The classification of different phonation types in emotional and neutral speech

Gudrun Klasmeyer and Walter F. Sendlmeier

Technical University, Berlin

ABSTRACT A general approach is proposed towards the objective measurement of different phonation types in fluent speech. Signal characteristics of the transglottal airflow 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.

The following experiment focuses on emotion-specific changes in acoustic correlates of voice quality. Emotional arousal can influence the speaker's voice quality or type of phonation and listeners use this knowledge to attribute affective states to the speaker. The forensic relevance of the investigation can be seen in speaker-identification tasks. From the experiment the conclusion cannot be drawn that voice quality must change under emotional arousal. Hence the analysis of acoustic speech signals cannot serve as a reliable method of measuring the speaker's affective state.

KEYWORDS voice quality; emotion; phonation; paralinguistic aspects.

INTRODUCTION

Besides semantic information spoken language contains evidential information which is conveyed by signs in speech which act as attributive markers (Laver 1994). These are used by the 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.

1350–1771

Forensic Linguistics 1(4) 1997

Most of the named attributive markers are used by listeners to identify a familiar speaker. In forensic situations speakers often have to be identified from short, recorded speech samples. In this case identification is more difficult, not only because non-cooperative speakers disguise their voices, but also because psychological markers of affective state or mood can change with the speaker's emotional situation.

In forensic practice the acoustic speech signal is analysed to support the identification process. But the correlation between attributive markers and acoustic parameters is rather complicated. Laver states that physical characteristics lie in a speaker's voice quality, while social markers include such features as accent and choice of vocabulary. Psychological markers are often taken to reside in a speaker's tone of voice (Laver 1994: 14). 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 which 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 the speaker's personal voice quality or state of health, but it could also be a psychological marker for fear.

This paper proposes a general approach towards the objective measurement of different phonation types in fluent speech.

Initially, conclusions about typical signal characteristics of the transglottal airflow are drawn from the production strategies of different 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 quasi-stationary signals, and hence are difficult to analyse 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 above, voice quality can be used by listeners as a basis on which to attribute speaker characteristics.

The following experiment focuses on emotion-specific changes in acoustic correlates of voice quality. Emotional arousal can influence the speaker's voice quality or type of phonation and listeners use this knowledge to attribute affective states to the speaker. The forensic relevance of the investigation can be seen in speaker-identification tasks. Experimental results prove that acoustic parameters of voice quality are affected in ways that lead to unambiguous attributions of speaker emotions in a listening test with naive listeners. From the experiment one cannot conclude that voice quality necessarily changes under emotional arousal, nor can the analysis of acoustic speech signals serve as a reliable method to measure the speaker's genuine affective state.

Readers who are familiar with the anatomy of the voice source and production strategies of different types of phonation could skip the introductory section of this paper and proceed to the section **Signal characteristics of acoustic speech signals.**

Anatomy and control mechanisms of the voice source

In human speech production the breathing organs produce an airflow which 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 chords. 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 chords, 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 which closes the ligamental glottis as a secondary effect of a force on the arytenoid cartilages, which is caused by tension of the lateral cricothyroid muscles in collaboration with tension in the lateral part of the thyroarytenoid muscles. These forces of longitudinal tension, medial compression and adductive tension are demonstrated in Figure 1, which shows a cross-section of the human larynx.

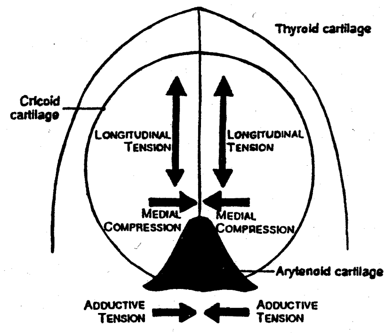

Figure 1 Schematic illustration of longitudinal tension, medial compression and adductive tension

Phonetic categories of the vo

The most basic differentiation
of the voice source is the voic
category, two main kinds of ph
glottis with completely relaxe
does not cause an audible sou
muscular forces which are not
this will lead to local turbul
audible as turbulent noise. A d
transglottal airflow or an incre
of turbulences, and the frictio
audible voiceless sound may b
followed by a sudden opening
has built up. The resulting au

All signals which are produc
chords fall into the category
airflow signal in voiced sound
less additional friction noise.
be further differentiated into p
production mechanisms. Th
phonation, whispery phonati
'glottal fry' or 'vocal fry') and
voice quality is often a combi
following production mechan
about typical characteristics c

Breathy phonation

Breathy phonation is produce
vocal chords, which are relax
flow, but they do not close com
turbulent noise and harmoni
the amount of transglottal ai
usually low.

Whispery phonation

Compared with breathy phon
tension of the vocal chords
consequence, the ligamental
while the cartilaginous part r
the average opening section
increased friction noise, w
components of the source sig

Phonetic categories of the voice source

The most basic differentiation of larynx functions into phonetic categories of the voice source is the voiced/unvoiced dimension. Within the voiceless category, two main kinds of phonation can be differentiated. A fully opened glottis with completely relaxed muscular forces, as is the case for breathing, does not cause an audible sound. If the opening section is constricted by muscular forces which are not strong enough to cause a complete closure, this will lead to local turbulences in the transglottal airflow which are audible as turbulent noise. A decrease of the opening section with constant transglottal airflow or an increase of transglottal flow will lead to an increase of turbulences, and the friction noise will be louder. The way in which an audible voiceless sound may be caused is a single complete glottal closure followed by a sudden opening and release of the subglottal pressure which has built up. The resulting audible signal is a transient impulse.

All signals which are produced with oscillating or partly oscillating vocal chords fall into the category of voiced sounds. Therefore the transglottal airflow signal in voiced sounds consists of periodic air pulses with more or less additional friction noise. The great variety of these voiced sounds may be further differentiated into phonetic categories according to their different production mechanisms. The most important categories are breathy phonation, whispery phonation, modal voice, creak (which is also called 'glottal fry' or 'vocal fry') and falsetto. In natural communication situations voice quality is often a combination of different types of phonation. The following production mechanisms serve as a basis from which conclusions about typical characteristics of transglottal airflow signals are drawn.

Breathy phonation

Breathy phonation is produced with relaxation of all muscular forces. The vocal chords, which are relaxed and thick, oscillate in the transglottal air flow, but they do not close completely at any time. The source signal contains turbulent noise and harmonic components; both are very weak, though the amount of transglottal airflow is high. The fundamental frequency is usually low.

Whispery phonation

Compared with breathy phonation, medial compression and longitudinal tension of the vocal chords are much higher in whispery phonation. In consequence, the ligamental part of the glottis stays more or less closed, while the cartilaginous part remains open. As compared to breathy voice, the average opening section of the glottis is much smaller. This leads to increased friction noise, which usually predominates the harmonic components of the source signal.

Modal voice
For modal voice, all muscular forces are moderate. In the absence of or with low subglottal pressure, the glottis is closed completely. The fundamental frequency of the oscillating vocal chords is dependent on muscular control of the larynx as well as on subglottal pressure. In fluent speech, muscular forces in the larynx are used to produce an intonation pattern. The most important control mechanism of the fundamental frequency seems to be longitudinal tension of the vocal chords. An increase of tension results in an increase of fundamental frequency.

Creak (glottal fry, vocal fry)
Creak is produced when the longitudinal tension of the vocal chords is minimal, whereas medial compression and adductive tension are high. In this case the vocal chords are thick and relaxed owing to the missing longitudinal tension. They open either at irregular temporal distances or at a nearly periodic oscillation of about one half of the speaker's 'normal' fundamental frequency. In the second case the air pulses can be heard as single events rather than as a tonal sound.

Falsetto
Falsetto is a phonation type in which the longitudinal tension is extremely high. This causes an extremely high fundamental frequency. The vocal cords are very thin and stiff, so only the edges can oscillate in the airflow.

Depending on the muscular forces required for different phonation types, some combinations are possible, while others cannot occur for physiological reasons.

SIGNAL CHARACTERISTICS OF ACOUSTIC SPEECH SIGNALS

Influence of the supraglottal cavities

As stated above, the source signal which is the object of the 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 anti-resonances. To investigate the source signal, the modifications which are caused by supraglottal components and mouth radiation have to be compensated in the acoustic speech signal. This is not so. A promising method of extracting the source signal from the acoustic speech signal works as follows: mouth radiation can be compensated by an integration filter (Wong *et al.* 1979). Resonances of the vocal tract can be compensated by inverse filtering with an optimized filter. The filter coefficients are calculated using an all-pole LPC during the closed glottis interval of the pre-emphasized acoustic speech signal (Wong *et al.* 1979).

If anti-resonances are also
which is more difficult t
(Makhoul 1975). Turbule
the most difficult proble
from turbulent noise orig
source characteristics mu
speech signal which are p
The vowel /a/ is the phon
therefore seems most ap
in acoustic speech signals

Parametric description of

The distinction between v
been adequately addressed
focuses on the objective de
whispery voice, modal v
parameters. The above-m
in the acoustic speech sig
harmonic and noise compo
derived from the time sign
vocal effort, voicing irr
fundamental frequency an
acoustic speech signals can

Modal voice
In modal voicing the force
compensated by the subglo
which are closed as long as
glottal opening is not abr
from zero to peak flow. At
which allows the vocal cho
forces support this closing
opening. The glottal openi
both resemble a sawtoot
followed by an abrupt dro
with modal phonation the
filtering techniques. The
consists of equally spaced s
at all higher harmonics wi
in 'normal' voices. Obvi
approximation of the tra
have any sharp edges. Ho
of the source signal.

If anti-resonances are also to be compensated, 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 which are produced with minimal vocal tract constriction. The vowel /a/ is the phoneme spoken with the most open vocal tract and therefore seems most appropriate for investigations of the voice source in acoustic speech signals.

Parametric description of the source signal

The distinction between voiced and voiceless sound segments has already been adequately addressed and solved in the research literature. This paper focuses on the objective description of voiced sounds, such as breathy voice, whispery voice, modal voice, creak and falsetto in terms of acoustic parameters. The above-mentioned 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.

Modal voice

In modal voicing the forces of all laryngeal control mechanisms have to be compensated by the subglottal pressure in order to open the vocal chords, 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 caused, which allows the vocal chords to return to a closed position. All laryngeal forces support this closing movement which is much faster than the glottal opening. The glottal opening function and the transglottal airflow function both resemble a sawtooth function with a gradually increasing flank followed by an abrupt drop and zero phase. In undistorted speech sounds with modal phonation the transglottal airflow can be calculated by inverse filtering techniques. The spectral characteristic of a sawtooth function consists of equally spaced spectral peaks at the fundamental frequency and at all higher harmonics with a spectral damping of about 12 dB per octave in 'normal' voices. Obviously the sawtooth function 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.

Decreased laryngeal forces If the muscular forces which would aid a closed glottis position were decreased, the result would be a steeper opening phase while the glottal closure would be less abrupt. This skewing of the glottal pulse is said to result in decreased loudness and is therefore a less effective way of voicing (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 which lie in the spectral region of formants have very little energy; therefore the source signal is hardly modified by the supraglottal cavities at all before it is radiated from the mouth.

Increased laryngeal forces If the muscular tension at the larynx is increased, the opening of the vocal chords is more difficult. Higher subglottal pressure is required to cause an opening of the vocal chords. Once the opening starts, the high subglottal pressure will lead to an increased transglottal airflow. In comparison to the sawtooth function discussed above, 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 these points are only valid for modal voicing, where the vocal chords vibrate completely. Falsetto, which is marked by extreme tension of the vocal chord and partial vibration only, must be the subject of a separate discussion.

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 both lead to steeper glottal pulses and therefore to less spectral damping of the

harmonics. It might make
two sub-categories: one in
effort are low or modera
high, because these voic
Extremely high vocal effo
increase of vocal effort can
which also causes an in
secondary effect could the
laryngeal forces, especially
the vocal chords, but in pr
production an increase c
increased tension of the b
increase of laryngeal ten
increase.

Irregular glottal cycles and a
schematic discussion above
voice qualities with high or
or FFT spectrograms. Probl
or if the fundamental freq
speech with marked prosod
in the spectrum then, w
components more difficul
problem will be introduce
of energy in separate fr
components versus turbu
below. The Hilbert envel
determine the compositio
these bands.

Falsetto
In falsetto only the thin e
glottal pulses to take on
shape. Therefore the spectr
Inverse filtering is usually i
too short. An unambiguous
frequency.

Creak (glottal fry, vocal f
Creak is characterized by
fundamental frequency of
frequency. In this way, it
signal.

harmonics. It might make sense to split the category modal voice into two sub-categories: 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 compensated by a relaxation of laryngeal forces, especially by a relaxation of the longitudinal tension of the vocal chords, 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 rather accompanied by an increase of laryngeal tension, adding to the fundamental frequency increase.

Irregular glottal cycles and abrupt changes of fundamental frequency In the schematic discussion above it proved relatively easy to differentiate between voice qualities with high or low laryngeal forces using inverse filtered signals or FFT spectrograms. Problems arise if either the glottal cycles are irregular or if the fundamental frequency changes abruptly, as is the case in fluent speech with marked prosody. Spectral lines at the harmonics will be smeared in the spectrum then, which makes the determination of harmonic components more difficult. An alternative parameter, dealing with this problem will be introduced in the section entitled **Spectral distribution of energy in separate frequency bands and detection of harmonic components versus turbulent noise within separate frequency bands,** below. The Hilbert envelope of separate frequency bands is used to determine the composition of harmonic and noisy components within these bands.

Falsetto

In falsetto only the thin edges of the vocal chords 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.

Creak (glottal fry, vocal fry)

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

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 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.

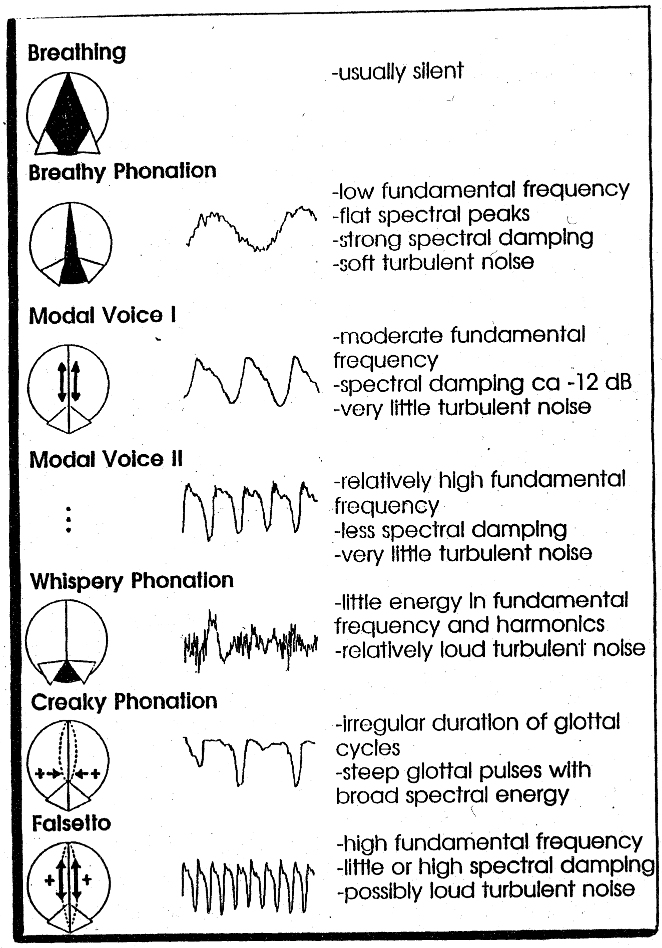
Breathing
-usually silent
Breathy Phonation
-low fundamental frequency
-flat spectral peaks
-strong spectral damping
-soft turbulent noise
Modal Voice I
-moderate fundamental frequency
-spectral damping ca -12 dB
-very little turbulent noise
Modal Voice II
-relatively high fundamental frequency
-less spectral damping
-very little turbulent noise
Whispery Phonation
-little energy in fundamental frequency and harmonics
-relatively loud turbulent noise
Creaky Phonation
-irregular duration of glottal cycles
-steep glottal pulses with broad spectral energy
Falsetto
-high fundamental frequency
-little or high spectral damping
-possibly loud turbulent noise

Figure 2 Typical signal characteristics for phonatory settings

Breathy phonation

Breathy voices are charact
glottis is fairly wide open a
peak transglottal flow is re
signal components which
quite soft, though the abs
values. The sound produ
forces are minimal in bre
damping of higher harmo
high bandwidth of the for

A short summary of typica
given in Figure 2.

ACOUSTIC PARAMET

In this section a brief defini
which can be used to det
signals. The parameters w
typical signal characteristics
of several acoustic paramet
quality.

Glottal pulses via synchro

Inverse filtering technique
speech signal can only be us
dependent damping or ph
closure resembles an imp
the acoustic speech signal
vocal tract, because the s
tract during this time. T
calculated during the clos
closure can be determined
from a time interval which
period duration of the ac
determine the exact point
an 18th order covariance
The filter coefficients for i
glottis interval of the mid
phoneme /a/. Due to diffe
different glottis cycles the
with regard to reasonable
the speech signal are filte

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 which could be heard as tonal sound quality are also quite soft, though the absolute air flow is registered at comparably high values. The sound production is ineffective. As stated above, 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.

A short summary of typical signal characteristics for phonatory settings is given in Figure 2.

ACOUSTIC PARAMETERS

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

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 following 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 this 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 which 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/. Due to different durations of the closed-glottis interval in different glottis cycles the length of the data window has to be adapted with regard to reasonable spectral shaping of the inverse filter. 100 ms of the speech signal are filtered. Only the middle period within which the

LPC coefficients were calculated was examined further. An example for an inverse filtered time signal from an utterance spoken with modal voice is given in Figure 3. (Transglottal airflow is presented in a negative direction, because this is often the case in electroglottogram displays.) As discussed above, 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.*

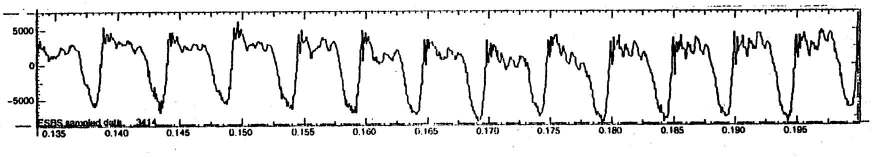

Figure 3 Inverse filtered time signal

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 that slightly phase-distorted signals can still be interpreted. While the inverse filtering

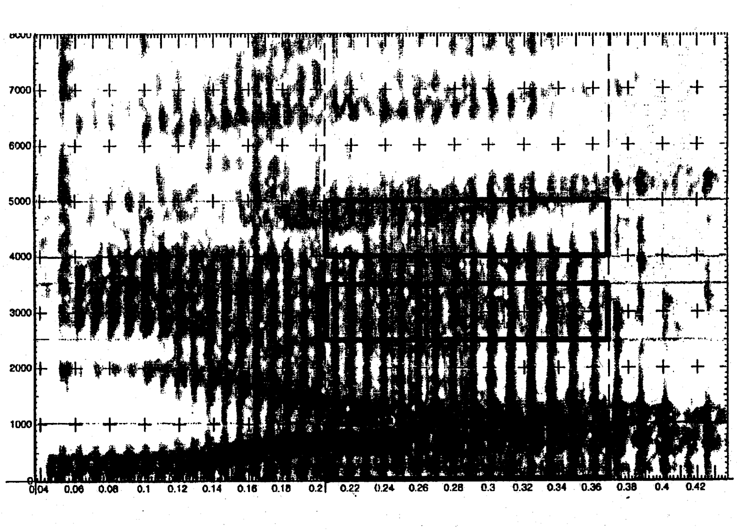

Figure 4 Wideband-Spectrogram, Low-band (0–1kHz), Mid-band (2.5–3.5kHz) and High-band (4–5kHz)

technique described in the
phase distortion, which ca
special measuring microph
to phase distortion. Freq
cause measuring mistakes
domain.

In order to estimate spec
segment without formant
wideband spectrogram as sh
are filtered from the segmen
fundamental frequency; the
kHz); and the high band (4–5
is calculated and divided by

A normal FFT of a quasi-s
is only appropriate for deci
the fundamental frequency
irregularities or steep fun
unsynchronously inverse filt
signal) can be calculated. T
with a Hanning window in t
frequency domain using an F
bands are calculated using s
1 kHz), the mid-band (2.5-
negative frequencies are set t
above 0 Hz are transformed
FFT algorithm. The Hilbert
calculated from the comple
predicted parts into absolu
usually mainly harmonic fo
Figure 6. Therefore a norm
mid-band and between the lo
a high correlation would i
components, whereas a low c
Actually correlation was sel
obvious harmonic compone
whether a separate band wa
made directly from the Hilb
of a harmonic and a turbule

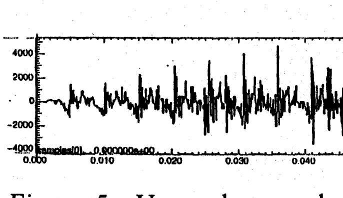

Figure 5 Unsynchronously

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 measuring mistakes in the time domain as well as in the frequency domain.

In order to estimate spectral energy distribution a quasi-stationary vowel segment without formant movements is chosen (phoneme /a/) using a wideband spectrogram as shown in Figure 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 whole segment's energy.

A normal FFT of a quasi-stationary segment of the acoustic speech signal is only appropriate for deciding whether the band contains harmonics of the fundamental frequency or turbulent noise, if there are no voicing irregularities or steep fundamental frequency changes. Otherwise an unsynchronously inverse filtered time signal as shown in Figure 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 mainly harmonic for all voice qualities. An example is shown in Figure 6. Therefore a normalized cross-correlation between the low and mid-band and between the low and highband 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 which 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 7 and 8 show examples of a harmonic and a turbulent high-band Hilbert envelope.

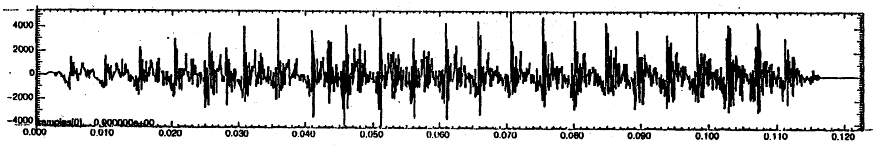

Figure 5 Unsynchronously filtered time signal (residual signal)

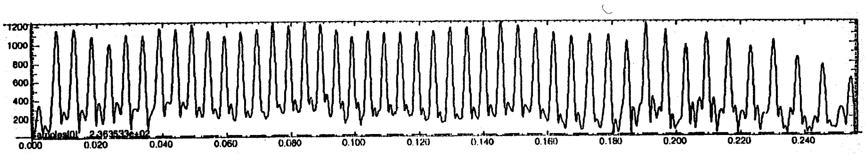

Figure 6 Low-band Hilbert envelope (harmonic band)

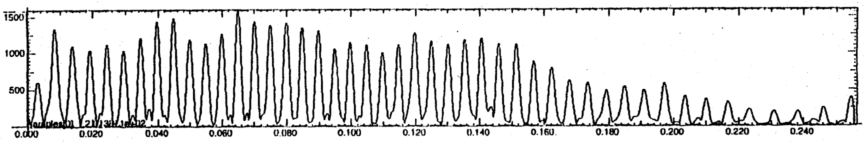

Figure 7 High-band Hilbert envelope (harmonic band)

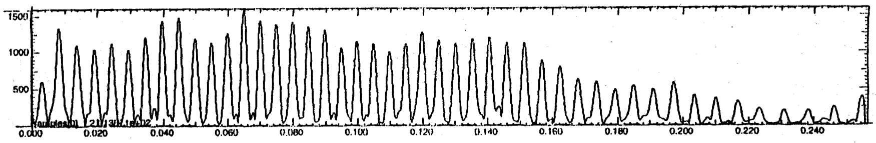

Figure 8 High-band Hilbert envelope (turbulent band)

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 9 gives an explanation: in the algorithm local maxima in the fundamental frequency contour are found. The average gradient is calculated in Hz/s by fitting 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 chords.

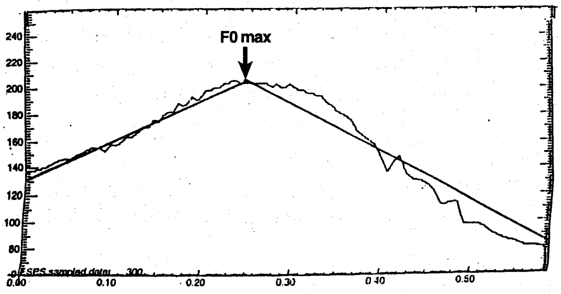

Figure 9 Fundamental frequency contour

Energy difference of vowe vocal effort

Steep pulses in the source s additional estimation of measurement in a vowel ar loudness range of voiceles loudness range of vowels; rough) reference. The energ the energy contour of the p utterance spoken with high (below).

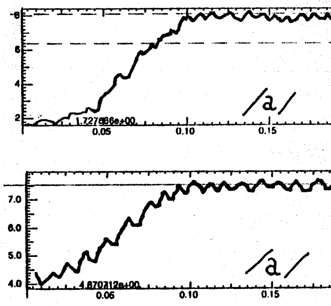

Figure 10 Energy contour

Voicing irregularities (jitte

It should be stated that a algorithm is required to m measurement the patient p frequency contours, where and fall corresponding to th 'jitter' has to be measured of the fundamental freque approximation of the funda 11 and subtracting this ap difference values are used t by the absolute period dura and Klasmeyer 1995). This identification (Wagner 199 so the relative jitter should regions with short period sampling rates or any kind values for high fundament

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 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 dB. Figure 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).

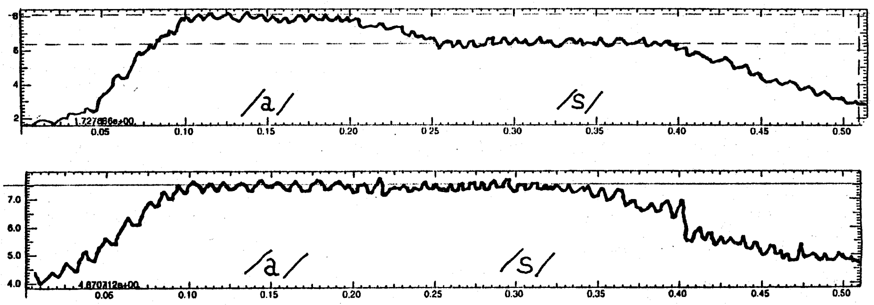

Figure 10 Energy contour of phonemes /a/, /s/

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 the 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 while taking account of this linguistic variation of the fundamental frequency. This is done by calculating a polynom approximation of the fundamental frequency contour as is shown in Figure 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 and 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, so the relative jitter values for high fundamental frequency segments are less reliable.

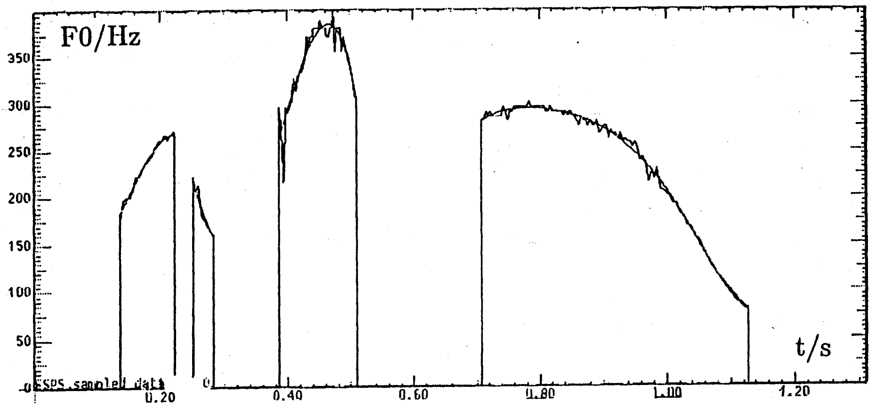

Figure 11 Polynom approximation of a fundamental frequency contour

EXPERIMENTAL PART

As mentioned in the introduction, voice quality can be interpreted as a physical marker which characterizes the speaker's personal vocal equipment or state of health and it can also be interpreted as a psychological marker which is influenced by the speaker's mood or affective state.

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

Recording of the database and recognition test

Ten 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 a 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 B&K measuring microphone. Each utterance was spoken several times in a neutral voice and several times with each of the following feigned emotions: 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 author and the 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 evalu
actor the selected sevent
was repeated three times
1 kHz tone between diff
were played to twenty na
were instructed to mar
overemphasized as 'unna
using eight categories: 'n
'boredom', 'disgust' and
content'. Only such sente
listeners were used for the
is extremely high. In a
recognition rate of 48 p
highest recognition rate
lowest rate was found f
1996).

The utterances used fo
evaluated as 'natural' and t

For speaker T forty-four
These were ten 'neutral'
sentences, ten 'angry' sent
spoken with fear. None c
above the threshold.

For speaker M forty-one
threshold of 80 per cent. T
sentences, three 'sad' se
sentences, four 'frightene
disgust.

For the female speaker
in the analysis. These were
(the other 'happy' utterar
sentences, ten 'angry' sent
spoken with fear. None o
above the threshold.

The utterances were pre
recognition rates were gene
low recognition rates for
expression of disgust cons
sentence spoken with a 'd
the utterances spoken w
threshold the last syllabl
sentences spoken with disg
low recognition rates for
fact, that vocal expressio
'sad' and 'bored' utteranc

content, but also to evaluate the naturalness of the utterances. For each actor the selected seventy sentences were randomized. Every sentence was repeated three times with three-second pauses and with a one-second 1 kHz tone between different emotions. The sentences from each actor were played to twenty naive listeners in separate sessions. The listeners were instructed to mark any utterances which sounded 'acted' or overemphasized as 'unnatural'. The emotional content was evaluated using eight categories: 'neutral', 'happiness', 'sadness', 'anger', 'fear', 'boredom', 'disgust' and 'unnatural or not recognizable emotional content'. Only such sentences recognized by at least 80 per cent 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 per cent for acoustic emotion portrayals. The highest recognition rate was found for 'anger' (78 per cent) and the lowest rate was found for 'disgust' (15 per cent) (Banse and 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 forty-four out of seventy sentences were used in the analysis. These were ten 'neutral' sentences, seven 'happy' sentences, three 'sad' sentences, ten 'angry' sentences, six 'bored' sentences and eight sentences spoken with fear. None of the sentences with 'disgust' were recognized above the threshold.

For speaker M forty-one out of seventy sentences reached the recognition threshold of 80 per cent. These were ten 'neutral' sentences, seven 'happy' sentences, three 'sad' sentences, eight 'angry' sentences, five 'bored' sentences, four 'frightened' sentences and four sentences spoken with disgust.

For the female speaker thirty-seven out of seventy sentences were used in the analysis. These were ten 'neutral' sentences, only one 'happy' sentence (the other 'happy' utterances were judged 'overemphasized'), four 'sad' sentences, ten 'angry' sentences, two 'bored' sentences and ten sentences spoken with fear. None of the sentences with 'disgust' were recognized above the threshold.

The utterances were preselected by the author 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 which were recognized above the threshold the last syllable ended in a 'retching' sound. Therefore the sentences spoken with disgust were not analysed any further. The relatively low recognition rates for sadness and boredom can be explained by the fact, that vocal expressions of sadness and boredom are quite similar, so 'sad' and 'bored' utterances were confused by the listeners.

Acoustic parameters

The acoustic parameters outlined above were used to analyse the emotional utterances. The shape of glottal pulses as well as spectral distribution of energy and the detection of harmonics vs. 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 because in the articulation of this vowel the supraglottal cavities are not expected to cause additional friction noise. The results reflect the type of phonation which 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 which 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.

Results

The experiment focuses on emotion-specific changes in voice quality. The speakers were not selected to represent a wide variety of different 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 inter-individual differences between the speakers do not stand out from the intra-individual differences within one emotion and hence are not discussed any further.

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 per cent. The type of phonation can therefore be classified as 'normal' modal voicing.

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 mainly 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 per cent, but, as the fundamental frequency reaches extremely high values at localized peaks, this also means that the systematic measuring error is high, since the

speech signals were sampl
classified as loud modal
fundamental frequency cc

Anger
The vowel energy is 1.5 t
also implies a greater loud
glottal pulses are very stee
utterances, voiced segments
bands. Spectral damping of
The physical energy in the
band, but considering the f
bands in angry speech sig
impression. As in happy u
higher than in the neutral
happy utterances, the highe
local fundamental frequenc
as shouted modal voicing v

Sadness
The maximum difference b
also means that the speech
looks very similar to a pu
(The speech signal itself do
is consistent with the energ
damping. For male speaker
but only turbulent compon
is determined by the low fr
energy in higher frequenc
often appears at the begin
Creak is equal to a relative
phonation in segments wit
breathy phonation.

Fear
The difference between vo
1.2 dB. This indicates a lim
very similar to the sad pulse
values. In contrast to sad u
in voiced segments, the utte
damping. This means that
and high band are very im
even the low band contains
frequency shows irregulari
of between 2 and 8 per ce

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.

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 bands 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.

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. 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 the sentence. Creak is equal to a relative jitter value of 40 or 50 per cent. The type of phonation in segments without voicing irregularities can be classified as breathy phonation.

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 which result in high relative jitter values of between 2 and 8 per cent. It should be mentioned that the utterances under investigation do not represent panic fear. They rather express whispery, conspicuous fear. 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 could indicate that fear induces narrow fricative articulation in contrast to the usually more open vowel articulation.

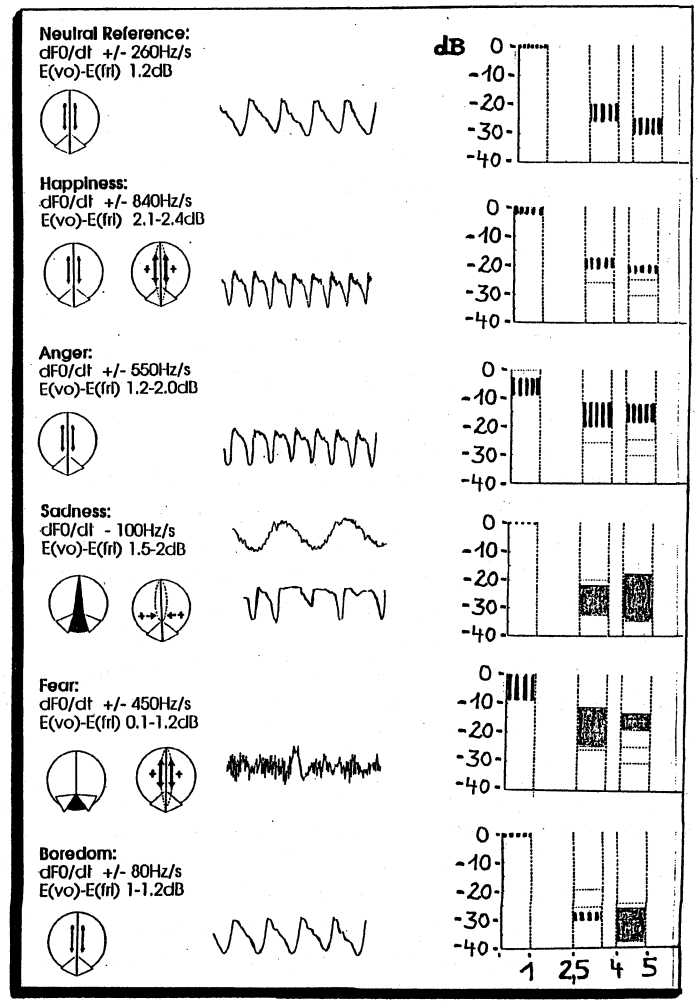

Figure 12 Results: Fundamental frequency and amplitude changes, type of phonation, pulse shapes and spectral distribution of energy for emotional voices

Boredom
The difference between vowe
indicates a moderate loudnes
function, as is the neutral mat
and high band is stronger thar
do have turbulent noise in th
relatively low energy these hi
perceived sound quality. In c
not show voicing irregularities
can be classified as 'normal'
fundamental frequency conto

The results are represented ir

SUMMARY

In the initial part of the paper s
were discussed with regard to
can be measured in the acoust
were introduced. To deal with
be regarded as quasi-stationa
using common algorithms fo
parameters are not orthogo
measuring similar phenomen
combination of these acousti
can be regarded as a reliabl
quality.

The experiment focuses on
quality parameters. The spea
variety of different personal v
were not more marked than
emotion.

Voice quality in 'neutral' ut
voicing for all speakers. In 'ha
as loud modal voicing with e
frequency contour. In angry utt
with strongly marked prosod
phonation. The type of phonat
appears at the beginning of
spoken with fear is (presur
fundamental frequency show
jitter values of about betwee
resembles sad speech. Voic
unambiguously recognized

Boredom

The difference between vowel and fricative energy is around 1.2 dB, which indicates a moderate loudness. The pulses are very similar to a sawtooth-function, as is the neutral material, though the spectral damping in the mid and high band is stronger than in the neutral utterances. 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; i.e. relatively monotonous speech.

The results are represented in Figure 12.

SUMMARY

In the initial part of the paper 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 introduced. To deal with the fact that some voice qualities can hardly be regarded as quasi-stationary signals and hence are difficult to analyse 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. The combination of these acoustic parameters allows plausibility checks, so it can be regarded as a reliable method for the characterization of voice quality.

The experiment focuses 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 inter-individual differences were not more marked than the intraindividual differences within one emotion.

Voice quality in 'neutral' utterances can be classified as 'normal' modal voicing for all speakers. In 'happy' utterances voice quality can 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 changes within the utterance. Creak appears at the beginning of voiced segments. Voice quality in sentences spoken with fear is (presumably) breathy or whispery falsetto. The fundamental frequency shows irregularities which result in high relative jitter values of about between 2 and 8 per cent. 'Bored' speech often resembles sad speech. Voice quality in those utterances that were unambiguously recognized as 'bored' speech do not show voicing

Boredom
The difference between vowel and fricative energy is around 1.2 dB, which indicates a moderate loudness. The pulses are very similar to a sawtooth-function, as is the neutral material, though the spectral damping in the mid and high band is stronger than in the neutral utterances. 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; i.e. relatively monotonous speech.

The results are represented in Figure 12.

SUMMARY

In the initial part of the paper 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 introduced. To deal with the fact that some voice qualities can hardly be regarded as quasi-stationary signals and hence are difficult to analyse 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. The combination of these acoustic parameters allows plausibility checks, so it can be regarded as a reliable method for the characterization of voice quality.

The experiment focuses 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 inter-individual differences were not more marked than the intraindividual differences within one emotion.

Voice quality in 'neutral' utterances can be classified as 'normal' modal voicing for all speakers. In 'happy' utterances voice quality can 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 changes within the utterance. Creak appears at the beginning of voiced segments. Voice quality in sentences spoken with fear is (presumably) breathy or whispery falsetto. The fundamental frequency shows irregularities which result in high relative jitter values of about between 2 and 8 per cent. '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.

The results of the experiment are relevant for forensic speaker identification, because emotional arousal can influence the speaker's voice quality. From the experiment the conclusion cannot be drawn that voice quality necessarily changes under emotional arousal, so the analysis of acoustic speech signals cannot serve as a reliable method to measure the speaker's affective state.

REFERENCES

Banse, R. and Scherer, K. R. (1996) 'Acoustic profiles in vocal emotion expression'. Accepted for publication in the *Journal of Personality and Social Psychology.*

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

Makhoul, J. (1975) 'Linear Prediction: A Tutorial Review', *Proceedings IEEE*, 63: 561–80.

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

Scherer, K. R. (1994) 'Affect Bursts', in S. M. H. van Goozen, N. E. van de Poll and J. A. Sergeant (eds), *Emotions: Essays on emotion theory*, Hillsdale, NJ: Erlbaum, 161–96.

Sundberg, J. (1994) 'Vocal fold vibration patterns and phonatory modes', STL-QPRS 2-3/94, KTH Stockholm.

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. and Gray, A. H. (1979) 'Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform', *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(4), 350–5.

Phonetic manifest
and physical stres
untrained police o

Marianne Jessen

University of Trier

ABSTRACT Two groups of subjects
were exposed to two types of psyc
activity (including memorizing event
precision shooting) and concentra
conditions and in two non-stress co
F0-mean, F0-standard deviation, an
F0 were determined both on the intr
F0-mean and intra-syllable F0-stan
psychological stress. Higher values f
unstressed speech and higher values
in stress management training leads
particular in the cognitive conditio

KEYWORDS psychological stre
frequency; syllable duration.

INTRODUCTION

The impact of psychological
phonetic manifestations have
(Scherer 1981; Hollien 1990:
parameters that were foun
psychological stress are the m
frequency (F0). Several other
levels of psychological stress
frequently in the literature or
tasks and other factors. The m
and timing, long-term spectral
and voicing irregularities.