18

Information Mining from Speech Signal

Milan Sigmund
Brno University of Technology
Czech Republic

1. Introduction

Language is the engine of civilization and speech is its most powerful and natural form 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. The representation of language as speech signals in digital form is, of course, of fundamental concern for all sub-fields of machinery speech processing. Speech data are characterized by a large variability. The production of connected speech is affected not only by the well-known coarticulation phenomena, but also by a large number of sources of variation such as regional, social, stylistic and individual ones. People speak differently according to their geographical provenance (accent or dialect) and according to factors such as the linguistic background of their parents, their social status and their educational background. Individual speech can vary because of different timing and amplitude of the movements of speech articulators. Moreover, the physical mechanism of speech undergoes changes, which can affect the nasal cavity resonance and the mode of vibration of the vocal cords. This is obvious, for instance, as a consequence of any laryngeal pathology, as when the speaker has a cold. Less obvious are changes in the fundamental frequency and phonation type, which are brought by factors such as fatigue and stress or in the long term by aging. A series of environmental variables like background noise, reverberation and recording conditions have also to be taken into account. In essence, every speech production is unique and this uniqueness makes the automatic speech processing quite difficult.

Information mining from speech signal as the ultimate goal of data mining is concerned with the science, technology, and engineering of discovering patterns and extracting potentially useful or interesting information automatically or semi-automatically from speech data. In general, data mining was introduced in the 1990s and has deep roots in the fields of statistics, artificial intelligence, and machine learning. With the advent of inexpensive storage space and faster processing over the past decade, data mining research has started to penetrate new grounds in areas of speech and audio processing.

This chapter deals with issues related to processing of some atypical speech and/or mining of specific speech information, issues that are commonly ignored by the mainstream speech processing research. Atypical speech can be broadly defined as speech with emotional content, speech affected by alcohol and drugs, speech from speakers with disabilities, and various kinds of pathological speech.

2. Speech Signal Characteristics

2.1 Information in speech

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 phonemes (sounds) per second. The phonemes are language-specific units and thus each language needs a declaration of its own phonetic alphabet. The numbers of phonemes commonly in use in each literary language vary between 30 and 50. 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 bits/sec for the average information rate of speech. In other words, the written equivalent of speech contains information equivalent to 60 bits/sec at normal speaking rate. This is in a contrast to the minimal bit rate of 64 kb/sec measured in digital speech signal at lowest acceptable speech quality obtained with 8 bits/sample at sampling rate 8 kHz. The high information redundancy of a speech signal is associated with such factors as the loudness of the speech, environmental condition, and emotional, physical as well as psychological state of the speaker. Many of these characteristics are also subjectively audible, but much of the phonetically irrelevant information is few distinguishable by untrained humans. However, some specific information hidden in speech signal can be detected using advanced signal processing methods only.

Word duration from the information point of view was studied in different European languages. Figure 1 shows the average word length in number of syllables and corresponding information (Boner, 1992).

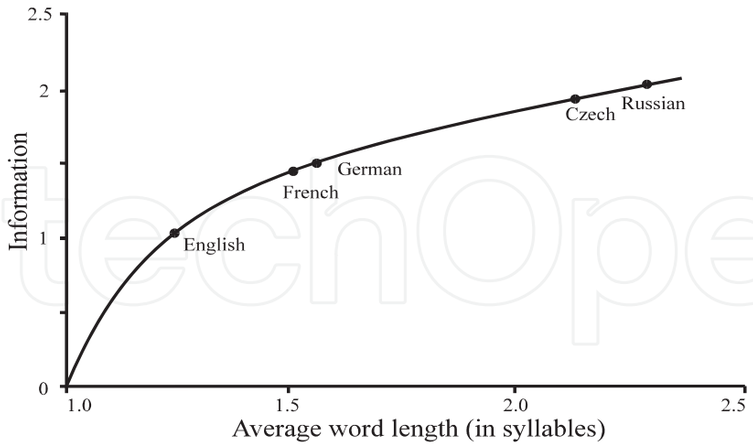


Fig. 1. Average word duration vs. information for some languages.

2.2 Phonemic notation of individual languages

With the growth of global interaction, the demands for communications across the boundaries of languages are increasing. In case of systems for speech recognition, before the

machine can understand the meaning of an utterance, it must identify which language is being used. Theoretically, the differences between different spoken languages are manifold and large. Although these differences can be found at various levels (e.g. phoneme inventory, acoustic realization of phonemes, lexicon, etc.) how to reliably extract these features for is still an unsolved problem. A brief review of approaches for language identification can be found, for instance, in (Yan et al., 1996) and (Matějka, 2009). Navratil applied a particularly successful approach based on phonotactic-acoustic features and presented new system for language recognition as well as for unknown language rejection (Navratil, 2001).

Speech processing research focused on mining of specific information from speech signal aims to develop analyzers that are task-, speaker- and vocabulary-independent so as to be easily adapted to a variety of applications for different languages. When porting an analyzer to a new language, certain system parameters or components will have to be changed, i.e. those incorporating language-dependent knowledge sources such as the selection of the phoneme set, the recognition lexicon (alternate word pronunciations), and phonological rules. Many language dependent factors are related to the acoustic confusability of the words in the language (such as homophone, monophone and compound word rates) and the word coverage of a given size recognition vocabulary. There are other parameters which can be considered language independent, such as the language model weight and word or phoneme insertion penalties. The selection of these parameters can vary however depending on factors such as the expected out-of-vocabulary rate. In this section we discuss the important characteristics for the most widespread European languages (i.e. English, German, and French).

Comparing French and English we may observe that for lexicons, the number of words must be doubled for French in order to obtain the same word coverage as for English. The difference in lexical coverage for French and English mainly stems from the number and gender agreement in French for nouns, adjectives and past participles, and the high number of different verbal forms for a given verb (about 40 forms in French as opposed to at most 5 in English). German is also a highly inflected language, and one can observe the same phenomena as in French. In addition, German has case declension for articles, adjectives and nouns. The four cases: nominative, dative, genitive and accusative can generate different forms for each case which often are acoustically close. For example, while in English there is only one form for the definitive article *the*, in German number and gender are distinguished, giving the singular forms *der, die, das* (male, female, neuter) and the plural form *die*. Declension case distinction adds 3 additional forms *des, dem, den* to the nominative form *der*. In German most word can be substantivized, thus generating lexical variability and homophones in recognition. The major reason of the poor lexical coverage in German certainly arises from word compounding. Whereas compound words or concepts in English are typically formed by a sequence of words (e.g. *the speech recognition problem*) or in French by adding a preposition (e.g. *le probléme de la reconnaissance de la parole*), in German words are put together to form a new single word (e.g. *Spracherkennungsproblem*) which in turn include all number, gender and declension agreement variations.

Looking at language-dependent features in lexica and texts, we can observe that the number of homophones is higher for French and German that for English. In German homophones arise from case sensitivity and from compound words being recognized as sequences of component words. A major difficulty in French comes from the high number of monophone

words. Most phonemes can correspond to one or more graphemic forms (e.g. the phoneme *ɛ* can stand for *ai, aie, aies, ait, aient, hais, hait, haie, haies, es, est*). The other languages have fewer monophones, and these monophones are considerably less frequent in the texts. Counting monophone words in newspaper texts, gave about 17% for French versus 3% for English (Lamel et al., 1995). In French, not only is there the frequent homophone problem where one phonemic form corresponds to different orthographic forms, there can also be a relatively large number of possible pronunciations for a given word. The alternate pronunciations arise mainly from optional word-final phonemes, due to liaison, mute *e* and optional word-final consonant cluster reduction. One particular feature of French is liaison. Liaison is where normally silent word final consonants are pronounced when immediately followed by a word initial vowel. This improves the fluency of articulation of natural French speech. Languages with a larger lexical variability require larger training text sets in order to achieve the same modeling accuracy.
For acoustic modeling we use the phoneme in context as basic unit. A word in the lexicon is then acoustically modeled by concatenating the phoneme models according to the phonemic transcription in the lexicon. The phonemes are language-specific units and thus each language needs a declaration of its own phonetic alphabet. The numbers of phonemes commonly in use in each literary language mentioned above are listed in Tab. 1.

Language	Phonemes
English	45
French	35
German	48

Table 1. Number of phonemes in some European languages.

The phoneme set definition for each language, as well as its consistent use for transcription is directly related to the acoustic modeling accuracy. The set of internationally recognized phonemic symbols is known as the International Phonetic Alphabet (IPA). This alphabet was first published in 1888 by the Association Phonétique Internationale. A comprehensive guide to the IPA is the handbook (IPA, 1999). In many EU countries, the SAMPA (Phonetic Alphabet, created within the Speech Assessment Methods) has been widely used recently. None of the above mentioned alphabets is directly applicable to Czech and other Slavic languages. It is because some sounds that are specific for Czech (not only the well-known *ř* but also some others, e.g. *ď, ť, ň*) are not included there. That is why it was necessary to define a Czech phonetic alphabet. The alphabet, denoted as PAC (Phonetic Alphabet for Czech) consists of 48 basic symbols that allows for distinguishing all major events occurring in spoken Czech language (Nouza et al., 1997). Typically, there are some tongue-twisting consonant clusters in Czech which are difficult to pronounce by non-Czechs, e.g. words such as *zmrznout* (English *to freeze*), *čtvrtek* (English *thursday*), *prst* (English *finger*), etc.

2.3 Basic model of speech production

Based on our knowledge of speech production, the appropriate model for speech corresponding to the electrical analogs of the vocal tract is shown in Figure 2. Such analog models are further developed into digital circuits suitable for simulation by computer.

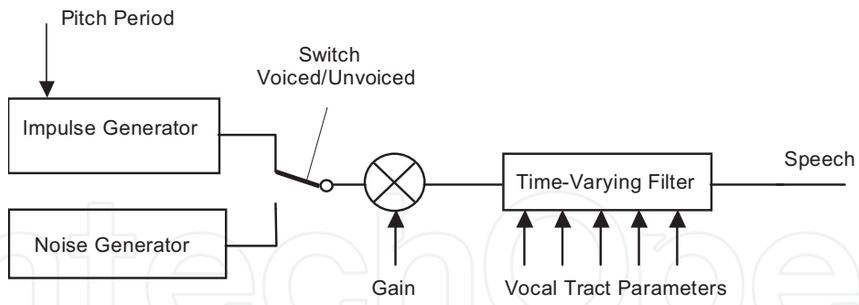


Fig. 2. Electrical model of speech production.

In modeling speech, the effects of the excitation source and the vocal tract are often considered independently. The actual excitation function for speech is essentially either a quasi-periodic pulse train (for voiced speech sounds) or a random noise source (for unvoiced speech sounds). In both cases, a speech signal $s(t)$ can be modeled as the convolution of an excitation signal $e(t)$ and an impulse response characterizing the vocal tract $v(t)$

$$s(t) = e(t) * v(t) \tag{1}$$

which also implies that the effect of lips radiation can be included in the source function (Quatieri, 2002). convolution of two signals corresponds to multiplication of their spectra, the output speech spectrum $S(f)$ is the product of the excitation spectrum $E(f)$ and the frequency response $V(f)$ the vocal tract.

$$S(f) = E(f)\, V(f) \tag{2}$$

The excitation source is chosen by a switch whose position is controlled by the voiced/unvoiced character of the speech. The appropriate gain G of the source is estimated from the speech signal and the scaled source is used as input to a filter, which is controlled by the vocal tract parameters characteristic of the speech being produced. The parameters of this model all vary with time.

Unvoiced excitation is usually modeled as random noise with an approximately Gaussian amplitude distribution and a flat spectrum over most frequencies of interest. More research has been done on voiced excitation because the naturalness of synthetic speech is crucially related to accurate modeling of voiced speech. It is very difficult to obtain precise measurements of glottal pressure or glottal airflow. The glottal airflow can be measured directly via electro-glottography, pneumotachography or photoglottography (Baken & Orlikoff, 2000). The mostly used electroglottography is a non-invasive method of measuring vocal fold contact during voicing without affecting speech production. The Electroglottograph (EGG) measures the variation in impedance to a very small electrical current between the electrodes pair placed across the neck as the area of vocal fold contact changes during voicing. Simultaneously with the glottal flow can be recorded also the speech pressure signal. The speech pressure signal includes information about glottal pulses waveform. Because of electroglottographs are quite expensive devices only the speech pressure signal is often recoded. The glottal airflow is then estimated from this signal. A typical glottal airflow $\Phi(t)$ of voiced speech in steady state is periodic and roughly resembles a half-rectified sine wave (see Fig. 3). From a value of zero when the glottis is closed, the

airflow gradually increases as the vocal folds separate. The closing phase is more rapid than the opening phase due to the Bernoulli force, which adducts the vocal folds (O´Shaughnessy, 1987).

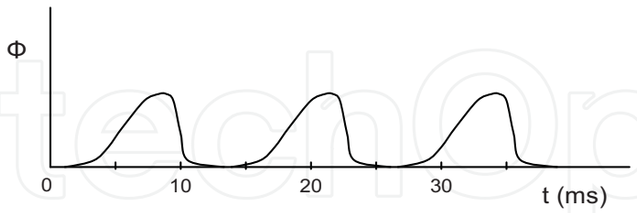


Fig. 3. Simplified glottal waveform during a voiced sound.

Figure 4 shows photography of the vocal folds during a voicing cycle when completely open and completely closed (Chytil, 2008). The vocal folds are typically 15 mm long in men and 13 mm in women. In general, the glottal source estimation has a great potential for use in identifying emotional states of speaker, non-invasive diagnosis of voice disorders, etc.

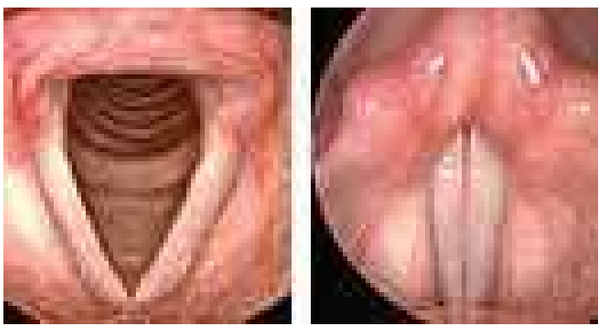


Fig. 4. Vocal folds in the open phase (left) and closed phase (right).

3. General Principles of Speech Signal Processing

The whole processing block chain common to all approaches to speech processing shows Fig. 5. The first step in the processing is the speech pre-processing, which provides signal operations such as digitalization, preemphasis, frame blocking, and windowing. Digitalization of an analog speech signal starts the whole processing. The microphone and the A/D converter usually introduce undesired side effects. Because of the limited frequency response of analog telecommunications channels and the widespread use of 8 kHz sampled speech in digital telephony, the most popular sample frequency for the speech signal in telecommunications is 8 kHz. In non-telecommunications applications, sample frequencies of 12 and 16 kHz are used. The second step, i.e. features extraction, represents the process of converting sequences of pre-processed speech samples $s(n)$ to observation vectors $\mathbf{x}$ representing characteristics of the time-varying speech signal. The properties of the feature measurement methods are discussed in great details in (Quatieri, 2002). The kind of features extracted from speech signal and put together into feature vector $\mathbf{x}$ corresponds to the final aim of the speech processing. For each application (e.g., speaker identification, gender selection, emotion recognition, etc.), the most efficient features, i.e. the features

carrying best the mining information, should be used. The first two blocks represent straightforward problems in digital signal processing. The subsequent classification is then optimized to the final expected information. In contrary to the blocks of features extraction and classification, the block of pre-processing provides operations that are independent on the aim of speech processing.

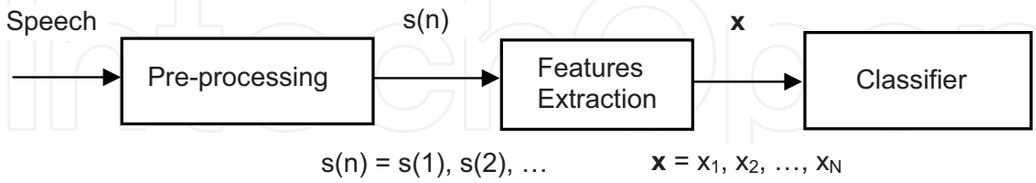


Fig. 5. Block diagram of the speech processing.

3.1 Preemphasis

The characteristics of the vocal tract define the current uttered phoneme. Such characteristics are evidenced in the frequency spectrum by the location of the formants, i.e. local peaks given by resonances of the vocal tract. Although possessing relevant information, high frequency formants have smaller amplitude with respect to low frequency formants. To spectrally flatten the speech signal, a filtering is required. Usually, a one coefficient FIR filter, known as a preemphasis filter, with transfer function in the *z*-domain

$$H(z) \quad = \quad 1 - \lambda \, z^{-1} \tag{3}$$

is used. In the time domain, the preemphasized signal is related to the input signal by the difference equation

$$\tilde{s}(n) \quad = \quad s(n) - \lambda \; s(n-1) \tag{4}$$

A typical range of values for the preemphasis coefficient is $\lambda \in [0.9\text{-}1.0]$. One possibility is to choose an adaptive preemphasis, in which λ changes with time according to the relation between the first two values of autocorrelation coefficients

$$\lambda \; = R(1) / R(0) \tag{5}$$

The effect of preemphasis on magnitude spectrum of short phoneme can be seen in Fig. 6.

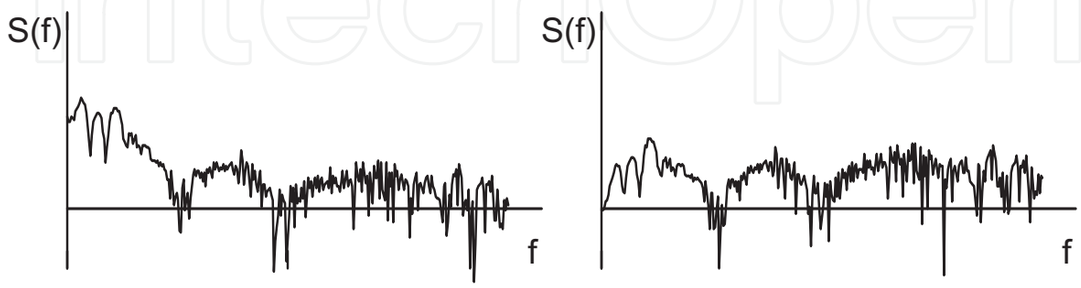


Fig. 6. Phoneme spectrum without preemphasis (left) and after preemphasis (right).

3.2 Frame blocking

The most common approaches in speech signal processing are based on short-time analysis. The preemphasized signal is blocked into frames of N samples. Frame duration typically ranges between 10-30 msec. Values in this range represent a trade-off between the rate of change of spectrum and system complexity. The proper frame duration is ultimately dependent on the velocity of the articulators in the speech production system. Figure 7 illustrates the blocking of a word into frames. The amount of overlap to some extent controls how quickly parameters can change from frame to frame.

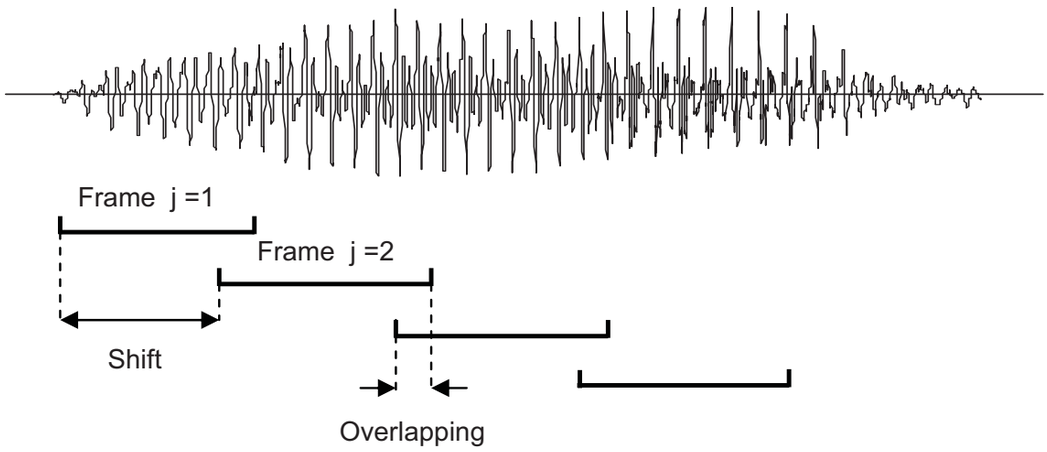


Fig. 7. Blocking of speech into overlapping frames.

3.3 Windowing

A signal observed for a finite interval of time may have distorted spectral information in the Fourier transform due to the ringing of the $\sin(f)/f$ spectral peaks of the rectangular window. To avoid or minimize this distortion, a signal is multiplied by a window-weighting function before parameter extraction is performed. Window choice is crucial for separation of spectral components which are near one another in frequency or where one component is much smaller than another. Window theory was once a very active topic of research in digital signal processing. The basic types of window function can be found in (Oppenheim et al., 1999). Today, in speech processing, the Hamming window is almost exclusively used. The Hamming window is a specific case of the Hanning window. A generalized Hanning window is defined as

$$w(n) = \frac{\alpha - (1-\alpha)\cos(2\pi n / N)}{\beta} \qquad \text{for} \quad n = 1,...,N \tag{6}$$

and $w(n) = 0$ elsewhere. α is defined as a window constant in the range <0,1> and N is the window duration in samples. To implement a Hamming window, the window constant is set to $\alpha = 0.54$. β is defined as a normalization constant so that the root mean square value of the window is unity.

$$\beta = \sqrt{\frac{1}{N}\sum_{n=1}^{N} w^2(n)} \tag{7}$$

In practice, it is desirable to normalize the window so that the power in the signal after windowing is approximately equal to the power of the signal before windowing. Equation (7) describes such a normalization constant. This type of normalization is especially convenient for implementations using fixed-point arithmetic hardware.

Windowing involves multiplying a speech signal $s(n)$ by a finite-duration window $w(n)$, which yields a set of speech samples weighted by the shape of the window. Regarding the length N, widely used windows have duration of 10-25 msec. The window length is chosen as a compromise solution between the required time and frequency resolution. A comparison between the rectangular window and the Hamming window, their time waveforms and weighted speech frame, is shown in Fig. 8.

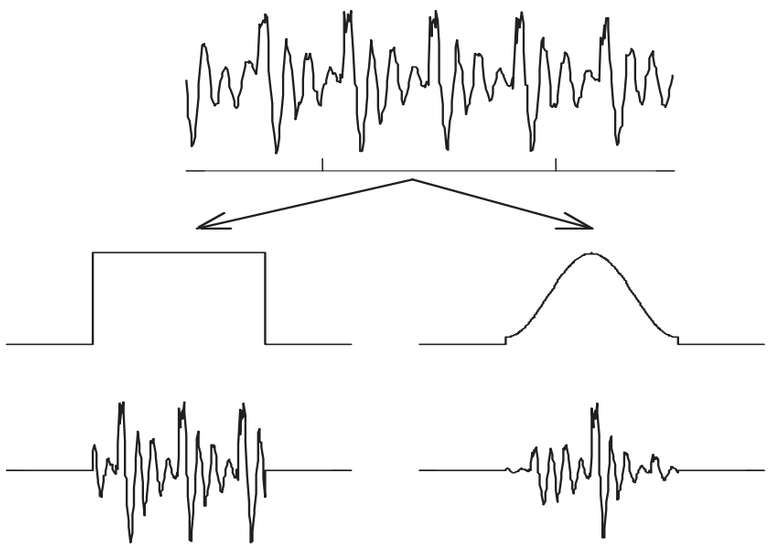


Fig. 8. Window weighting functions and the corresponding frames cut out from a speech signal by the rectangular window (left) and by the Hamming windows (right).

4. Effect of Stress on Speech Signal

The most emotional states of a speaker can be identified from the facial expression, speech, perhaps brainwaves, and other biological features of the speaker. In this section, the problem of speech signal under psychological stress is addressed. Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity and deterioration of performance. Psychological stress has a broad sense and a narrow sense effect. The broad sense reflects the underlying long-term stress and the narrow sense refers to the short-term excitation of the mind that prompts people to act. In automatic recognition of stress, a machine would not distinguish whether the emotional state is due to long-term or short-term effect so well as it is reflected in facial expression. Stress is more or less present in all professions in today's hectic and fast-moving society. The negative influence of stress

on health, professional performance as well as interpersonal communication is well known. A comprehensive reference source on stressors, effects of activating the stress response mechanisms, and the disorders that may arise as a consequence of acute or chronic stress is provided, for example, in the Encyclopedia of Stress (Fink, 2007).
Stress may be induced by external factors (noise, vibration, etc.) and by internal factors (emotion, fatigue, etc.). Physiological consequences of stress are, among other things, changes in the heart rate, respiration, muscular tension, etc. The muscular 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.9. The accepted term for the speech signal carrying information on the speaker's physiological stress is "stressed speech".

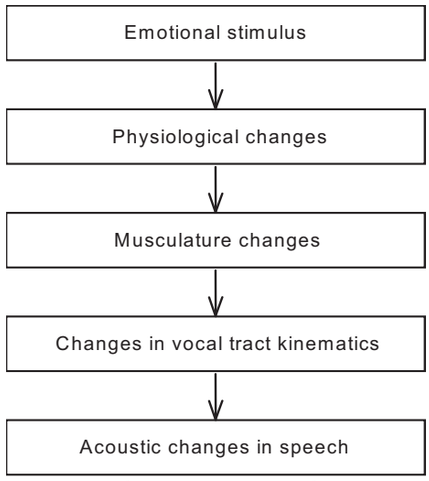


Fig. 9. Model of how emotion causes changes in speech.

Assessment of speaker stress has applications such as sorting of emergency telephone message, telephone banking, and hospitals. Stress is recognized as a factor in illness and is probably implicated in almost every type of human problem. It is estimated that over 50% of all physician visits involve complaints of stress-related illness.

4.1 Stressed speech databases

The evolution of algorithms for recognition of stressed speech is strictly related to the availability of large amount of speech whose characteristics cover all the variability of specific information required for the application target. However, it is really difficult to obtain realistic voice samples of speakers in various stressed states, recorded in real situations. "Normal people" (as well as professional actors) cannot simulate real case stress perfectly with their voices.
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 SUSC-0 (Haddad et al., 2002). The advantage of this database is that an objective measure of workload was obtained, and that physiological stress measures (heart rate, blood pressure, respiration, and transcutaneous pCO_2) were recorded simultaneously with the speech signal. However, such

extreme situations as crashed aircraft occur seldom in everyday life. The most frequently mentioned corpus in the literature is the SUSAS (Speech Under Simulated and Actual Stress) database of stressed American English described in (Hansen & Ghazale, 1997) 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 into many aspects of emotions including stress, and it also collected emotion databases. Their website provides access to a number of databases and research materials. The German database of emotional utterances including panic was recorded at the Technical University of Berlin. A complete description of the database called Berlin Database of Emotional Speech can be found in (Burkhardt et al., 2005). A list of existing emotional speech data collections including all available information about the databases such as the kinds of emotions, the language, etc. was provided in (Ververidis & Kotropoulos, 2006).
For our studies conducted within research into speech signals we created and used our own database. The most suitable event 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 persons were 31 male pre-graduate and post-graduate students, mostly Czech native speakers. The created database called ExamStress consists of two kinds of speech material: stressed speech collected during the state exams and neutral speech recorded a few days later, both spoken by the same speakers. The students were asked to give information about some factors, which can correlate with stress in influencing the voice, e.g. the number of hours of sleep during the previous night, the use of (legal) drugs or alcohol shortly before examination, etc. This information was added to the records in the database. The recording platform is set up to store the speech signals live in 16-bit coded samples at a sampling rate of 22 kHz. Thus, the acoustic quality of the records is determined by the speaking style of the students and the background noise in the room. A complete description of the ExamStress database can be found in (Sigmund, 2006). In some cases the heart rate *HR* of students was measured simultaneously with the speech recordings in both stressed and neutral 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. On average, the *HR* values obtained for stressed state were almost doubled compared to the neutral state (such values usually occur if a person is under medium physical activity).

4.2 Changes in time and frequency domain

From various emotion analyses reported in the literature, it is known that emotion causes changes in three groups of speech parameters: a) voice quality; b) pitch contour; c) time characteristics. To get the quantitative changes of speech parameters, we applied in first study some simple features that had not been specifically designed for the detection of stressed speech, such as vowel duration, formants and fundamental frequency (Sigmund & Dostal, 2004).
Duration analysis conducted across individual vowel phonemes shows the main difference in the distribution of vowel "a". By contrast, the small differences in the distribution of vowels "e" and "i" seem to be irrelevant for the detection of emotional stress (Fig. 10).

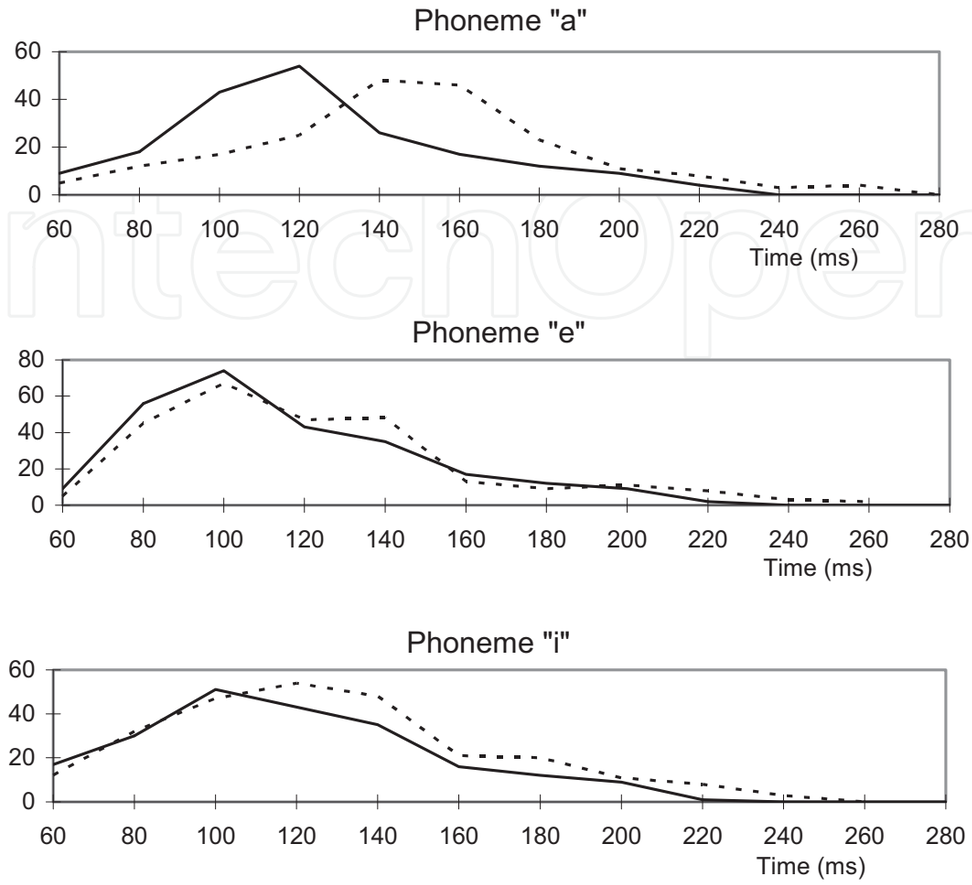


Fig. 10. Distribution of duration for the vowels ”a“, “e“ and “i“ (solid lines are for normal speech, dotted lines for speech under stress).

In general, more significant results are given by formants. Formant values were obtained via a formant-tracking algorithm based on peak-picking. 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 2 shows the average formant values for phoneme “i“.

	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

Table 2. Formant changes in spectrum for phoneme “i“ (all in Hz).

Further, the characteristics of pitch were estimated. The fundamental frequency F_0 contours were calculated on the frame-by-frame basis using the center-clipping autocorrelation method (Rabiner, 1993). 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. Table 4.11 shows the results obtained for three male speakers. Figure 11 illustrates the F_0 distribution obtained for speaker "Kl" in Tab. 3.

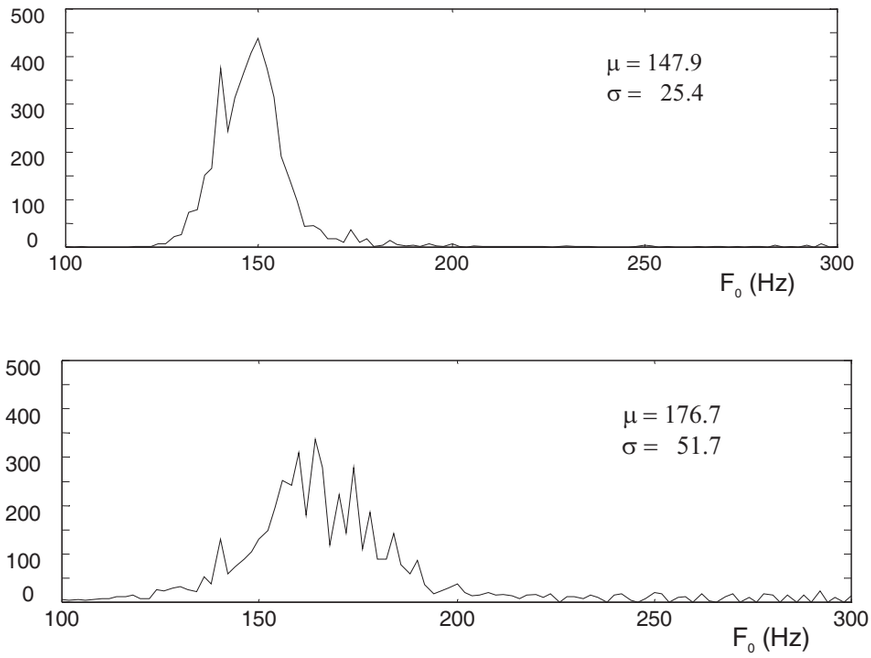


Fig. 11. Pitch distribution for speaker "Kl" (upper graph is for normal speech, lower graph is for speech under stress).

Normal speech has a single narrow high peak at around 148 Hz, alcoholic speech a somewhat broader peak while stressed speech is broader still. The area under each curve is related to the number of frames of speech observed, which is directly related to the speaking rate. The curves are comparable because they were obtained from speaking/reading the same text.

	Speaker "De"		Speaker "Fl"		Speaker "Kl"	
	Mean	Dev.	Mean	Dev.	Mean	Dev.
Normal Speech	127	16	142	13	148	25
Stressed Speech	162	25	243	61	177	52

Table 3**.** Mean values and standard deviation of F_0 distributions (all in Hz).

The current most commonly used short-term spectral measurements are cepstral coefficients and their frequency-warped alternative coefficients. To compute the cepstrum, we first compute the log spectral magnitudes and next the inverse Fourier transform (IFT) of the log spectrum. The output signal is a set of cepstral coefficients

$$cc(\tau) = \text{IFT}\{\log \mid \text{FT}[s(n)] \mid\} \tag{8}$$

called as the cepstrum of signal $s(n)$. The low-order terms of the cepstrum correspond to short-term correlation in the speech signal (vocal tract information). The local maxima in the higher order terms demonstrate long-term correlation or periodicity in the waveform (excitation information). Experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified. When one of the components of this sound falls outside this bandwidth, it can be individually distinguished (Zwicker, 1999). The subjective nonlinear perception of frequency had led to an objective computational model that converts a physically measured spectrum into a psychological "subjective spectrum". Used mapping of acoustic frequency f to the so-called mel scale for subjective pitch is

$$pitch = 2595 \log\left(1 + \frac{f}{700}\right). \tag{9}$$

The mel scale attempts to map the perceived frequency of a tone onto a linear scale. This scale is often approximated as a linear scale from 0 to 1000 Hz and then a logarithmic scale beyond 1000 Hz. The algorithm for estimation of the mel-warped cepstral coefficients can be found, for instance, in (Rabiner & Juang, 1993).

In our experiments we focused the cepstral analysis on the data set within the vowel class. The first 12 mel-cepstral coefficients *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 corresponded vowels were compared. The most effective indicator of stress seems to be the 9th mel-cepstral *mcc*(9) coefficient computed from the vowel "u". Table 4 shows the mean values of *mcc*(9) obtained for three various speakers. This indicator gives a higher value in case of stressed speech.

	Speaker "De"					
Test #	1		2		3	
Speech	N	S	N	S	N	S
mcc(9)	-0.191	-0.096	-0.206	-0.150	-0.121	0.056
	Speaker "Fl"					
Test #	1		2		3	
Speech	N	S	N	S	N	S
mcc(9)	-0.152	0.160	-0.101	-0.025	-0.024	0.002
	Speaker "Kl"					
Test #	1		2		3	
Speech	N	S	N	S	N	S
mcc(9)	-0.157	-0.094	-0.177	-0.107	0.110	0.187

Table 4. Mean values of the 9th mel-cepstral coefficient for the vowel "u" (N denotes normal speech and S is for stressed speech).

Figure 12 illustrates ten values of the coefficient *mcc*(9) obtained for speaker ″De“ in the Test # 3 mentioned in Tab. 4. The mean values in Tab. 4 were calculated from 10 speech frames (each of 40 msec).

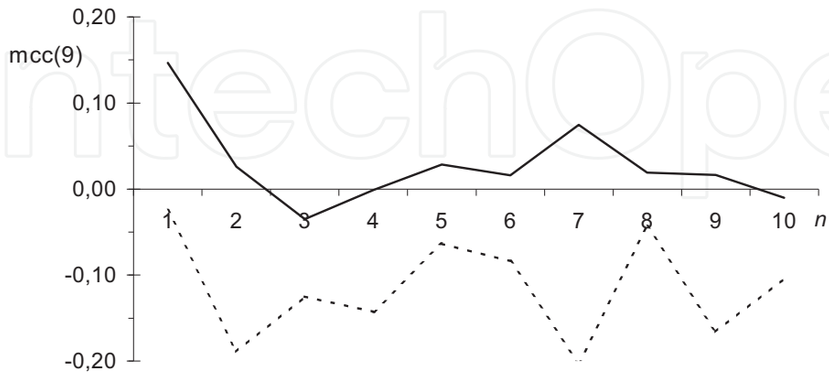


Fig. 12. Values of coefficient *mcc*(9) in ten corresponding speech frames for normal speech (dotted line) and stressed speech (solid line).

4.3 Changes in glottal pulse excitation

In our experiments, glottal pulses were obtained from 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 (Alku, 1992). Other techniques for obtaining glottal pulses from speech signal can be found, for example, in (Bostik & Sigmund, 2003). The block diagram of the IAIF is shown in Fig. 13. This method operates in two repetitions, hence the word iterative in the name of the method. The first phase (blocks LPC 1st order, filter $H_1^{-1}(z)$, LPC 12th order, filter $H_2^{-1}(z)$) generates an estimate of glottal excitation, which is subsequently used as input of the second phase (blocks LPC 4th order, filter $H_3^{-1}(z)$, LPC 12th order, filter $H_4^{-1}(z)$) to achieve a more accurate estimate. The steps of the method are described in detail below. Firstly, the input speech signal is analyzed by first-order LPC predictor. This step gives an initial estimate of the effect of glottal flow on the speech spectrum. Using the obtained filter $H_1^{-1}(z)$ of 1st order, the input signal is inversely filtered. This step effectively removes the spectral tilt caused by the spectrum of the excitation signal. The output of the previous step is analyzed by the LPC predictor of 12th order to obtain a model of the vocal tract transfer function. The order of the LPC analysis is related to the number of formants to be modeled. The input signal is then inversely filtered by filter $H_2^{-1}(z)$ using the inverse of the 12th order model from the previous step. This yields the first estimate of the glottal pulse derivative and completes the first repetition. The second repetition runs analogously.

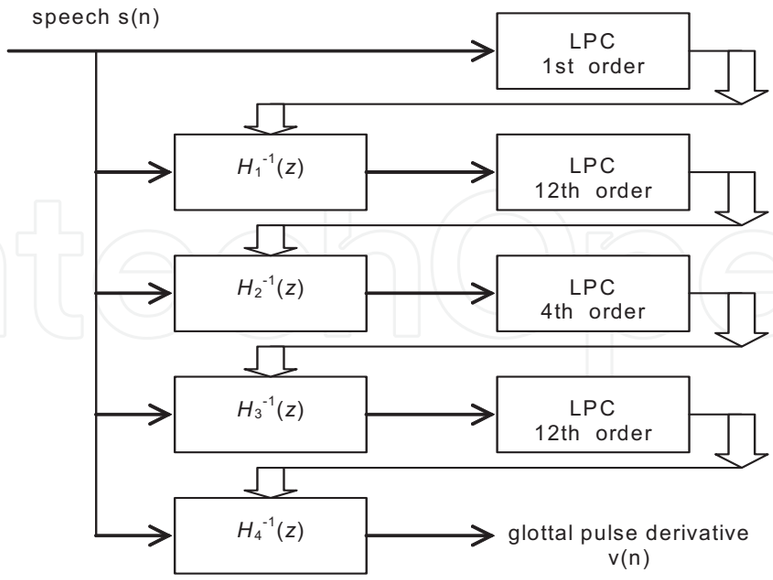


Fig. 13. Block diagram of the IAIF algorithm.

Figure 14 shows a typical waveform $s(n)$ of the vowel "a" and its 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. 4) 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 of one fundamental period in length. Selected segments were fixed in one of the two phases and overlaid.

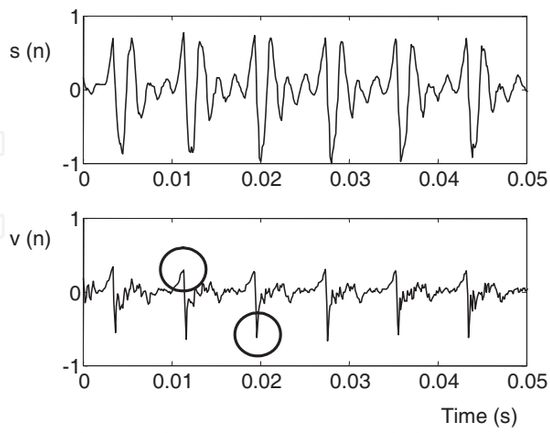


Fig. 14. Example of a speech signal (upper graph) and the corresponding glottal pulse derivative (lower graph).

Based on the graphical interpretation, a two-dimensional distribution matrix was generated. 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 distribution matrix obtained was displayed as a gray scale image where the maximum and minimum values of the matrix are black and white. An example of such an image created from about 4000 segments can be seen in Fig. 15. In this case, the fixation point for all segments was in each period the maximum of the glottal pulse derivative waveform (upper circle in Fig. 14).

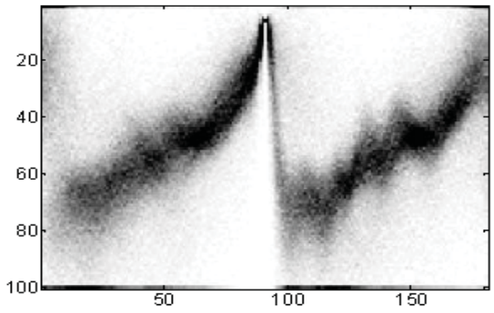


Fig. 15. Illustration of a distribution matrix of glottal pulses derivative waveform interpreted in gray scale.

The ultimate goal in our experiments was to find common speech characteristics of stressed speech based on distribution matrices of glottal pulses. In order to compare the distribution matrices automatically with each other, it is inevitable to find a useful description (a few significant features) of the matrices. An effective criterion seems to be straight cuts made at a reference position. Figure 16 shows the positions of applied cuts and the form of the intersection for two speakers in both neutral and stressed state. For the stressed state, the distribution matrix seems to be "blacker" than for the neutral state; it means that if the speaker is under stress, the derivative waveforms of the glottal pulses produced are more concentrated about the average waveform and the distribution form in the cuts is more asymmetric; in the cut the mass of the distribution is concentrated on the right of the figure. These effects are obvious in almost any speaker. Another type of cuts provides less useful information (Bostik, 2005).

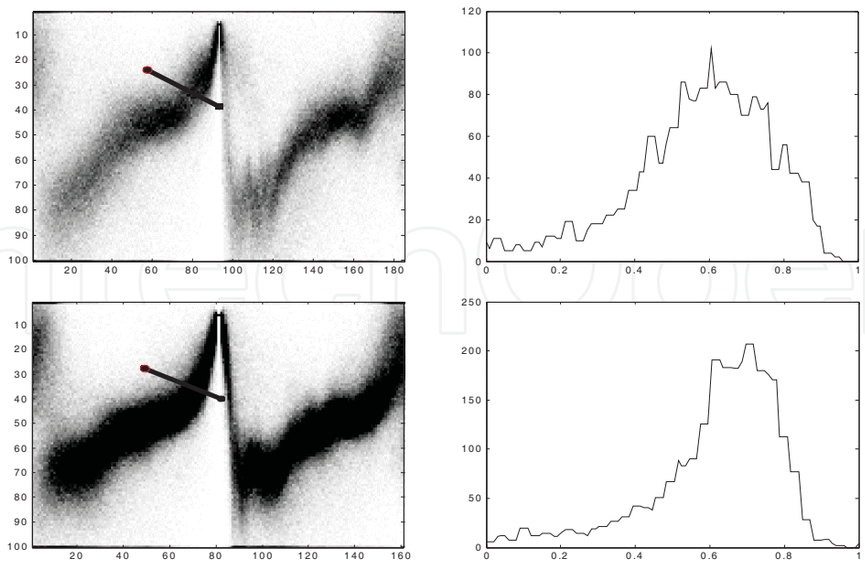


Fig. 16. Graphical samples of distribution matrices and their comparative cuts estimated for the vowel "a".

An experiment with mathematical description of applied cuts resulted in the use of two effective parameters: α and k. The first parameter, α, is defined by

$$\alpha = \frac{S_1}{S_2 + S_3}, \tag{10}$$

where S_1, S_2, and S_3 are the sub-areas of the cut located symmetrically to the maximum of the cut and bounded graphically by lines in 20%, 40%, 60%, and 80% of the total width of the cut, as illustrated in Fig. 17.

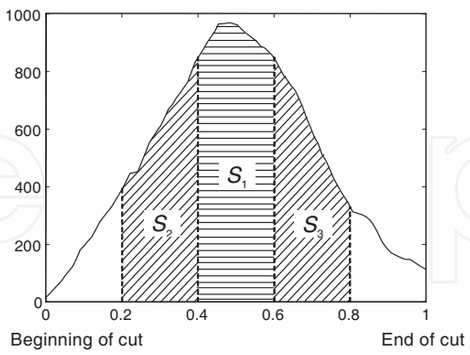


Fig. 17. Definitions of the sub-areas S_1, S_2, and S_3 in a distribution matrix cut.

The second parameter, k, is defined as

$$k = \frac{\mu_4}{\sigma^4} - 3, \tag{11}$$

where μ_4 is the fourth central moment and σ is the standard deviation. For our experiment, 31 male speakers from the ExamStress database were used. Approximately 2000 voiced segments of 5 vowels were extracted from the speech data of each speaker for each state. The IAIF algorithm was applied to those segments to estimate the glottal pulse waveforms. Two distribution matrices for both neutral and stressed state were calculated for each speaker and the speaker's state was estimated using the distribution parameters of cuts in a binomial classification (stress/no stress). The classification was performed by using the Mahalanobis distance measure. Table 5 (Sigmund et al., 2008) shows samples of parameters α and k computed from the speech signals of vowel phonemes for a group of ten male speakers in neutral state (denoted N) and stressed state (denoted S).

Speaker	Parameter a		Parameter k	
	N	S	N	S
M1	1.09	1.20	2.91	3.21
M2	0.85	1.92	2.66	3.50
M3	1.32	1.17	3.04	3.31
M4	0.93	1.18	2.88	2.96
M5	0.78	0.88	2.12	2.86
M6	0.76	1.40	2.45	3.30
M7	0.55	0.83	2.30	2.63
M8	1.03	1.40	2.68	3.19
M9	0.91	1.05	2.82	2.90
M10	0.87	1.32	2.71	3.25

Table 5. Values of applied parameters for neutral (N) and stressed (S) speech.

In the stressed state, slightly higher values of both applied parameters are indicated in most cases. The stress recognition rate in the speaker dependent recognition achieved 88%. In the speaker independent experiments without neutral reference speech data the recognition rate decreased to 72%.

5. Effect of Alcohol on Speech Signal

The term alcohol refers generically to compounds with a hydroxyl group [-OH]. In our work, alcohol refers only to ethanol also called ethyl alcohol. This is the specific compound found in alcoholic beverages. Research of alcohol detection from speech signal was started worldwide by accident of the tanker Exxon Valdez in March 1989. A suspicion arose the captain was influenced by alcohol during the accident, but it was impossible to prove it, because blood alcohol tests were executed too late. A tape with recordings of a dialogue between the captain and terrestrial radio communication station was the only material, which could clarify the situation. Therefore an intensive research of alcohol influence to speech signal followed and the suspicion was confirmed 2 years later (Brenner & Cash,

1991). Subsequently, insurance offices and security organizations began to support next research in this field.

There are two main ways of reporting alcohol concentration in the body, blood-alcohol concentration (BAC) and breath-alcohol concentration (BrAC). Of these two, BAC enjoys some primacy, and in fact, BrAC is very often converted to an expression of equivalent BAC. Alcoholic intoxication causes changes in emotional state and changes in psychomotorics in short-term point of view. It means that the recognition of alcohol influence from speech will be superimposed by the recognition of emotional state. It was proved that emotional information in a speech signal is mainly carried by excitation rather than by the vocal tract in linear modeling of speech. So if we want to separate the influence of alcohol from that of emotion, we must concentrate on vocal tract information. Vocal tract parameters and their changes can represent the quality of psychomotorics in fact. Psychomotorical changes are noticeable on levels of over 0.5 ‰ of blood-alcohol concentration (BAC). Exceeding the level of 1.5 ‰ BAC, changes in psychomotorics are so distinct that speech defects are audible by the human ear.

There are not many available corpora designed to allow the study of speech signal carrying information on the speaker's alcohol intoxication. The German database of alcoholic speech called Alcohol Language Corpus was recorded at the University of Munich (Schiel et al., 2008).). In our research, we used alcoholic voices from a small own database collected at the Brno University of Technology. This database was created by recording 25 speakers (13 males and 12 females) aged 18 to 50 years, who twice said a set of 5 utterances for each given phrase: one set at a level of 0.0 ‰ BAC (sober) and one set at a level of 0.5 ‰ to 1.0 ‰ BAC. The texts of recorded utterances were chosen by an empirical criterion, they are mostly words containing liquids ("r" and "l"), which are relatively difficult to pronounce. The values of BAC were measured by the Drivesafe breathalyzer. Thus, the alcoholic database contains records of sober speakers, records of speakers influenced by alcohol and the approximated BAC values of speakers.

We made several sets of measurements to detect alcohol intoxication in speech signal. Alcohol-induced changes in speech were observed in both the short-time and the long-time domains. First, an analysis of fundamental frequency F_0 was performed. Table 6 gives the mean F_0 values for speakers of both genders in sober state and after alcohol consumption. From the sober condition to a measurable alcohol level, the mean F_0 increased for 21 speakers, decreased for 3 speakers and for one male speaker it remained unchanged. The magnitude of change was greater for the increases in F_0 than for the decreases. The maximum increase in F_0 was 18 Hz for a female speaker.

	Male Speakers		Female Speakers	
	0.5 - 0.8 ‰	0.8 - 1.0 ‰	0.5 - 0.8 ‰	0.8 - 1.0 ‰
F_0 alcoholized	123.5	121.1	209.5	214.0
F_0 sober	121.4	116.2	205.8	207.4
Difference	2.1	4.9	3.7	6.6

Table 6. Changes in fundamental frequency F_0 for 25 speakers (all in Hz).

Because of the fact that an increase in the mean F_0 value can also be caused by other stimuli, fundamental frequency alone is not sufficient as an alcohol indicator. Further, a comparative observation of the significant speech features was performed to find among them the best candidate for alcohol identification. All records from the alcoholic database were parameterized in terms of linear predictive coefficients (LPC), cepstral coefficients (CC), PARCOR coefficients (PC), log area ratio coefficients (LAR), and delta parameters of each of the above features denoted by the prefix "Δ". More details to these speech features can be found e.g. in (Quatieri, 2002) and (Rabiner & Juang, 1993). The classification task was simplified using only two categories of speaker's state. The first state stands for sober speaker (0.0 ‰ BAC) and the second state for intoxicated speaker with an alcohol level of over 0.5 ‰ BAC. The utterance means were computed using dynamic time warping (Rabiner & Juang, 1993) averaging for each speaker, each state and each type of feature. We observed the dispersion inside each state and dispersion between both states. The ratio of inter-state to intra-state dispersion could represent a global ability of a feature to distinguish the two states.

	LPC	CC	PC	LAR	ΔLPC	ΔCC	ΔPC	ΔLAR
D_{inter}	107.40	20.34	17.61	58.75	49.00	6.74	6.96	22.78
D_{intra}	33.71	5.13	5.33	15.65	22.98	3.07	3.15	10.14
D_{inter}/D_{intra}	3.19	3.97	3.30	2.24	2.13	2.19	2.21	2.24

Table 7. Ratios of inter-state dispersion D_{inter} to intra-state dispersion D_{intra} for various speech features.

Table 7 shows word-dependent results for the key word "Laura", which seems to be very suitable word for this purpose. Although the best absolute ratio is given by cepstral coefficients, the log area ratio coefficients provide an approximately equal score in both direct and delta forms. The results for delta parameters are independent of timing in the pronunciation and thus more important. An objective function as a difference between the inter-state dispersion and intra-state dispersion was proposed in (Menšik, 1999). Using this criterion, the most effective phoneme to detection alcohol in speech seems to be consonant "r", especially on the boundaries between the trilled "r" and vowels. These results are particularly interesting from the point of view of acoustic theory, which sometimes cites "r" as an example of the many-to-one relationship between articulatory configuration and acoustic results.

At present, the two main spheres in which alcohol testing from voice becomes meaningful are vehicular traffic and workplaces. The increasing availability of digital speech processing techniques shifts the trend toward instrumental analysis in alcohol and speech research.

6. Conclusion

Human voice is the key tools that human use to communicate. In addition to the intended messages, a significant part of information contained in speech signal refers to the speaker. These phonologically and linguistically irrelevant speaker-specific information make speech recognition less effective but can be used for speaker recognition and analysis of the speaker's emotional and health state. Such a speech cue would allow an analysis without the physical presence of the speaker. While examining stress or alcohol we are only concerned with the physically measurable characteristics of the speech signal. Besides these changes in the spoken language the content of the language, e.g. repetition of selected words, structure of the sentence, etc. is also very important for speech analysis made by psychologists, psychiatrists and other experts.

Information mining from speech signal includes many ways of applying machine learning, speech processing, and language processing algorithms to benefit and serve commercial applications. It also raises and addresses several new and interesting fundamental research challenges in the areas of prediction, search, explanation, learning, and language understanding. Effective techniques for mining speech, audio, and dialog data can impact numerous business and government applications. The technology for monitoring conversational speech to discover patterns and generate alarms is essential for intelligence and law enforcement organizations as well as for enhancing call center operation.

7. References

Alku, P. (1992). An automatic method to estimate the time-based parameters of the glottal pulseform, *Proceedings of ICASSP'92*, pp. 29–32, ISBN 0-7803-0532-9, San Francisco, March 1992, IEEE Press, Piscataway, USA

Baken, R. J. & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice,* Singular Publishing Group, ISBN 1-56593-869-0, San Diego, CA, USA

Boner, A. (1992). *Spracherkennung mit Computer*, AT Verlag, ISBN 3-85502-435-9, Aarau, Switzerland

Bostik, M. & Sigmund, M. (2003). Methods for estimation of glottal pulses waveforms exciting voiced speech, *Proceedings of Eurospeech'03*, pp. 2389-2392, ISSN 1018-4074, Geneva, September 2003, International Speech Communication Association, Grenoble

Bostik, M. (2005). *Voice Analysis for Stress Recognition*, Ph.D. Thesis, Brno University of Technology

Brenner, M. & Cash, J. R. (1991). Speech analysis as an index of alcohol intoxication - the Exxon Valdez accident. *Aviation, Space, and Enviromental Medicine,* Vol. 62, No. 9, (September 1991), pp. 893-898, ISSN 0095-6562

Burkhardt, F.; Paeschke, A.; Rolfes, M; Sendlmeier, W. & Weiss, B. (2005). A database of German emotional speech, *Proceedings of Eurospeech'05*, pp. 1517-1520, ISSN 1018-4074, Lisbon, September 2005, International Speech Communication Association, Grenoble

Fink, G. (2007). *Encyclopedia of Stress*, Academic Press, ISBN 978-0-12-088503-9, London, New York

Haddad, D.; Walter, S.; Ratley, R. & Smith, M. (2002). *Investigation and Evaluation of Voice Stress Analysis Technology*. Project Report: Rome Laboratory, NY, USA

Hansen, J. H. & Ghazale, S. E. (1997). Getting started with SUSAS, *Proceedings of Eurospeech'97*, pp. 1743-1746, ISSN 1018-4074, Rhodes, Greece, September 1997, International Speech Communication Association, Grenoble

Chytil, P. (2008). *Voice Analysis for Detection of Diseases*, Ph.D. Thesis, Brno University of Technology

IPA-International Phonetic Association (1999). *A Guide to the Use of the International Phonetic Alphabet*, University Press, ISBN 0-521-63751-1, Camridge

Lamel, L.; Adda-Decker, M. & Gauvain, J. L. (1995). Issues in large vocabulary, multilingual speech recognition, *Proceedings of Eurospeech'95*, pp. 185-188, ISSN 1018-4074, Madrid, September 1995, International Speech Communication Association, Grenoble

Matějka, P. (2009). *Advancing Phonotactic and Acoustic Language Recognition*, Ph.D. Thesis, Brno University of Technology

Menšik, R. (1999). Recognition of Alcohol Influence on Speech, *Proceedings of Workshop on Text, Speech and Dialogue*, pp. 384-387, ISBN 978-3-540-66494-9, Plzeň, September 1999, Springer-Verlag, Heidelberg

Navratil, J. (2001). Spoken language recognition – a step toward multilinguality in speech processing. *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 6, (September 2001), pp. 678-685, ISSN 1063-6676

Nouza, J.; Psutka, J. & Uhlíř, J. (1997). Phonetic alphabet for speech recognition of Czech. *Radioengineering*, Vol. 6, No. 4, (April 1997), pp. 16-20, ISSN 1210-2512

Oppenheim, A. V.; Schafer, R. W. & Buck, J. R. (1999). *Discrete-Time Signal Processing*, Prentice Hall, ISBN 0-13-7549202, Englewood Cliffs, NJ, USA

O´Shaughnessy, D. (1987). *Speech Communication - Human and Machine*, Addison-Wesley Publishing, ISBN 0-201-16520-1, New York, USA

Quatieri, T. F. (2002). *Discrete-Time Speech Signal Processing*, Prentice Hall, ISBN 0-13-242942-X, Englewood Cliffs, NJ, USA

Rabiner, L. R. & Juang, B. H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, ISBN 0-13-015157-2, Englewood Cliffs, NJ, USA

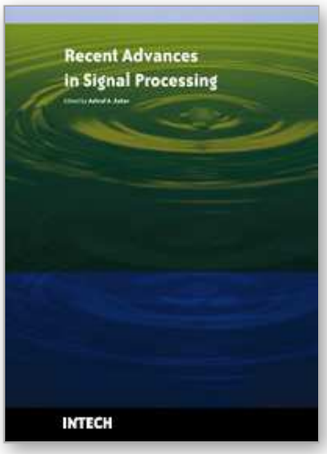
Schiel, F.; Heinrich, Ch.; Barfüßer, S. & Gilg, T. (2008). ACL - Alcohol Language Corpus, *Proceedings of LREC 2008*, pp. 1-5, ISBN 2-9517408-4-0, Marrakesh, Marocco, May 2008, European Language Resources Association, Paris

Sigmund, M. & Dostal, T. (2004). Analysis of emotional stress in speech, *Proceedings of IASTED Internat. Conf. on Artificial Intelligence and Applications,* pp. 317-322, ISBN 0-88986-404-7, Innsbruck, Austria, February 2004, ACTA Press, Calgary

Sigmund, M. (2006). Introducing the database ExamStress for speech under stress, *Proceedings of 7th IEEE Nordic Signal Processing Symposium*, pp. 290-293, ISBN 1-42244-0413-4, Reykjavik, Iceland, June 2006, IEEE Signal Processing Society, Piscataway

Sigmund, M.; Prokes A. & Brabec, Z. (2008). Statistical analysis of glottal pulses in speech under psychological stress, *Proceedings of EUSIPCO,* pp. 1-5, Lausanne, Switzerland, August 2008, EURASIP, Leuven

IntechOpen



Recent Advances in Signal Processing

Edited by Ashraf A Zaher

ISBN 978-953-307-002-5

Hard cover, 544 pages

Publisher InTech

Published online 01, November, 2009

Published in print edition November, 2009

The signal processing task is a very critical issue in the majority of new technological inventions and challenges in a variety of applications in both science and engineering fields. Classical signal processing techniques have largely worked with mathematical models that are linear, local, stationary, and Gaussian. They have always favored closed-form tractability over real-world accuracy. These constraints were imposed by the lack of powerful computing tools. During the last few decades, signal processing theories, developments, and applications have matured rapidly and now include tools from many areas of mathematics, computer science, physics, and engineering. This book is targeted primarily toward both students and researchers who want to be exposed to a wide variety of signal processing techniques and algorithms. It includes 27 chapters that can be categorized into five different areas depending on the application at hand. These five categories are ordered to address image processing, speech processing, communication systems, time-series analysis, and educational packages respectively. The book has the advantage of providing a collection of applications that are completely independent and self-contained; thus, the interested reader can choose any chapter and skip to another without losing continuity.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen