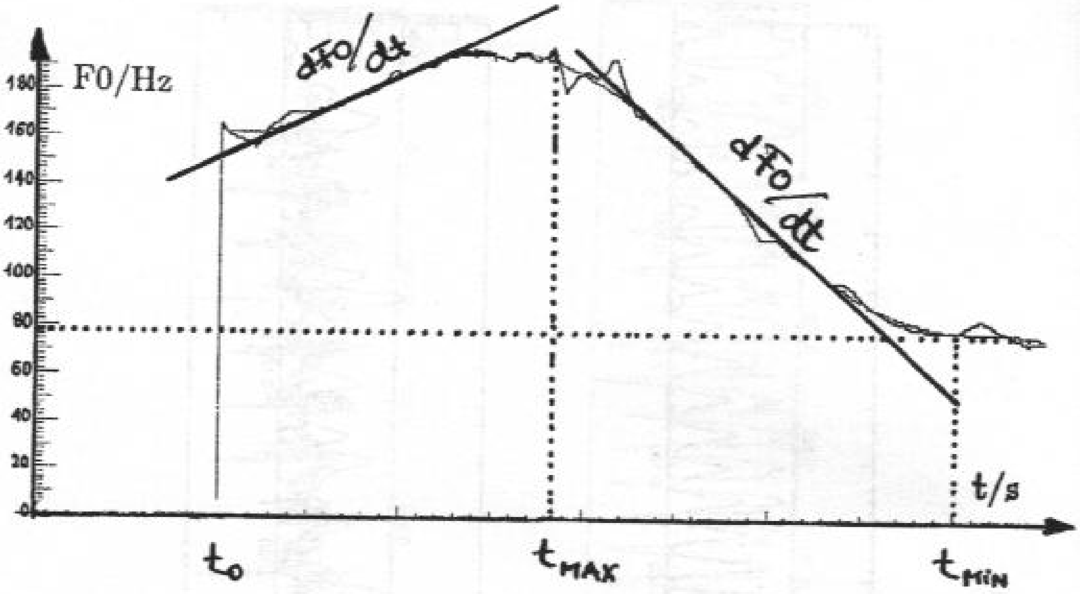


Figure 15–9. Fundamental frequency contour.

energy measurement in a vowel and an adjacent voiceless fricative. The possible loudness range of voiceless fricatives is more limited than the possible loudness range of vowels; therefore the fricative energy can be used as (a rough) reference. The energy difference is measured in decibels. Figure 15–10 shows the energy contour of the phoneme /a/ followed by the phoneme /s/ in an utterance spoken with high vocal effort (above) and with low vocal effort (below).

3.5. Voicing Irregularities (Jitter)

It should be stated that a very reliable fundamental frequency detection algorithm is required to measure voicing irregularities. In clinical voice measurement, a patient produces isolated vowels with flat fundamental frequency contours, whereas in fluent speech there is a continuous rise and fall corresponding to the intonation pattern. This means the parameter jitter has to be measured taking linguistic variation of the fundamental frequency into account. This is done by calculating a polynomial approximation of the fundamental frequency contour as is shown in Figure 15–11 and subtracting this approximation from the measured values. The difference values are used to calculate the absolute jitter, which is divided by the absolute period duration to calculate a relative jitter value (Rosken & Klasmeyer, 1995). This jitter algorithm is also used in forensic speaker identification (Wagner, 1995). In emotional speech, period durations vary; so the relative jitter should be calculated for short segments only. In such regions with short period durations, the measuring error due to low sampling rates or any kind of noise is much higher, therefore, the relative jitter values for high fundamental frequency segments are less reliable.

4. EXPERIMENT

As mentioned at the chapter opening, voice quality can be interpreted as a physical marker that characterizes a speaker's personal vocal equipment or state of health; it can also be interpreted as a psychological marker influenced by the speaker's mood or affective state.

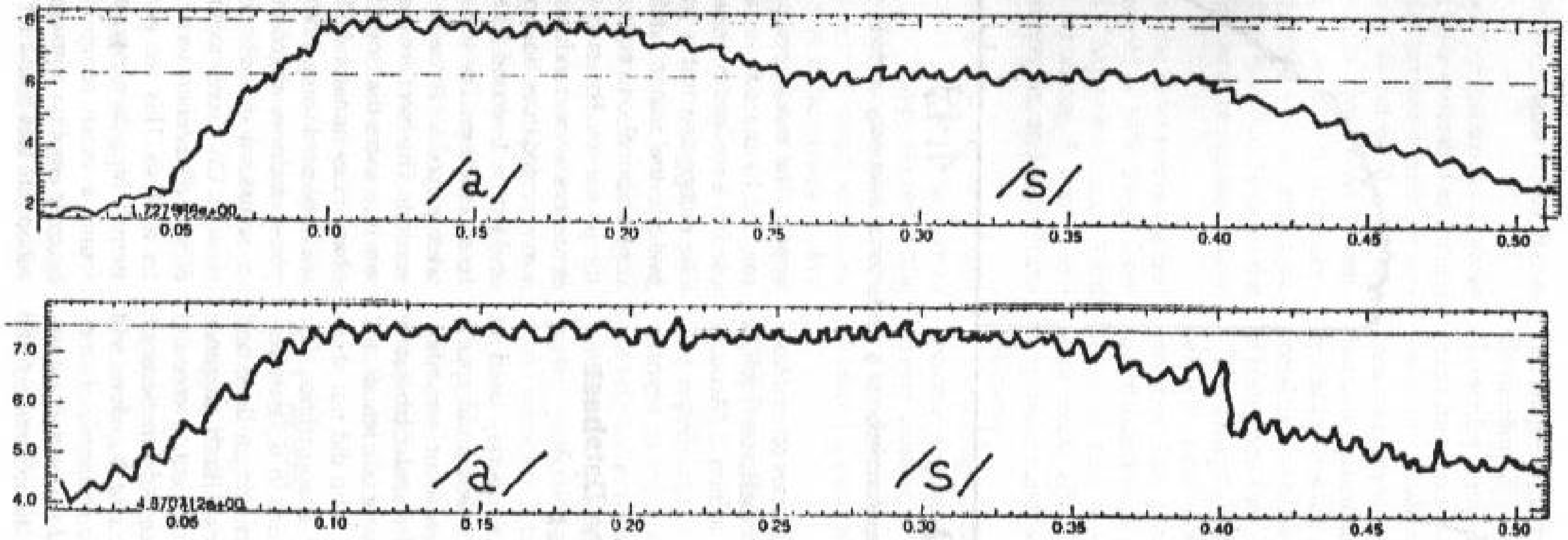


Figure 15–10. Energy contour of phonemes /a/ and /s/.

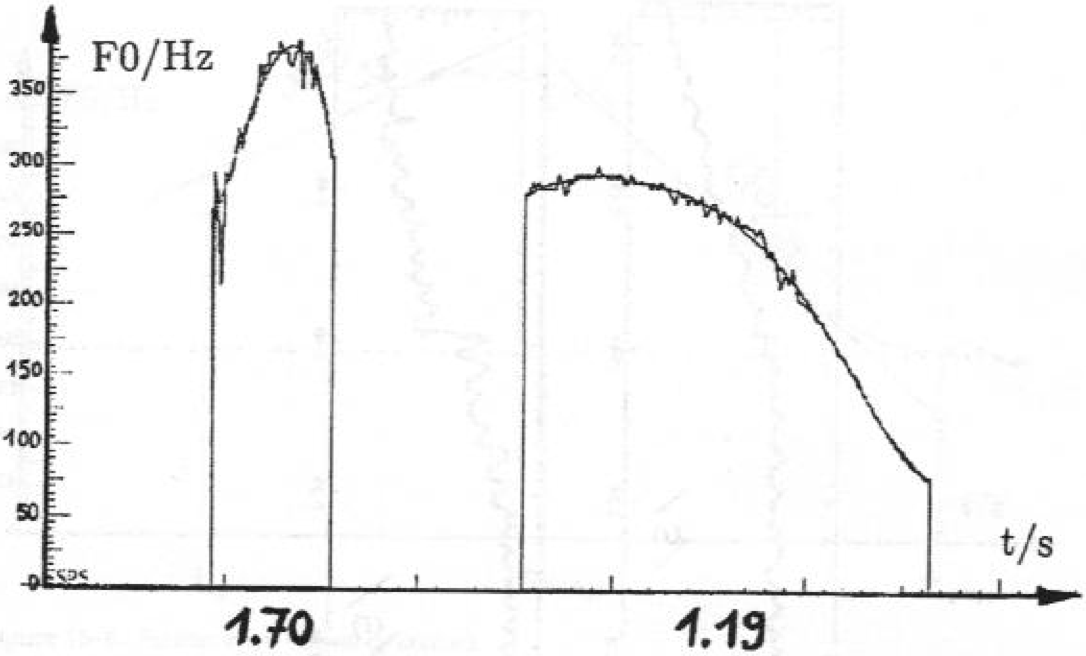


Figure 15–11. Polynomial approximation of a fundamental frequency contour.

The following experiment focuses on emotion-specific changes in the acoustic voice quality parameters introduced in section 3, "Acoustic Parameters."

4.1. Recording of the Database and Recognition Test

Seventy short sentences frequently used in everyday communication, which could appear in all emotional contexts without semantical inconsistency, were recorded under laboratory conditions. The speakers were one female and two male drama students, who did not show any voice pathologies or abnormalities. The sentences were DAT-recorded in separate sessions in an anechoic room using a Bruel and Kiger measuring microphone. Each utterance was spoken several times in a neutral voice and several times with the feigned emotions of: happiness, sadness, anger, fear, boredom, and disgust. To achieve realistic portrayals of emotion, the actors were asked to imagine a situational context in which the sentence could appear. The most appropriate realization was selected by the authors and each actor to be used in a recognition test with naive listeners. The recognition test is important not only to evaluate the recognizability of the emotional content, but also to evaluate the naturalness of the utterances. For each actor, the selected 70 sentences were randomized. Every sentence was repeated three times with 3-second pauses and with a 1-second 1 kHz tone between different emotions. The sentences from each actor were played to 20 naive listeners in separate sessions. The listeners were instructed to mark any utterances that sounded acted or overemphasized as unnatural. The emotional content was evaluated using 8 categories: neutral, happiness, sadness, anger, fear, boredom, disgust, or unnatural or not recognizable emotional content. Only sentences recognized by at least 80% of all listeners were used for the parameter analysis. This recognition threshold is extremely high. In a psychological study, Scherer found a mean recognition rate of 48% for acoustic emotion portrayals. The highest recognition rate was found for anger (78%) and the lowest rate was found for disgust (15%) (Banse & Scherer, 1996).

The utterances used for parameter analysis in the present study were evaluated as natural and the emotional content was deemed unambiguous. For speaker T, 44 out of 70 sentences were used in the analysis. These were 10 neutral sentences, 7 happy sentences, 3 sad sentences, 10 angry sentences, 6 bored sentences, and 8 sentences spoken with fear. None of the sentences with disgust were recognized above the threshold.

For speaker M, 41 of the 70 sentences reached the recognition threshold of 80%. These were 10 neutral sentences, 7 happy sentences, 3 sad sentences, 8 angry sentences, 5 bored sentences, 4 frightened sentences, and 4 sentences spoken with disgust.

For the female speaker, 37 out of 70 sentences were used in the analysis. These were 10 neutral sentences, only 1 happy sentence (the other happy utterances were judged overemphasized), 4 sad sentences, 10 angry sentences, 2 bored sentences, and 10 sentences spoken with fear. None of the sentences with disgust were recognized above the threshold.

The utterances were preselected by the authors and the actors; so high recognition rates were generally expected in the listening test. The relatively low recognition rates for disgust can be explained by the fact that vocal expression of disgust consists of brief affect bursts rather than of a long sentence spoken with a disgust-specific voice quality (Scherer, 1994). In the utterances spoken with disgust that were recognized above the threshold, the last syllable ended in a retching sound. Therefore, the sentences spoken with disgust were not analyzed any further. The relatively low recognition rates for sadness and boredom can be explained by the vocal expressions of sadness and boredom being quite similar; so sad and bored utterances were confused by the listeners.

4.2. Acoustic Parameters

The acoustic parameters previously outlined were used to analyze the emotional utterances. The shape of glottal pulses, as well as spectral distribution of energy and the detection of harmonics versus turbulent noise within different frequency bands, were measured in the vowel /a/ in all sentences. The reason why measurement was restricted to the vowel /a/ is that in the articulation of this vowel, the supraglottal cavities are not expected to cause additional friction noise. The results reflect the type of phonation that is a suprasegmental quality.

The energy difference between vowel and adjacent fricative was measured in the whole sentence. This parameter was used to check the plausibility of the parameters mentioned before. A high energy difference is a sign of great vocal effort that also correlates with steep pulses and little spectral damping of the harmonics.

The average gradient of separate increase or decrease segments of the fundamental frequency contour and the relative perturbation of period durations (jitter) were measured in all voiced segments.

4.3. Results

The experiment focuses on emotion-specific changes in voice quality. The speakers were not selected to represent a wide variety of personal voice qualities. All speakers had completed professional speech training in the same school of acting; they were of the same age and did not show any voice abnormalities. The interindividual differences between the speakers do not stand out from the intraindividual differences within one emotion and, hence, are not discussed any further.

4.3.1. Neutral

In the neutral reference utterances, the vowel energy is 1.2 dB higher than the fricative energy. The glottal pulses are similar to a sawtooth function. Most signal energy is concentrated in the low band below 1 kHz. Energy in all bands is mostly harmonic. Jitter is below 1%. (See Figure 15–12). The type of phonation can therefore be classified as normal modal voicing.

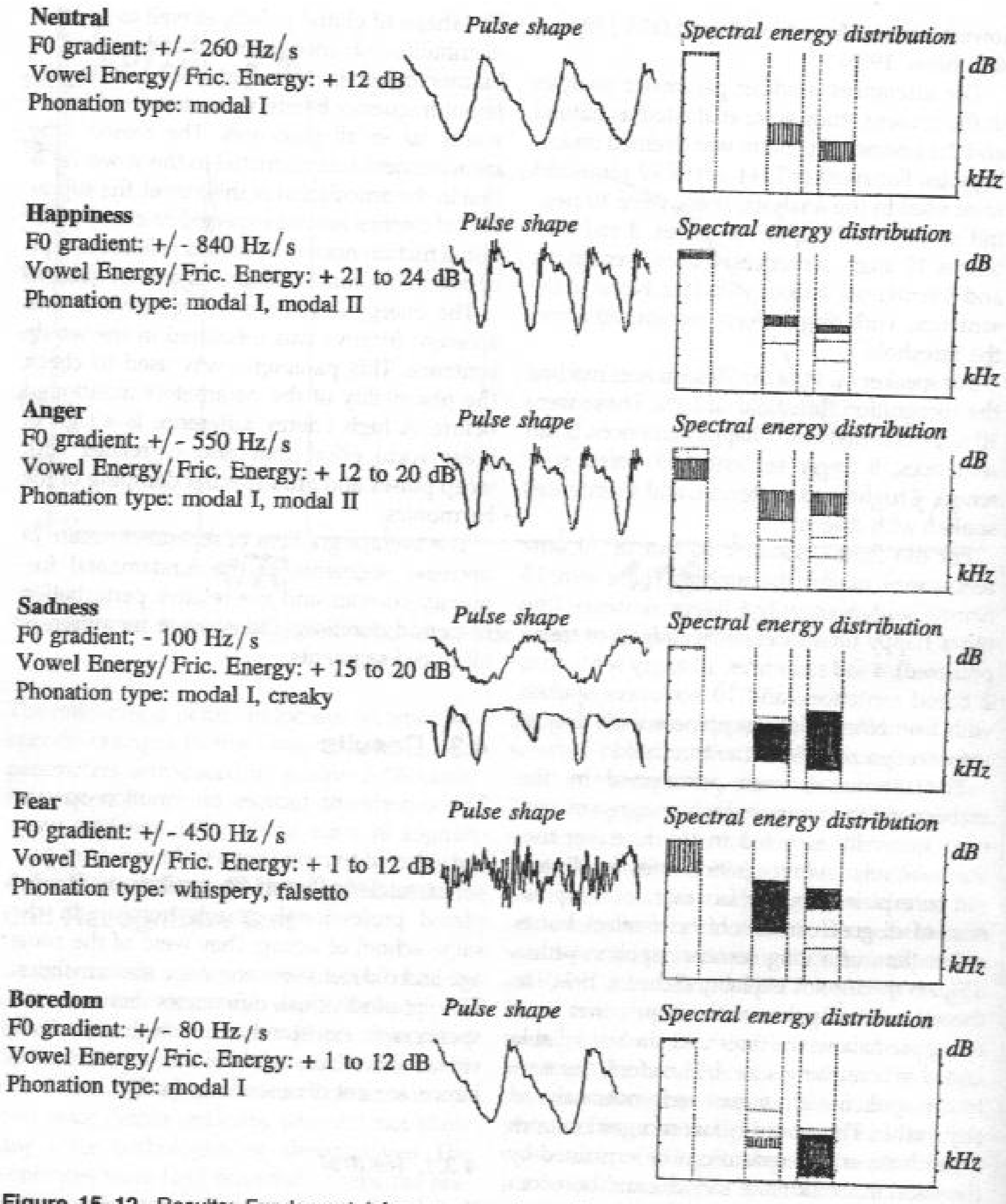


Figure 15–12. Results: Fundamental frequency and amplitude changes, type of phonation, pulse shapes, and spectral distribution of energy for emotional voices.

4.3.2. Happiness

In happy utterances, the vowel energy is 2.1 to 2.4 dB higher than the fricative energy, which also implies a greater loudness than in the neutral utterances. The glottal pulses are steeper than the neutral reference templates. Voiced segments contain largely harmonic energy in all frequency bands. Spectral damping of the harmonics is less than in neutral utterances. The average relative jitter value is slightly above 1%, but as the fundamental frequency reaches extremely high values at localized peaks, this also means that the systematic measuring error is high, as the speech signals were sampled at 16 kHz. The type of phonation can be classified as loud modal voicing with extremely fast changes in the fundamental frequency contour.

4.3.3. Anger

The vowel energy is 1.5 to 2 dB higher than the fricative energy, which also implies a greater loudness than in the neutral reference material. The glottal pulses are very steep with long closed-glottis intervals. As in happy utterances, voiced segments contain mainly harmonic energy in all frequency bands. Spectral damping of the harmonics is less than in neutral utterances. The physical energy in the mid and high band is not as high as in the low band, but considering the frequency-dependent loudness, the mid and high bands in angry speech signals are perceptibly important for the sound impression. As in happy utterances, the average relative jitter is slightly higher than in the neutral reference utterances. But, as was shown for happy utterances, the higher values might be caused by systematic errors at local fundamental frequency peaks. This type of phonation can be classified as shouted modal voicing with strongly marked prosody.

4.3.4. Sadness

The maximum difference between vowel and fricative energy is 1dB. This also means that the speech is not very loud. The inverse filtered signal looks very similar to a pure sine wave with obvious noise components. (The speech signal itself does not differ very much from that either.) This is consistent with the energy measurement, which shows marked spectral damping. Unlike the female speaker, for male speakers, the mid and high bands contain no harmonic, but only turbulent components. The sound impression of voiced segments is determined by the low frequency components, because there is not much energy in higher frequency regions. Most utterances show creak, which often appears at the beginning of voiced segments within a sentence. Creak is equal to a relative jitter value of 40% or 50%. The type of phonation in segments without voicing irregularities can be classified as breathy phonation.

4.3.5. Fear

The difference between vowel and fricative energy varies between 0.1 and 1.2 dB. This indicates a limited loudness. The inverse filtered signal looks very similar to the sad pulse shapes. The spectral distribution shows different values. In contrast to sad utterances, which have strong spectral damping in voiced segments, the utterances spoken with fear have very little spectral damping. This means that the turbulent noise components in the mid and high band are very important for the perceived sound quality. Not even the low band contains pure harmonic components. The fundamental frequency shows irregularities that result in high relative jitter values of between 2% and 8%. It should be mentioned that the utterances under investigation do not represent panic fear. They rather express whispery, suspicious fear. The fundamental frequency is high with little variation in the sense of increases or decreases for linguistic purposes. This type of phonation can be classified as breathy or whispery falsetto. But this conclusion might be wrong, because the turbulent noise components could also originate in an imprecise articulation with fricative constrictions during vowel articulation. The marked elongation of fricatives in the utterances could indicate that fear induces narrow fricative articulation in contrast to the usually more open vowel articulation.

4.3.6. Boredom

The difference between vowel and fricative energy is about 1.2 dB, which indicates a moderate loudness. The pulses are very similar to a sawtooth function, as is the neutral material, although the spectral damping in the mid and high band is stronger than in the neutral utterances. Unlike the female speaker, the male speakers do have turbulent noise in the higher frequency bands, but owing to the relatively low energy, these higher frequencies seem less important for the perceived sound quality. In contrast to sad utterances, bored speech does not show voicing irregularities or creaky phonation. This type of phonation can be classified as normal modal voicing with little dynamics in the fundamental frequency contour, that is, relatively monotonous speech.

5. SUMMARY

In the initial part of this chapter signal characteristics of the transglottal airflow were discussed with regard to the question of how different voice qualities can be measured in the acoustic speech signal. Several acoustic parameters were discussed. To deal with the fact that some voice qualities can hardly be regarded as quasistationary signals and hence difficult to analyze using common algorithms for stationary signals, some of the discussed parameters are not orthogonal, but represent varying approaches to measuring similar phenomena in the frequency and time domains. The combination of these acoustic parameters allows plausibility checks, so that the process can be trusted as a reliable method for the characterization of voice quality. The experiment presented focused on emotion-specific changes in the acoustic voice quality parameters. The speakers were not selected to represent a wide variety of different personal voice qualities. The interindividual differences were not more marked than of the intraindividual differences within one emotion.

Voice quality in neutral utterances could be classified as normal modal voicing for all speakers. In happy utterances, voice quality could be classified as loud modal voicing with extremely rapid changes in the fundamental frequency contour. In angry utterances the speakers used shouted phonation with strongly marked prosody. Sad utterances were spoken with breathy phonation. The type of phonation often changed within the sad utterances. Creak appears at the beginning of voiced segments. In frightened utterances the average fundamental frequency is high with little variation for linguistic purposes. Voice quality in sentences spoken with fear is (presumably) breathy or whispery falsetto. The fundamental frequency shows irregularities that result in high relative jitter values of about between 2% and 8%. Bored speech often resembles sad speech. Voice quality in those utterances that were unambiguously recognized as bored speech, do not show voicing irregularities or creaky phonation as the sad utterances do. This type of phonation can be classified as normal modal voicing with little dynamics in the fundamental frequency contour.

REFERENCES

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614.

Klasmeyer, G., & Sendlmeier W. F. (1995). Objective voice parameters to characterize the emotional content in speech. *Proceedings of the International Conference of Phonetic Sciences (Stockholm), 1,* 181–185.

Laver, J. (1994). *Principles of phonetics.* Cambridge, UK: Cambridge University Press.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings IEEE, 63,* 561–580.

Rosken, W., & Klasmeyer, G. (1995). Erfassung von F0-Irregularitäten in gesprochener Sprache als messbarer Parameter zur Beschreibung von Stimmqualitäten. *Fortschritte der Akustik, DAGA 95,* 1047–1050.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research, *Psychological Bulletin, 1986, 99,* 143–165.

Scherer, K. R. (1994). Affect bursts. In S. M. H. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161–196). Hillsdale, NJ: Erlbaum.

Sundberg, J. (1994). Vocal fold vibration patterns and phonatory modes. *Speech Transmission Lab* (Quarterly Progress Status Report 2–3/94). Stockholm: Royal Institute of Technology.

Wagner, I. (1995). A new jitter-algorithm to quantify hoarseness: An exploratory study, *Forensic Linguistics, 2*(1), 18–21.

Wong, D. J., Markel, J. D., & Gray, A. H. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics,*