VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice Habilitační a inaugurační spisy, sv. 229
ISSN 1213-418X

Milan Sigmund

VOICE ANALYSIS AND RECOGNITION

BRNO UNIVERSITY OF TECHNOLOGY
Faculty of Electrical Engineering and Communication

Doc. Ing. Milan Sigmund, CSc.

VOICE ANALYSIS AND RECOGNITION

ANALÝZA A ROZPOZNÁVÁNÍ HLASU

A THESIS OF A TALK FOR THE PROFESSORIAL APPOINTIVE
PROCEDURE IN THE STUDY FIELD OF
ELECTRONICS AND COMMUNICATIONS

VUTIUM

BRNO 2007

KEY WORDS

Speech signal processing, speaker recognition, voice analysis, speech under psychological stress, feature extraction, time domain, frequency domain, glottal excitation

KLÍČOVÁ SLOVA

Zpracování řečových signálů, rozpoznávání mluvčích, analýza hlasu, řeč pod psychologickým stresem, extrakce příznaků, časová oblast, kmitočtová oblast, hlasivkové buzení

ISBN 978-80-214-3396-0
ISSN 1213-418X

CONTENTS

Author’s introduction

Name:	Milan Sigmund, Doc., Ing., CSc.
Born:	1959 in Ivančice, Czech Republic

Education and degrees:

1975-79	Secondary School of Electrical Engineering, Brno
1979-84	Brno University of Technology
Ing. (M.Sc.)	Faculty of Electrical Engineering
	Thesis “Photopletysmograph for Measurement of the Heart Rate”
1986-89	Brno University of Technology
CSc. (Ph.D.)	Faculty of Electrical Engineering
	Thesis “Word Recognition Using Statistical Approach” (1989)
Doc. (Assoc. Prof.)	Brno University of Technology (2000)
	Faculty of Electrical Engineering and Communication
	Thesis “Speaker Recognition”

Professional experience:

1986 - 1989 Institute of Radio Electronics, FEE BUT Brno, Postgraduate Student
1990 - 1999 Institute of Radio Electronics, FEE BUT Brno, Assistant Professor
since 2000 Institute of Radio Electronics, FEEC BUT Brno, Associate Professor

Research activities:

Speech signal processing, speech segmentation, features extraction, word recognition.
Speaker recognition, voice analysis for diagnostics, detection of psychological stress.
Supervisor of 6 Ph.D. students (2 international: Dutch and US).

Publication activities:

Author 1 international monograph (published in Germany), author or co-author 5 study texts (3 international), 8 journal publications (3 international), 32 presentations on international conferences, 29 presentations on Czech and Slovak conferences, 14 special research reports (4 international).

International mobility and co-operation:

1994-95	UAS Wiesbaden, Department of Computer Science, Germany, 2 terms
1996-98	Scintilla AG, Solothurn, Switzerland, 2x 6 months
2001-03	UAS Wiesbaden, Department of Computer Science, Germany, 3 terms
2001-06	special lectures at UAS Pforzheim and TU Vienna
1997-2000	Council member of European Association for Education in Electrical and Information Engineering
since 2005	Member of Evaluation Board of the German Academic Exchange Service
since 2006	Member of Program Committee for the annual IASTED International Conference “Artificial Intelligence and Applications”

Teaching:

B.Sc. and M.Sc. lectures: Pulse and Digital Techniques, Signals and Systems, Speech Signal Analysis and Synthesis, Electronics in German

Ph.D. lectures: Speech Signal Processing for Speaker Recognition, Modern Digital Wireless Communications

1 INTRODUCTION

1.1 Speech Signal

Speech and language are tools that humans use to communicate or share thoughts, ideas, and emotions. Speech is talking, one way that a language can be expressed. Language may also be expressed through writing, signing, or even gestures. There are several ways of characterizing the communication potential of speech. According to information theory, speech can be represented in terms of its message content. An alternative way of characterizing speech is in terms of the signal carrying the message information, i.e. the acoustic waveform.

A central concern of information theory is the rate at which information is conveyed. For speech, this rate is given by taking into consideration the fact that physical limitations on the rate of motion of the articulators require that humans produce speech at an average rate of about 10 sounds (phonemes) per second. Assuming a six-bit numeric code to represent all the phonemes and neglecting any correlation between pairs of adjacent phonemes, we get an estimate of 60 bit/sec for the average information rate of speech. In other words, the written equivalent of speech contains information equivalent to 60 bit/sec. The above estimate does not take into account such factors as emotional state of the speaker, the loudness of the speech, etc. Although information theory ideas have played a major role in sophisticated communications systems, we will consider throughout this work the speech representation based on the acoustic signal, which has been most useful in practical applications.

As shown in Fig. 1.1, these representations can be classified in two broad groups, namely waveform representations and parametric representations. Parametric representations are concerned with representing the speech signal as the output of a model for speech production. The parameters of this model are conveniently classified as either excitation parameters (i.e. related to the source of speech sounds) or vocal tract response parameters (i.e. related to individual speech sounds).

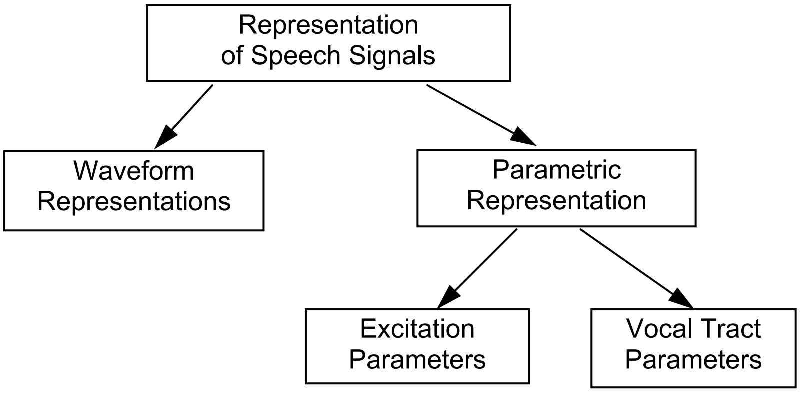


Fig. 1.1 Acoustic representations of speech signals.

Successful speech signal processing requires knowledge and expertise in a wide range of disciplines such as digital signal processing, physics (acoustics), pattern recognition, information theory, linguistics, computer science, physiology, etc. The basis of all speech processing tasks is speech analysis. On the other hand, all sub-fields of speech processing can be thought of as self-contained problems. They have reached a certain maturity and found their position on the market but a number of problems have remained unsolved.

1.2 Some History of Speech Processing

The history of speech processing is very old. Already in ancient Greece priests used a simple mechanical synthesizer in a statue's mouth to produce signals resembling speech. Research into automatic speech processing by machine has been done for almost five decades. To be able to appreciate for the amount of progress achieved over this period, it is worthwhile to briefly review some research highlights. Regarded as significant milestones in speech research can be the following:

1779 - mechanical production of five vowels (Ch. G. Kratzenstein)
1791 - mechanical speech model generating intelligible utterances (J. W. Kempelen)
1876 - acoustic signal was transmitted by telephone (A. G. Bell)
1918 - empirical model of human speech recognition was devised (H. Fletcher)
1939 - a speech synthesizer (Voder) was demonstrated at the World Fair in New York
1934 - the term "high fidelity" began to be used
1937 - pulse-code modulation (PCM) was invented (H. Reeves)
1948 - digital methods of spectrum estimation were introduced (M. Bartlett)

The 1950s

- earliest attempts to devise systems for automatic speech recognition by machine
- speech research was focused only on English

1952 - the first formant synthesizer was presented at a conference in London (G. Fant)
1952 - a system for isolated digit recognition for single speaker was built at Bell Labs
1959 - a speaker-independent vowel recognizer for 10 vowels was constructed at MIT

The 1960s

- special-purpose hardware for speech recognition was developed and the first companies were founded which built, marketed and sold speech recognition products
- speech research spread word-wide to include also other languages; significant Japanese and Russian studies were published

1962 - speaker recognition by voiceprint analysis was begun (L. Kersta)
1966 - the application of linear prediction (LP) to speech was formulated (F. Itakura)
1968 - the dynamic programming for time alignment was proposed (T. Vintsyuk)

The 1970s

- isolated word recognition was the key focus of research in this decade
- Japanese research showed how dynamic programming methods could be successfully applied and how to use an appropriate distance measure based on LP spectrum

1971 - the Markov model of word was defined (T. Vintsyuk)
1972 - the PARCOR technique of linear prediction was presented (Saito and Itakura)
1975 - line spectrum representation was proposed as a modification of LP (F. Itakura)

The 1980s

- a decade in which personal computers became ubiquitous (in 1981 IBM introduced the first personal computer), in the early 1980s there appeared quite a few single-chip DSPs (Intel, NEC, Texas Instruments)
- speech research was characterized by a shift in technology from template-based approaches to statistical modeling methods
- work on neural networks was motivated by the wish to emulate pattern recognition by human brain

1980 - mel-cepstrum as a parameter of speech was used (Davis and Mermelstein)
1988 - the European Speech Communication Association (ESCA) was established

The 1990s

- the decade of a new era in interconnection (multimedia PCs, laptop, Internet, cordless telephone)
- importance of biometrics is growing enormously

1990 - perceptual linear predictive (PLP) speech analysis was introduced (N. Morgan and H. Hermansky)
1993 - specialized journal, *Speech and Audio Processing*, began to be published
1994 - relative spectra (RASTA) processing of speech was introduced (H. Hermansky)
1995 - the *European Language Resources Association* (ELRA) was established
1999 - ESCA became a truly international association in the global field of speech science changing its name to ISCA (*International Speech Communication Association*)

The 2000s

- the years of an enormous spreading of mobile phones (new multimedia services, significant quality improvements at low bit rates ≤ 8 kbit/s)
- singing and expressive speech (including happiness, sadness, anger, fear etc.) synthesis is commercially available
- techniques for speech classification into emotional and pathological states are developed

2003 - mixture of hidden Markov models for emotion classification was presented (Fernandez and Picard)
2006 - perceptual evaluation of audio quality (PEAQ) was defined (Huber and Kollmeier)

1.3 Voice Recognition by Humans

People can reliably identify familiar voices. About 2-3 seconds of speech is sufficient to identify a voice, although performance decreases for unfamiliar voices. One review of human speaker recognition [1] notes that many studies of 8-10 speakers (work colleagues) yield in excess of 97% accuracy if a sentence or more of the test speech is heard. Performance falls to about 54% when duration is shorter than 1 second and/or distorted e.g., severely high pass or low pass filtered. Performance also falls significantly if training and test utterances are processed through different transmission systems. A study using voices of 45 famous people in 2 sec test utterances found only 27% recognition in an open-choice test, but 70% recognition if listeners could select from six choices [1]. If the utterances were increased to 4 seconds, but played backward (which distorts timing and articulatory cues), the accuracy resulted to 57%. Widely varying performance on this backward task suggested that cues to voice recognition vary from voice to voice and that voice patterns may consist of a set of acoustic cues from which listeners select a subset to use in identifying individual voices.

Recognition often falls sharply when speakers attempt to disguise their voices e.g., 59-81% accuracy depending on the disguise vs. 92% for normal voices [2]. This is reflected in machines, where accuracy decreases when mimics act as impostors. Humans appear to handle mimics better than machines do, easily perceiving when a voice is being mimicked. If the target (intended) voice is familiar to the listener, he often associates the mimic voice with it. Certain voices are more easily mimicked than others, which lends further evidence to the theory that different acoustic cues are used to distinguish different voices.

1.4 Applications of Automatic Voice Recognition

Voice recognition is the general term used to include all of the many different applications of discriminating people based on the sound of their voices. The area of voice recognition can be divided into five distinct sub-fields, as shown in Fig. 1.2 .

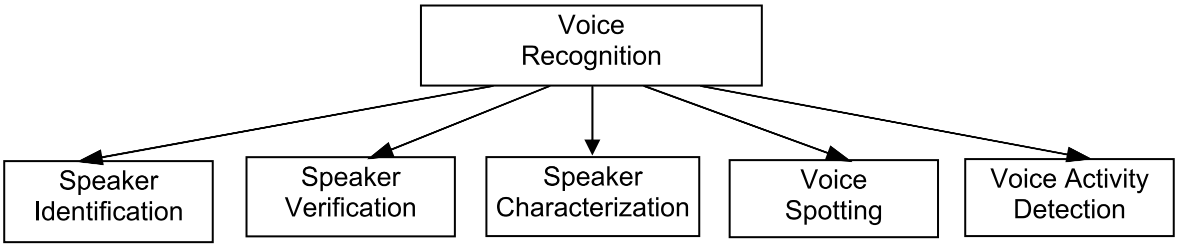


Fig. 1.2 Areas of application for voice recognition.

Speaker Identification aims to identify a speaker who belongs to a group of users through a sample of his speech. **Speaker Verification** aims to verify the identity of the speaker through a comparison of some samples of his speech with the references of the speaker he claims to be. One of the most obvious use of speaker verification techniques is caller authentication over the telephone network. The user claims his identity, most of the time by dialling (or saying) a personal code number. Then, either a code word is required to authentify the speaker, or his utterance of the code number is used for verification purposes. The main applications for speaker verification over the telephone network are twofold: the first kind is home banking, the second kind is access to licensed databases. It is obvious that both fields do no require the same level of security; it is usually less costly to let someone unauthorized have access to a database, than to allow somebody to operate some transaction that can involve substantial amounts. However, in many countries, it is possible to pay by credit card over the phone, without any other verification of the customer's identity than the consistence between the customer's name, his credit card number and its expire date (all of them being on the credit card itself).

The sub-area **Speaker Characterization** includes some specific applications as selection of sex, age, geographical provenience, and other demographic factors. The age estimation of unknown speaker's voice recorded from telephone calls is one of the most frequent tasks in speaker profiling. Other tasks of selection of speech patterns according to specified characteristics of the speaker are proposed by the technology providers to detect the speaker's current emotional state (mood state identification) and any pathologies (health state identification) using speech samples. Several investigations have shown that voice quality can change under the influence of smoking or the use of alcohol. One of the most common consequences of smoking and drinking is premature ageing of the mucous membrane covering the vocals muscle, resulting in a hoarse voice quality. Excessive drinking may eventually result in brain damage, which may in turn lead to severe speech disorders. The use of drugs can have a similar effect. In those cases it would be more appropriate to speak of pathological speech. A profitable direction for speaker characterization would be so as to focus research on idiosyncratic voice differences.

In combination with other technologies, further areas of application for voice recognition include aids for the disabled persons, learning technologies, virtual conferencing and virtual impersonation.

2 SPEAKER RECOGNITON

2.1 Ideal Speaker Recognition

The purpose of the discussion about an ideal voice recognition system is to find the theoretical limits of voice recognition performance when all the practical restrictions are lifted. An ideal system should be unaffected by processes as follows:

- changes in the speaker physical state (e. g. illness, cold),
- changes in the speaker emotional state (e. g. stress, onset of anger),
- changes in the speaker voice due to aging of the speaker,
- utterance variations (e. g. fast talking versus slow talking rates),
- noise etc.

While ideal systems should be the ultimate goal of voice recognition systems, there are practical considerations that make the achievement of this goal difficult. The uniqueness of an individual's voice is a consequence of both the physical features of the person vocal tract and the person mental ability to control the muscles in the vocal tract. The physical features of an individual vocal tract consist of the overall length of the tract, the height and width of the tract at different positions and the size and shape of the tongue, teeth and lips. The density of the tissue in the vocal tract also affects the sounds that the individual can produce. The physical dimensions of a vocal tract determine the range of possible sounds that can be made. It is not easy for an individual to change voluntarily these physical characteristics. However, they may change somewhat with ageing. An ideal voice recognition system would use only physical features to characterize speakers, since these features cannot be easily changed. However, it is obvious that investigators cannot simply measure the vocal tract dimensions of an unknown speaker. Thus, numerical values for physical features or parameters would have to be derived from digital signal processing parameters extracted from the speech signal. Using this strategy, a comparison of voices can be carried out as follows: the physical parameters of known speakers are determined either by processing shown in Fig. 2.1 or by using physical measuring devices, e.g. X-ray.

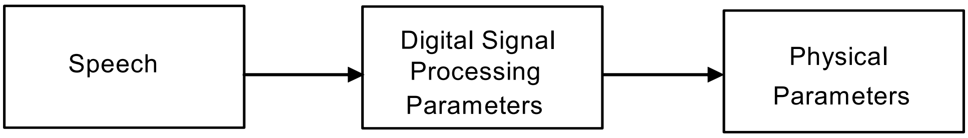


Fig. 2.1 Ideal parameter extraction.

Some signal processing parameters are as follows:

- fundamental frequency,
- formants frequency,
- cepstral coefficients,
- spectral moments.

Some signal physical parameters are as follows:

- vocal tract length, width and breadth,
- size of tongue,
- size of teeth,
- tissue density.

An ideal system would also be able to compensate for the common changes of physical parameters due to ageing of the speaker or temporary changes in physical condition due illness or stress. As an example, Fig. 2.2 illustrates the relationship between fundamental frequency of speech (i.e., DSP parameter) and membranous length (i.e., physical parameter). The fundamental frequency is scaled primarily according to the membranous length of the vocal folds, whereas mean airflow, sound power, glottal efficiency, and amplitude of vibration include another scale factor that relates to overall larynx size. There was predicted an inverse relationship between fundamental frequency F_0 and membranous length L_m with fixed tension and fixed mass per unit length. The hyperbola has the form [3]

$$F_0 = 1700 / L_\mathrm{m} \tag{2.1}$$

where L_m is in mm. For example a fundamental frequency of 170 Hz corresponds with adult female membranous length L_m= 10 mm.

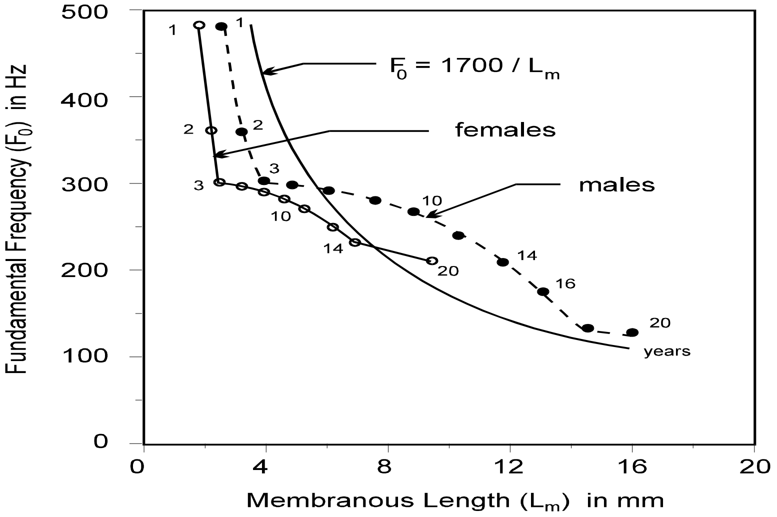


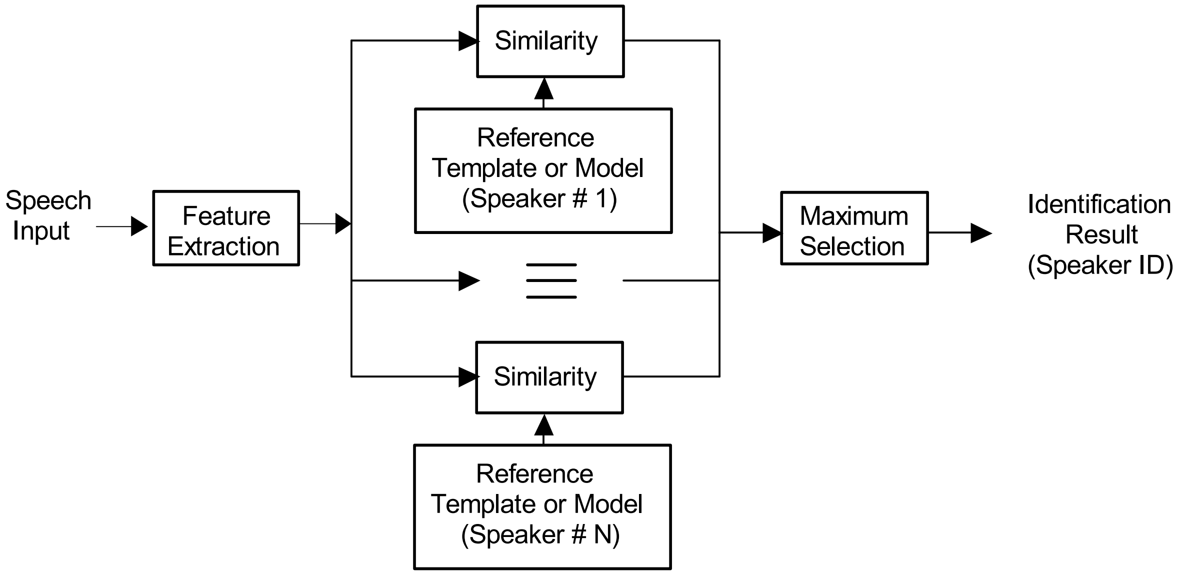
Fig. 2.2 Mean speaking fundamental frequency F_0 as a function of membranous length L_m .

Since many independent, continuously valued physical parameters of the vocal tract exist, it is unlikely that two speakers, even if they sounded very similar to each other, would have the same values for all parameters. Suppose that vocal tracts could be effectively represented by 10 independent physical features, with each feature taking on one of 10 discrete values. If the vocal tract could be modelled that accurately, then 10^{10} individuals in the population (i.e., 10 billion) could be distinguished. Today's world population amounts to approximately 6 billion ($6 \cdot 10^9$) individuals.

Because of the complexity of speech signals, it is not possible to perform the transformation from signal processing parameters to physical parameters with the accuracy necessary for speaker recognition. As a compromise, the signal processing parameters themselves are currently used in non-ideal voice recognition systems.

2.2 Principles of Speaker Recognition

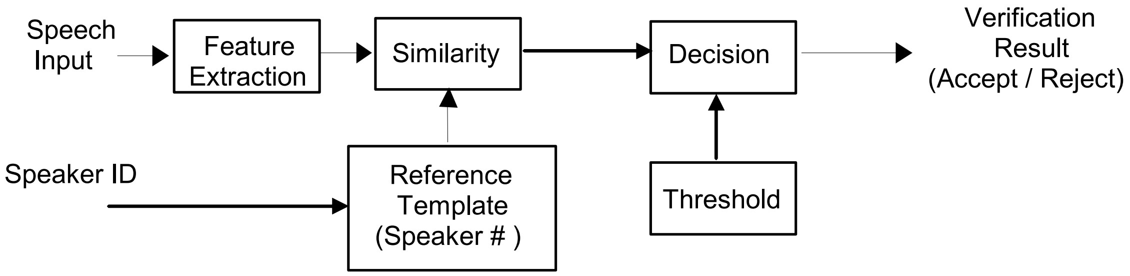
Speaker recognition covers two main different areas: speaker identification and speaker verification. In speaker identification, a speech utterance from an unknown speaker is analyzed and compared with models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. Figure 2.3 shows the structure of speaker identification system.



Speech Input
Feature Extraction
Similarity
Reference Template or Model (Speaker # 1)
Similarity
Reference Template or Model (Speaker # N)
Maximum Selection
Identification Result (Speaker ID)

Fig. 2.3 Basic structure of speaker identification system.

Speaker verification aims to verify the identity of the speaker through a comparison of some samples of his speech with the references of the speaker he claims to be. If the match is above a certain threshold, the identity claim is verified. A high threshold makes it difficult for impostors to be accepted by the system, but at the risk of rejecting the genuine person. Conversely, a low threshold ensures that the genuine person is accepted consistently, but at the risk of accepting impostors. Figure 2.4 shows the structure of speaker verification system.



Speech Input
Feature Extraction
Similarity
Decision
Verification Result (Accept / Reject)
Speaker ID
Reference Template (Speaker #)
Threshold

Fig. 2.4 Basic structure of speaker verification system.

Speaker recognition is one area of artificial intelligence where machine performance can exceed human performance - using short test utterances and a large number of speakers, machine accuracy often exceed that of humans. This is especially true for unfamiliar speakers, where the training time for humans to learn a new voice well is very long compared with that for machines.

2.3 Evaluation of Parameters

The current most commonly used short-term speech parameters for speaker recognition are LPC-derived cepstral coefficients [4] and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed form LPC coefficients [5] and therefore provides a steadier representation from one repetition to another of a particular speaker's utterances. As for the regression coefficients, typically, the first- and second-order coefficients, that is, derivatives of the time functions of cepstral coefficients [6] are extracted at every frame period to represent spectral dynamics (the delta- and delta-delta-cepstral coefficients).

The goal of parameter evaluation should be to determine the smallest set of parameters which contain as much useful information as possible. The penalties for choosing parameters incorrectly include poor recognition performance, excessive processing time and storage space. Typical voice recognition systems use a set of parameters that may be represented by a feature vector

$$\mathbf{x} = [x_1, x_2 \ldots x_N] \ ,$$

where x_1, x_2, etc. are individual features. The same parameters are calculated at different time positions in an utterance. One common measure of effectiveness for individual features is called the F-ratio, which compares inter- and intra-speaker variances:

$$F = \frac{Variance\ of speaker\ means}{Mean\ intraspeaker\ variance} \tag{2.2}$$

The F-ratio for each feature n can be determined as follows

$$F_n = \frac{\dfrac{J}{I-1}\sum_{i=1}^{I}\left(S_{i,n} - U_n\right)^2}{\dfrac{1}{J-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{i,j,n} - S_{i,n}\right)^2} \quad , \tag{2.3}$$

where $x_{\mathrm{i,j,n}}$ is the value of the n-th feature for the i-th speaker during the j-th frame. If J vectors have been collected for each of I number of speakers, then $S_{i,n}$ estimates the value of the n-th feature for the i-th speaker.

$$S_{i,n} = \frac{1}{J}\sum_{j=1}^{J} x_{i,j,n} \tag{2.4}$$

The average of the n-th feature over all frames of all speakers is represented by

$$U_n = \frac{1}{I}\sum_{i=1}^{I} S_{i,n} \tag{2.5}$$

Features with larger F-ratios will be more useful for voice recognition. The F-ratio tends to be high for features for which one or two speakers are very different from the rest, which suggests that F-ratios are most useful in eliminating poor features rather than choosing the best. However, F-ratios are only valid for the set of data from which they were calculated. Features that appear to be useful for one set of speakers may be worthless for another set of speakers.

3 EFFECTS OF PSYCHOLOGICAL STRESS ON SPEECH

3.1 Emotional Speech

Spoken language comes from our inside. Factors such as mood, emotion, physical characteristics and further pragmatic information are contained in speech signals. Many of these characteristics are also audible. An emotional speech with high content differs in some parameters from a normal speech [7]. In recent years, the interest for automatically detection and interpretation of emotions in speech has grown and vocal emotions have also tended to be studied in isolation. About 25% of information contained in a clean speech signal refer to the speaker. These phonologically-linguistically irrelevant speaker characteristics make speech recognition less effective but can be used for speaker recognition (ca. 15% of information) and analysis of the speaker's emotional and health state (ca. 10% of information).

In today's hectic and fast-moving society, stress is more or less present in all professions. Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity and deterioration of performance. Stress may be induced by external factors (noise, vibration, lack of sleep, etc.) and by internal factors (emotion, fatigue, etc.). Physiological consequences of stress are among others changes of the heart rate and respiratory, (e.g., increased respiration rate, irregular breathing), musculature changes (increased muscle tension), etc. The increased muscle tension of vocal cords and vocal tract may, directly or indirectly, have an adverse effect on the quality of speech. The entire process is extremely complex, and is shown in a simplified model in Fig. 3.1. The accepted term for speech signal carrying information on the speaker's physiological stress is "stressed speech".

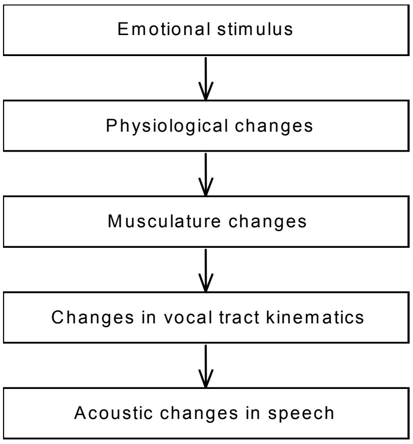


Fig. 3.1 Model how emotion causes changes in speech.

With increasing demand for speech technology systems, there is an increasing need for processing of emotion and other pragmatic effects (simulation in synthetic speech, elimination in robust speech recognition). In some cases, it is very important to detect the emotional state of a person (e.g., stress, fatigue or use of alcohol) from his/her voice.

3.2 Stressed Speech Data

There are not many corpora designed to allow the study of speech under stress. It is extraordinary difficult to obtain realistic voice samples of speakers in various states of stress, recorded in realistic situations. "Normal" people (as well as professional actors) cannot simulate stress perfectly with their voices. However, there are approaches how simulate stress by using: vocal noises, quick question-answer-quizzes with the possibility to win something, business negotiation about an important contract, etc.

A typical corpus of extremely stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. Such speech signals together with other corresponding biological factors are collected for example in the NATO corpus RSG10. On the other hand, such extreme situations occur seldom in everyday life. The mostly mentioned corpus in the literature is the SUSAS (Speech Under Simulated and Actual Stress) database of stressed American English described in [8] and distributed by Linguistic Data Consortium at the University of Pennsylvania. For the French speech, the Geneva Emotion Research Group at the University of Geneva conducts research in many aspects of emotions including stress and also collected other emotion databases. German database of emotional utterances spoken by actors was recorded at the Technical University Berlin. A complete description of the database called "Berlin Database of Emotional Speech" can be found in [9]. Most of the few studies reported in literature concern with English. For the Czech language, no research in emotional speech is known and no appropriate available database exists.

3.3 Database ExamStress

However, for our studies conducted within the research of speech processing in stress we used our own database. The most suitable situation with realistic stress took place during the final state examinations at Brno University of Technology held in oral form in front of a board of examiners. The test speakers were 31 male pre-graduated and post-graduated students. The created database ExamStress [10] consists of two kinds of speech material: "Defence" and "Pre-Defence" and is continuously completed. Main features of the database are summarized in Table 3.1.

Table 3.1 Description of the ExamStress database.

Database part	Defence	Pre-Defence
Stressor	Exam nerves, fatigue	Exam nerves, fatigue
Language	Czech	Czech
Speech type	Spontaneous, read sentences	Read sentences
Total length	180 minutes	30 minutes
Speakers	31	19
Gender	Male	Male
Occupation	Student	Student
Microphone	C 417	AKG
Quantization	16-bit linear	16-bit linear
Sampling rate	22 kHz	22 kHz
Quality	Good	Very good

The Defence Part
The speech data were collected during state exams. This material contains stressful phrases (improvisations relating to unknown technical problems) and other phrases with lower stress (during discussions relating to known technical problems). The recorded utterances were manually examined (including both examination of the waveform and parameter contour, and listening). In this way, a number of pauses and irrelevant extraneous voices were eliminated. Furthermore, short portions of 3 to 5 minutes of fluent stressed speech were selected, cut out and written down. A few days later, the same speakers read those written texts. In both states, the same text was used (spontaneous or read speech).

The Pre-Defence Part
The speech data were collected 10 minutes before the state exam started. All speakers were asked to read the same text with an approx. length of 1 minute (at a normal speaking rate). This material was also expected to contain stressful phrases (mostly less stressful than during the defence). A few days later, the same speakers read this text again.

3.4 Subjective and Objective Stress Classification

It is possible to measure stress objectively and can the speech signal be used as a possible indicator of stress? It can be expected that a student defending his diploma or doctoral thesis is under examination nerves. While in case of non-stressed databases, the distinguishing among different phonemes or words is quite simple, in case of stress-databases, determining the stress-level of speech data is usually very difficult. To estimate the level of the student's stress subjective listening tests and objective heart rate measuring were done.

For listening tests, three stress-levels were established (1 – low, 2 – medium, and 3 – high stress-level). Results of a non-professional listening test for 24 listeners and short speech portions selected for 16 speakers are summarized in Table 3.2. The value in the column Stress-level is a weighted average of stress-level for all listeners and the value in the column Std is the standard deviation of the stress-level.

Table 3.2 Results of non-psychologists listening test.

Speaker	Stress level	Std	Speaker	Stress level	Std
M1	2.36	0.48	M10	2.64	0.61
M2	2.71	0.45	M11	1.79	0.77
M3	1.29	0.59	M12	2.29	0.45
M4	1.71	0.70	M13	1.93	0.80
M5	2.00	0.65	M14	2.07	0.59
M7	2.00	0.76	M15	1.21	0.41
M8	1.07	0.26	M19	1.93	0.59
M9	1.21	0.41	M22	1.36	0.48

In some cases the heart rate *HR* of students was measured simultaneously with the speech recordings in both, stressed and normal state. A comparison of these measured data proves the influence of exam nerves on the speaker's emotional state. The oral examination seems to be a reliable stressor. Figure 2 shows a typical run of the *HR* curve in the first 900 seconds after the start of the examination. As expected, the *HR* is the highest at the beginning and then slowly decreases. On average the *HR* values obtained for stressed state was near doubled compared to the normal state (such values usually occur if a person is under medium physical activity). Surprising is the rapidly short-time increases at the beginning of recordings in normal state. Probably, this represents an initial effect of "stress due to the attached measuring equipment" which is similar to the well-known "stress for physicians".

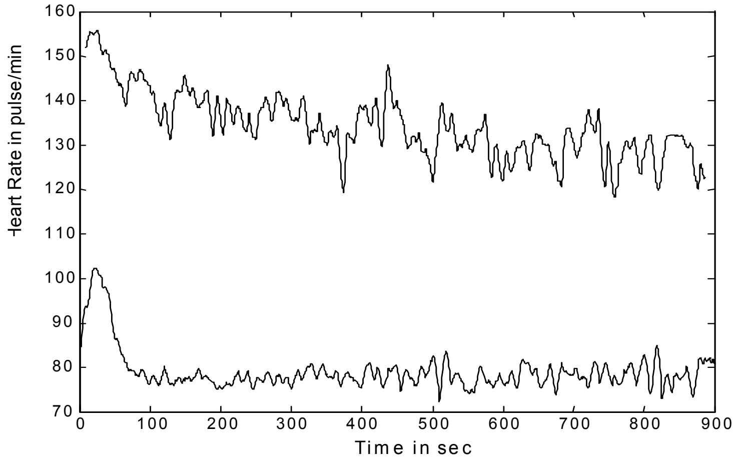


Fig. 3.2 Typical heart rate during the exam (upper curve) and in normal state (lower curve).

3.5 Analysis in Time and Frequency Domains

The often used methods for identifying stress and other emotional states usually start from the time distribution of single phonetic parts of words or sentences. Speech influenced by psychological stress can be identified e.g. by different time lengths of the concrete phonemes or by different time lengths of speech pauses between words. Duration analysis conducted across individual Czech vowel phonemes [11] shows the main difference in the distribution of vowel "a". By contrast, the small differences in the distribution of vowel "e" seem to be irrelevant for the detection of emotional stress (Fig. 3.3).

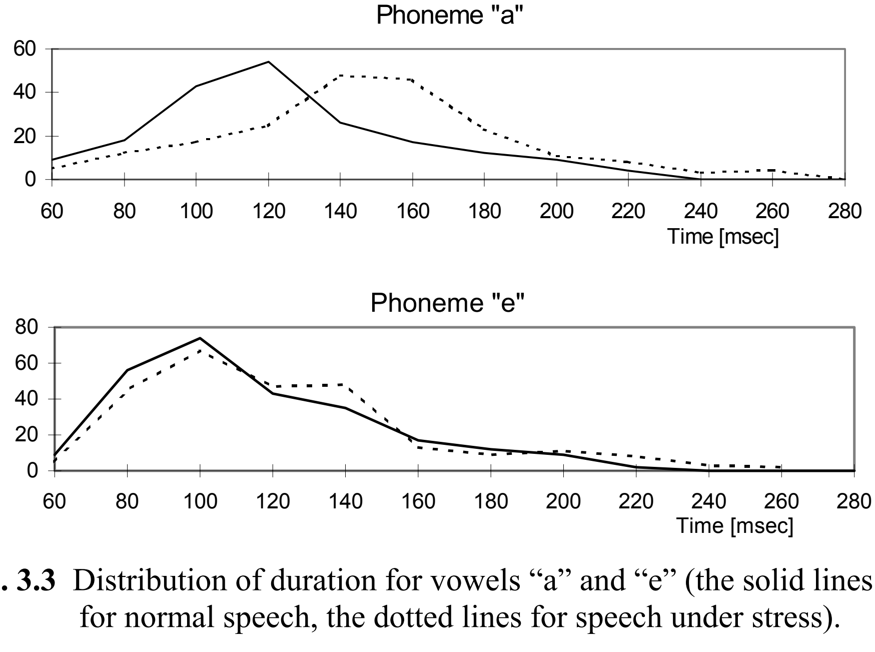


Fig. 3.3 Distribution of duration for vowels "a" and "e" (the solid lines are for normal speech, the dotted lines for speech under stress).

In general, more significant results are given by formants in the spectral domain. The analysis of vocal tract spectrum focused on formant positions F_i and formant bandwidths B_i for selected vowel phonemes shows that only changes in the first and the second formants are significant. In stressed speech, both low formants F_1 and F_2 were shifted to higher frequencies as a rule. Table 3.3 illustrates the average formant values for the Czech phoneme “i“.

Table 3.3 Formant changes in spectrum for phoneme “i” (all in Hz).

	F_1	B_1	F_2	B_2	F_3	B_3	F_4	B_4
Normal	409	52	1981	218	2630	489	3356	371
Stressed	525	98	2068	142	2672	462	3347	383

Statistical evaluation was also used to examine the distribution function of the pitch period. The fundamental frequency F_0 contours were calculated on the frame-by-frame basis using the center-clipping autocorrelation method [12]. From this information the distribution of F_0 values was obtained separately for the stressed and normal speech, and the mean F_0 values and standard deviations were calculated. In all cases, the average fundamental frequency increased and the range of fundamental frequency enlarged when the speaker was involved in a stressful situation [13]. Figure 3.4 shows the F_0 distribution for a typical young male speaker. The curves are comparable because they were obtained from speaking the same text.

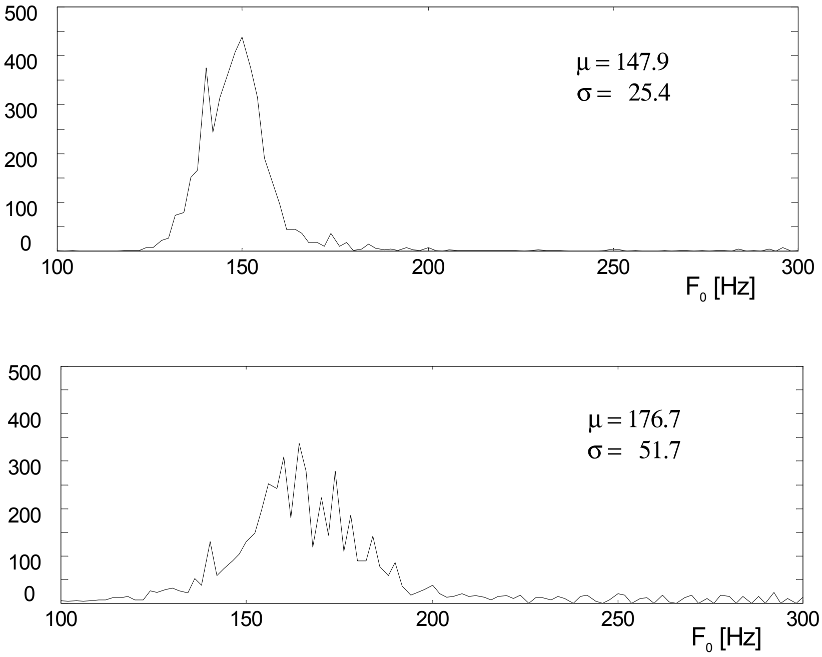


Fig. 3.4 Pitch distribution for a male speaker (upper graph is for normal speech, lower graph is for speech under stress).

The effect of changes in speech due to the emotional state of speaker on long-time spectrum [14] can be observed in Fig. 3.5. The dashed line gives the spectrum of emotional speech spoken under stress, the solid line gives the spectrum obtained from the same text read in normal state of speaker and the dotted line also gives the spectrum from the same text read by a tired speaker. Thus in all three cases the identical speech was spoken by one speaker in various states of mind. The psychological state (stress) affects the spectrum more than the physical state (fatigue).

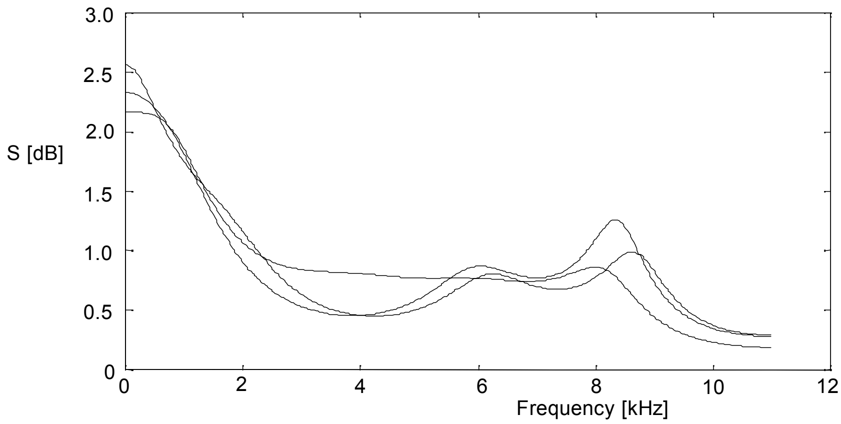


Fig. 3.5 Long-time spectrum variability within speaker for normal and emotional speech (LPC order 8, speech duration 114 sec).

In our experiments we also focused on the cepstral analysis within the vowel class. The first 12 standard mel-cepstral coefficients [15] *mcc*(1) to *mcc*(12) were estimated for all individual basic Czech vowels cut out from a speech spoken normally and under stress. Finally, the same coefficients obtained from corresponding vowels were compared. The most effective indicator of stress seems to be the 9th mel-cepstral coefficient computed from the vowel "u". Table 3.4 shows the mean values of *mcc*(9) obtained for three various speakers. This indicator gives a higher value in case of stressed speech.

Table 3.4 Mean values of the 9th mel-cepstral coefficient for the vowel "u" (N denotes normal speech and S is for stressed speech).

Speaker M1						
	Test 1		Test 2		Test 3	
Speech	N	S	N	S	N	S
mcc(9)	-0.191	-0.096	-0.206	-0.150	-0.112	0.029
Speaker M2						
	Test 1		Test 2		Test 3	
Speech	N	S	N	S	N	S
mcc(9)	-0.152	0.160	-0.101	-0.025	-0.024	0.002
Speaker M5						
	Test 1		Test 2		Test 3	
Speech	N	S	N	S	N	S
mcc(9)	-0.157	-0.094	-0.177	-0.107	0.110	0.187

3.6 Analysis of Glottal Excitation

Glottal source estimation has great potential for use in identifying of emotional states, non-invasive diagnosis of voice disorders, etc. Voiced speech is typically modeled as the output of a linear and time-invariant filtering process. A quasi-periodic glottal flow signal at the glottis, denoted by $g(n)$, acts as an acoustic source and excites the vocal tract filter of impulse response $h(n)$. The output speech pressure waveform measured in front of the lips can be expressed as convolution

$$s(n) = g(n) * h(n) \tag{3.1}$$

For the glottal pulse estimation, several techniques can be found e.g. in [16]. The well known and effective method for analysis of the estimated glottal pulses is the Liljencrant-Fant's (LF) approximation [17]. The LF model can also be used for speech signal synthesis. Using the parameters of this model it is possible to imitate the voice of a specific person. Some parameters of glottal pulses, obtained by the LF model, are especially suitable for emotional speaker state investigation. Glottal pulse approximation using the LF model is illustrated in Fig. 3.6.

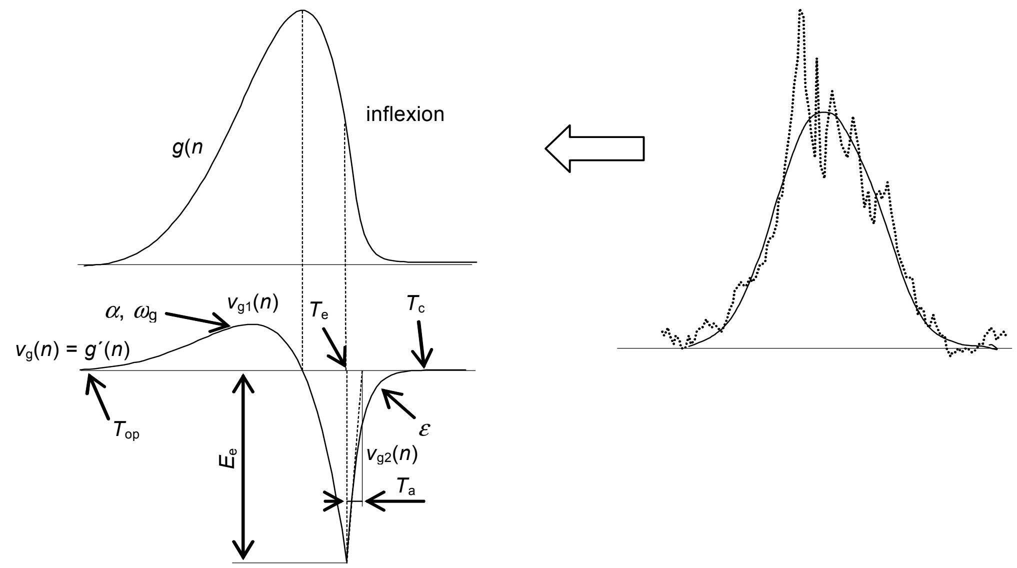


Fig. 3.6 Liljencrant-Fant's approximation of glottal pulse and its derivative (lower graph).

The first derivative $v_\mathrm{g}(n)$ of the approximation function $g(n)$ consists of two parts $v_{\mathrm{g}1}(n)$ and $v_{\mathrm{g}2}(n)$ described by Eq. (3.2) and Eq. (3.3).

$$v_{\mathrm{g}1}(n) = -E_\mathrm{e} \frac{\sin\left[\omega_\mathrm{g}\left(n - T_\mathrm{op}\right)\right]}{\sin\left[\omega_\mathrm{g}\left(T_\mathrm{e} - T_\mathrm{op}\right)\right]} e^{\alpha(n - T_\mathrm{e})} \qquad \text{for } \; T_\mathrm{op} \le n \le T_\mathrm{e} \tag{3.2}$$

$$v_{\mathrm{g}2}(n) = \frac{-E_\mathrm{e}}{\varepsilon T_\mathrm{a}} \left[e^{\varepsilon(T_\mathrm{e} - n)} - e^{\varepsilon(T_\mathrm{e} - T_\mathrm{c})}\right] \qquad \text{for } \; T_\mathrm{e} < n < T_\mathrm{c} \tag{3.3}$$

Variables T_op, T_e, T_c and T_a are important time values and their meaning can be clear from Fig. 3.6. Approximation is limited to the time interval $T_\mathrm{op} \le n \le T_\mathrm{c}$. The remaining variables E_e, ω_g, α and ε are the LF parameters sought. It is possible to obtain them by some of the iterative methods. The parameters are determined by criteria of the minimal average quadratic deviation of the approximating and the approximated function.

The LF approximation was applied to speech data recorded in normal and stressed state of the speaker. Records of both states were phonetically identical. For each phoneme, selections of ten sets containing six segments were created. The obtained results show that only some of the LF parameters E_e, ω_g, α and ε are suitable for stress detection [18]. Results for two phonemes spoken by one male speaker are plotted in the diagrams in Fig. 3.7.

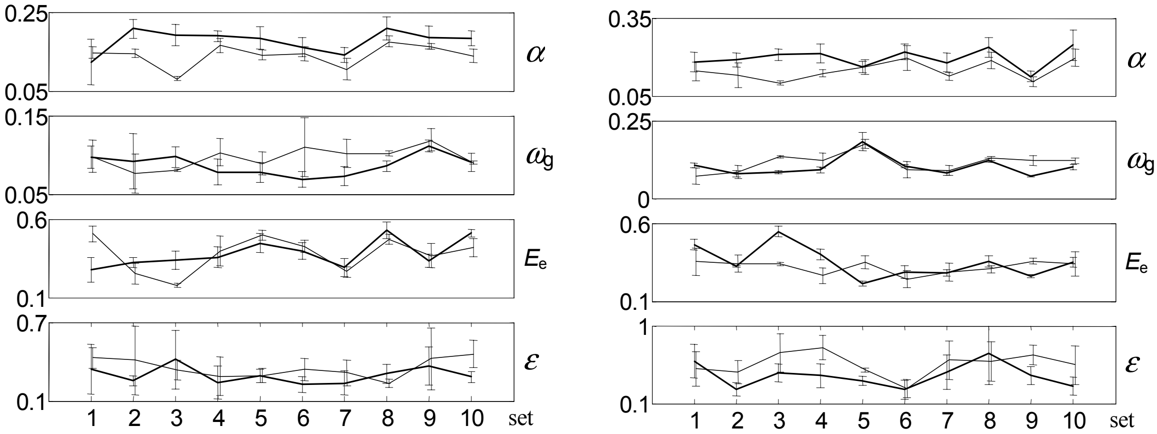


Fig. 3.7 Average values of the LF parameters and their scatter for sets of phonemes "a" (left) and "e" (right). Bold lines represent normal state.

In our statistical experiments, the glottal pulses were obtained from the speech by applying the IAIF (Iterative Adaptive Inverse Filtering) algorithm which is one of the most effective techniques for extracting excitation from a speech signal [19]. Figure 3.8 shows a typical time waveform $s(n)$ of the vowel "a" and it corresponding glottal pulse derivative $v(n)$ estimated using IAIF. In order to minimize the influence of voice intensity (i.e. loud vs. soft voice), the amplitude normalization was used before applying the IAIF-procedure. For the analysis, a pitch synchronous selection of segments from the obtained glottal pulse waveform was used. A position determining the special phase of the glottis (circles in Fig. 3.8) such as the maximum and the minimum of the glottal pulse derivative waveform was marked for every segment. The waveform was multiplied by rectangular window with length of one fundamental period. Selected segments were fixed in one of those two phases and overlaid.

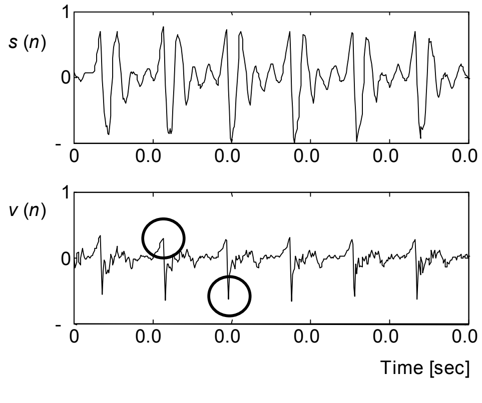


Fig. 3.8 Example of a speech signal (upper graph) and corresponding glottal pulses derivative (lower graph).

Based on the graphical interpretation, a two dimensional distribution matrix was generated as it can be seen in Fig. 3.9. The amplitude-time space is divided into small elements via horizontal and vertical lines (180 intervals on the time axis, 100 intervals on the amplitude axis). The number of glottal pulse waveform lines going through a cell is equal to the numerical value of the corresponding element of the distribution matrix 100x180.

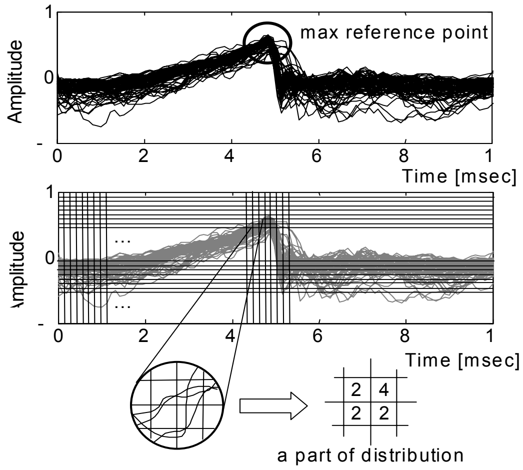
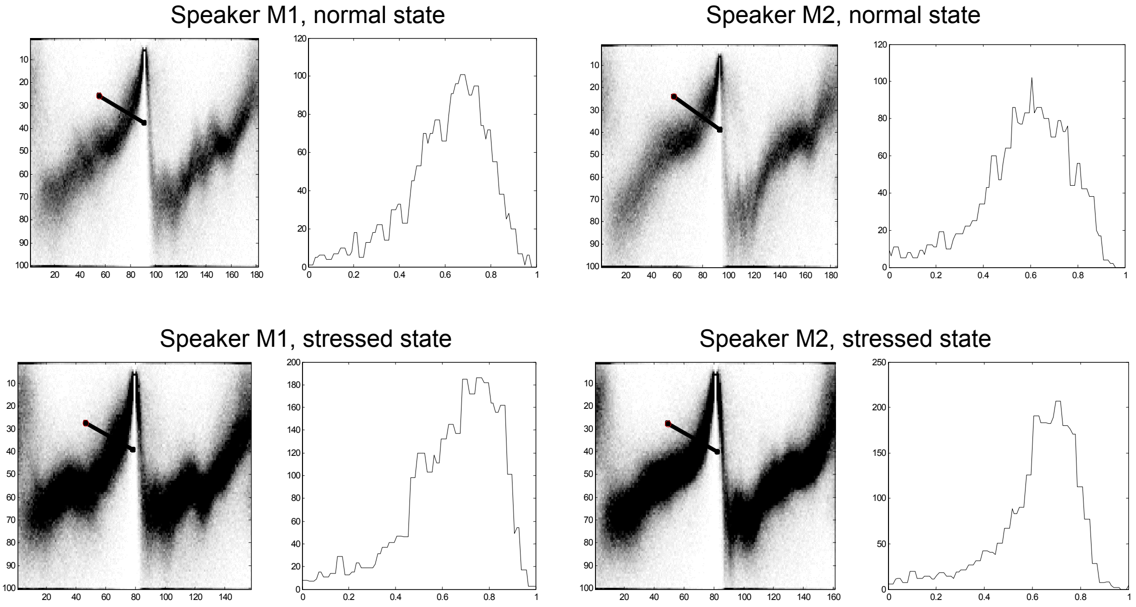


Fig. 3.9 Generating the distribution matrix of glottal pulses derivative.

The obtained distribution matrix was displayed as a gray scale image, where the maximum and minimum values of the matrix are black and white. In order to compare distribution matrices automatically with each other, it is inevitable to find a useful description (few significant features) of the matrices.


Speaker M1, normal state

Speaker M2, normal state

Speaker M1, stressed state

Speaker M2, stressed state

Fig. 3.10 Graphical samples of distribution matrices and their comparative cuts for two speakers.

An effective criterion seems to be straight cuts made at a reference position [20]. Figure 3.10 shows the positions of applied cuts and the form of the intersection for two speakers in both, normal and stressed state. In this case, the fixation point for all segments is in each period the maximum of the glottal pulses derivative waveform (upper circle in Fig. 3.8). For stressed state, the distribution matrix seems to be "blacker" than for normal state; it means if the speaker is under stressed conditions, derivative waveforms of produced glottal pulses are more concentrated to the average waveform and the distribution form in cuts is more asymmetric. These effects are obvious for almost any speaker. An experiment with a mathematical description of given cuts resulted in use of parameters based on the statistical moments of 3rd and 4th order.

4 CONCLUSION

Presented work deals with voice analysis and recognition. Voice recognition is the complement of speech recognition. In speech recognition, the processing tries to extract linguistic information from the speech signal to the exclusion of personal information. Conversely, voice recognition requires the processing to focus on the characteristics unique to the individual, disregarding the actual word spoken.

The first part of the work introduces the speech signal carrying the message information, provides a brief historical overview of speech processing, and specifies the application area of voice recognition. In the second part of the work, techniques for speaker identification and speaker verification are provided including information about the evaluation of efficient voice parameters.

The final part of the work is focused on feature extraction methods that are useful in emotion recognition. In particular, speech spoken by speaker under psychological stress was investigated. The most interesting common features are the pitch, the formants, the long-time spectrum, and the mel-cepstral coefficients. New and original features that are based on voice production model have been defined and optimised using statistical analysis of glottal pulses. The IAIF algorithm was applied to estimate the glottal pulse waveforms. The stress detection rate in the speaker dependent recognition achieved 88%. The presented stress is proven objectively by the increased heart rate of speakers. To analyse the speakers psychologically we need suitable databases with specific realistic speech signals. A new database of speech under stress was created for use in our experiments consisting of data collected during oral final examinations at our university. For the Czech language, this is the first appropriate database with realistic stressed speech. In our future research we will try to develop algorithms for on-line detection and quantification of stress.

Although many advances and successes in speaker recognition have recently been achieved, there are still many problems to which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. Because of differences among national languages and various "speaking behaviour" relating to each nation, it is necessary to perform research for each language separately. Some of the mentioned areas will be solved in the close future in cooperation with M.Sc. and Ph.D. students, which is a guarantee of interconnection between the research and teaching activities at the BUT. The increase in application opportunities has resulted in increased interest in voice recognition research. Speaker recognition is nowadays regarded by market projections as one of the more promising technologies of the future.

References

[1] LANCKER, D.; KREIMAN, J.; EMMOREY, K. Familiar voice recognition: Patterns and parameters. *Journal of Phonetics*. 1985, vol. 13, no. 1, pp. 19-38.

[2] REICH, A.; DUKE, J. Effects of selected vocal disguises upon speaker identification by listening. *Journal of the Acoust. Soc. Am*. 1979, vol. 66, no. 4, pp. 1023-1028.

[3] TITZE, I. R. Physiologic and acoustic differences between male and female voices. *Journal of the Acoust. Soc. Am*. 1989, vol. 85, no. 4, pp.1699-1707.

[4] JUANG, B. H.; RABINER, L. R. *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice Hall, 1993.

[5] SIGMUND, M. Efektivní určování spektra řečových signálů. In *Proceedings Nové smery v spracovaní siglálov VII*. Tatranské Zruby: VA L. Mikuláš, 2004, pp. 153-155.

[6] SIGMUND, M. *Voice Recognition by Computer*. Marburg: Tectum Verlag, 2003.

[7] JOHNSTONE, T.; SCHERER, K. The effects of emotions on voice quality. In *Proceedings of 14th International Congress of Phonetic Science*. San Francisco, 1999, pp. 2029-2032.

[8] HANSEN, J. H.; GHAZALE, S. E. Getting started with SUSAS. In *Proceedings of Eurospeech'97*. Rhodes: ESCA, 1997, pp. 1743-1746.

[9] BURKHARDT, F.; PAESCHKE, A.; ROLFES, M.; SENDLMEIER, W.; WEISS, B. A database of German emotional speech. In *Proceedings Interspeech 2005*. Lissabon: ISCA, 2005, pp. 1517-1520.

[10] SIGMUND, M. Introducing the database ExamStress for speech under stress. In *Proceedings of 7th IEEE Nordic Signal Processing Symposium (NORSIG 2006)*. Reykjavik: University of Iceland, 2006, pp. 290-293.

[11] SIGMUND, M.; DOSTÁL, T. Analysis of emotional stress in speech. In *Proceedings of Artificial Intelligence and Applications*. Innsbruck: IASTED, 2004, pp. 317-322.

[12] PSUTKA, J.; MÜLLER, Z.; MATOUŠEK, J.; RADOVÁ, V. Mluvíme s počítačem česky. Praha: Academia, 2006.

[13] SIGMUND, M. Untersuchungen zum Zusammenhang zwischen Sprachsignal und Sprecher unter Stress. In *Proceedings of KONVENS 2004*. Vienna: Medical University of Vienna, 2004, pp. 189-192.

[14] SIGMUND, M. Spectral Characteristics of vocal tract for speaker recognition. In *International Journal of Computer Science and Network Security*. 2006, vol.6, no.1A, pp. 17-19.

[15] SIGMUND, M.; SEVERŇÁK, O. Eine neue Sprachdatenbank mit der Sprache unter Stress. In *Proceedings of ESSV*. Bonn: University of Bonn, 2001, pp. 323-328.

[16] BOŠTÍK, M.; SIGMUND, M. Methods for estimation of glottal pulses waveforms exciting voiced speech. In *Proceedings of Eurospeech 2003*. Geneva: ISCA, 2003, pp. 2389-2392.

[17] ISELI, M. R.; ALWAN, A. Inter- and intra-speaker variability of glottal flow derivative using the LF model. In *Proceedings of ICSLP*. Beijing: ISCA, 2000, vol. 1, pp. 477-480.

[18] BOŠTÍK, M.; SIGMUND, M. Speaker stress detection by analysis of glottal excitation. In *Proceedings of MAVEBA*. Firenze: University of Firenze, 2003, pp. 87-90.

[19] ALKU, P. An automatic method to estimate the time-based parameters of the glottal pulse form. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*. San Francisco: IEEE, 1992, pp. 29-32.

[20] SIGMUND, M.; DOSÁL, T. Detection of psychological stress by analysing of glottal pulse waveform. In *Proceedings of Artificial Intelligence and Applications*. Innsbruck: IASTED, 2007, pp. 526-529.

Souhrn

Řeč slouží primárně k přímé akustické komunikaci mezi lidmi. Prostřednictvím řečového signálu ji můžeme věrně zaznamenat, archivovat a přenášet, ale také analyzovat technickými prostředky. Veškeré informace obsažené v čistém řečovém signálu (signál bez poruch a pozadí) představují ze 75% informace o obsahu zprávy a z 25% informace o nositeli zprávy - o mluvčím. Přitom cca. 15% z celkové informace reprezentuje identitu mluvčího a cca. 10% se váže ke krátkodobé charakteristice mluvčího odrážející také jeho aktuální psychický stav. Zatímco při rozpoznávání řeči se snažíme z řečového signálu získat čistě lingvistické informace nezávisle na mluvčím, rozpoznávání hlasu je zaměřeno na personální informace bez ohledu na aktuálně vyslovená slova.

Předložená práce se pokouší poskytnout ucelený náhled do problematiky signálové analýzy hlasu. V první kapitole jsou nejdříve uvedeny nejčastější způsoby reprezentace řečových signálů. Pak následuje stručný historický přehled rozvoje celkové disciplíny zpracování řečových signálů až do současného stavu. Dále je zmíněno rozpoznávání hlasu u lidí přirozenou cestou. V závěru úvodní kapitoly jsou představeny oblasti praktického uplatnění automatické analýzy hlasu. Druhá kapitola je zaměřena na dosud nejvíce propracované podoblasti analýzy hlasu zabývající se identifikací a verifikací mluvčích. K tomu účelu je také uveden postup pro výběr a hodnocení vhodných parametrů řečového signálu a úvaha o ideálním rozpoznávání mluvčích.

Třetí kapitola pojednává o specielní problematice zkoumání vlivu stresu na tvorbu řeči a jeho projevu v řečovém signálu. Nejdříve je uveden princip působení emocí na vytvářenou řeč. Důležitým předpokladem vývoje a testování nových metod je dostupnost vhodných řečových dat. V práci jsou uvedeny některé zahraniční databáze emotivní a stresové řeči a představena první česká databáze reálné stresové řeči ExamStress, která vznikla na VUT v Brně. V dalších částech jsou prezentovány vlastní výsledky získané při analýze stresové řeči. Pro detekci stresu v závislosti na mluvčích a na textu se ukázalo výhodné zkoumat změny mel-cepstrálních koeficientů a prvních dvou formantových kmitočtů. Pro detekci stresu nezávisle na mluvčích a na textu byly použity statistické hodnoty rozložení základního tónu řeči a dlouhodobé spektrum získané pomocí koeficientů lineární predikce. Slibné výsledky se ukazují zejména při analýze hlasivkových pulzů. Zde byl nejdříve použit Liljencrant-Fantův model pro aproximaci získaných pulzů a vyhodnoceny změny jeho parametrů v důsledku stresu. Dále byly vytvořeny distribuční matice z průběhů první derivace hlasivkových pulzů, na definovaných pozicích provedeny jejich řezy a statisticky vyhodnoceny. Úspěšnost detekce stresu se pohybuje kolem 88%. Objektivní potvrzení přítomnosti stresu bylo prokázáno výrazně zvýšeným srdečním pulzem, který byl snímán souběžně s řečovým signálem.

I když byly dosaženy měřitelné výsledky slibující možnosti praktického využití v komerční, bezpečnostní i forenzní sféře, zůstává celá řada nedořešených problémů, které vyplývají ze složitosti procesu tvorby a vnímání řeči. V dalším výzkumu se zaměříme na on-line detekci stresu a pokusíme se detekovaný stres kvantifikovat. Vzhledem k rozdílům mezi národními jazyky nejsou všechny metody a výsledky analýzy hlasu zcela přenositelné a pro aplikační účely je žádoucí provádět výzkum v každém jazyce. Naše pracoviště je dosud jediným pracovištěm v České republice, které se zabývá zkoumáním stresu pomocí signálové analýzy hlasu.