Brno University of Technology
Faculty of Electrical Engineering and Computer Science
Institute of Radio Electronics

Ing. Milan Sigmund, CSc.

SPEAKER RECOGNITION

Identifying People by their Voices

Habilitation Thesis

Brno 2000

ISBN: 80-214-1590-8

Contents

About Author **4**

Introduction **5**

1 Principles of Speaker Recognition **6**
1.1 Speaker Recognition by Humans 6
1.2 Areas of Automatic Speaker Recognition 6
1.3 Ideal Voice Recognition 8

2 Feature Parameters **10**
2.1 Parameters for Speaker Recognition 10
2.2 Evaluation of Parameters 12

3 Long-Time Spectrum of Vocal Tract **13**
3.1 Estimation of Vocal Tract Long-Time Spectrum 13
3.2 Speech Normalization by Long-Time Spectrum 15
3.3 Conclusions 15

4 Effects of Emotional Stress on Speech **16**
4.1 Stressed Speech Data 16
4.2 Detection of Stressed Speech 16
4.3 Conclusions 18

5 Conclusion **19**

Alphabetical List of Principal References **20**

Czech Abstract (Souhrn) **21**

Milan Sigmund was born in Ivančice, Czech Republic, in 1959. He graduated with the M.Sc. degree (Ing.) in biomedical engineering and the Ph.D. degree (CSc.) in speech signal processing, both from the then Technical University of Brno. Currently, he is an assistant professor at the Institute of Radio Electronics of the Faculty of Electrical Engineering and Computer Science, Brno University of Technology.

His main research interests lie in the areas of speech signal processing with a special focus on automatic speaker recognition. His current goal is to develop a robust speaker processing system optimised to operate under all sorts of adverse acoustic conditions, including speaker and environmental variability. In his teaching activities he gave various courses on digital signal processing, pulse and non-linear techniques.

Introduction

Language is the engine of civilisation, and speech is its most powerful and natural form. Textual language has become extremely important in modern life, but speech has dimensions of richness that text cannot approximate. For example, the health, sex and attitude of a person are all naturally and subliminally communicated by that person's speech. Such linguistic information has social value and serves important communicative functions in our everyday lives.

The purpose of speech is communication. There are several ways of characterising the communication potential of speech. According to information theory, speech can be represented in terms of its message content. An alternative way of characterising speech is in terms of the signal carrying the message information, i.e. the acoustic waveform. 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.

The increase in application opportunities has resulted in increased interest in speaker recognition research. Over the past three decades, a wide variety of speech processing techniques have been proposed and speech recognition has been in the centre of attention in the whole world. While researching into voice recognition, we found relatively little literature, and much of what we did find consisted of highly technical fragments of research published in journals and conference proceedings, to which most people do not have access. The purpose of this work is to provide an interpretative overview and perspective of voice recognition tasks and evaluation methodology. This topic is discussed in the habilitation thesis but it has been omitted here for lack of space. It is in this context that we present our own results in some special areas of voice recognition, e.g. disclosure of professional imitator, stressed speech analysis, and effect of alcohol on speech.

The design of any automatic speech processing system requires a large amount of spoken data to obtain reliable acoustic models and/or adequate language models for specific tasks. Thus during last years the design of adequate speech databases has been an important point of interest in the speech recognition community. Nowadays it is possible to find large and well-defined phonetic corpora and speech databases for specific tasks for widely used languages like English, French or Japanese [2]. For the English languages most speech corpora are distributed by the „Linguistic Data Consortium" (LDC). For non English languages, the „International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment" (COCOSDA) was established in 1990 to encourage and promote international interaction and cooperation in the foundation areas of spoken language processing. Unfortunately, the currently available databases do not meet the need of all research areas, so many non-standard test databases are still used.

The methods developed are useful not only from the technical point of view but also as a teaching aid because the students can process and model their own speech signal using these methods. Voice can be visualised in time and frequency domain, extracted features can be related with speech in its original acoustic form which all help the students understand the complex phenomena in speech signal analysis and recognition.

1 Principles of Speaker Recognition

1.1 Speaker 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 [7] 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 highpass or lowpass 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 [7]. If the utterances were increased to 4 sec, 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 [9]. 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.

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 exceeds 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.

1.2 Areas of Automatic Speaker Recognition

Speaker recognition is the general term used to include all of the many different applications of discriminating people based on the sound of their voices. There are many terms to distinguish the main different areas of application as follows:

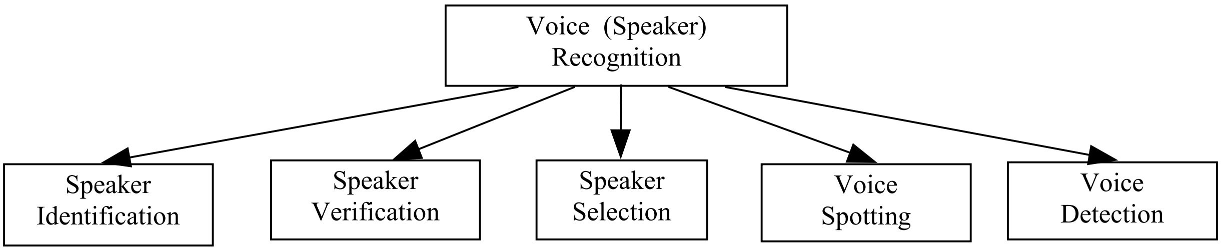

Fig. 1.1 Areas of voice (speaker) recognition.

Speaker identification aims to identify a speaker who belongs to a group of users through a sample of his speech. 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 1.2 shows the basic structure of speaker identification system.

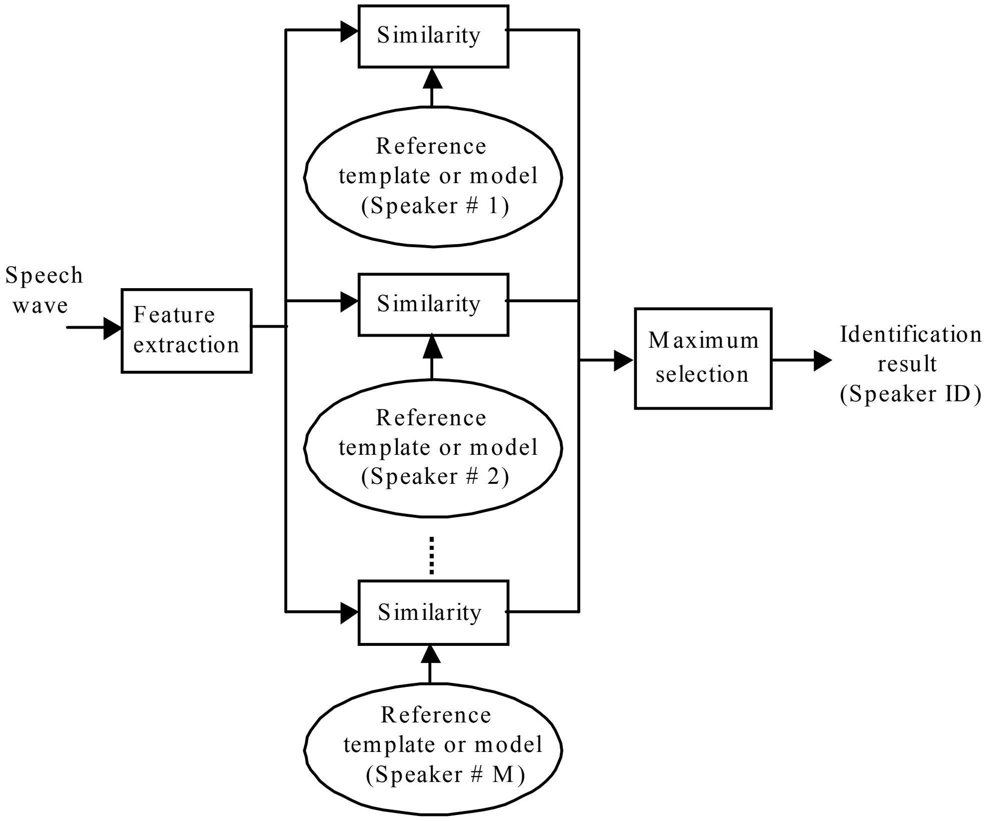
Speech wave
Feature extraction
Similarity
Reference template or model (Speaker # 1)
Similarity
Reference template or model (Speaker # 2)
Similarity
Reference template or model (Speaker # M)
Maximum selection
Identification result (Speaker ID)

Fig. 1.2 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 1.3 shows the basic structure of speaker verification system.

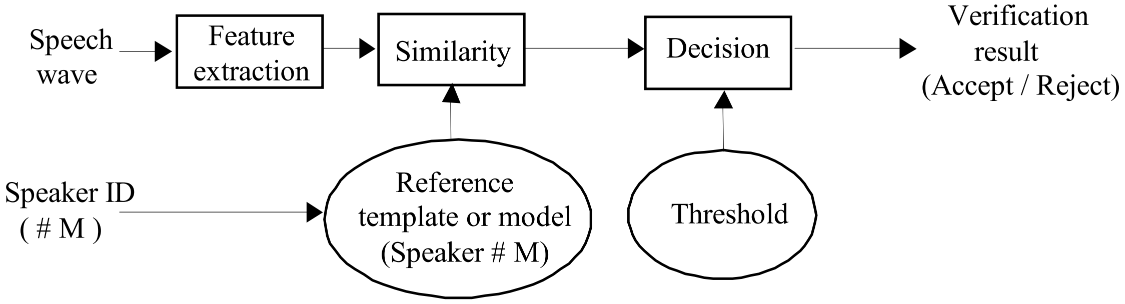
Speech wave
Feature extraction
Similarity
Decision
Verification result (Accept / Reject)
Speaker ID (# M)
Reference template or model (Speaker # M)
Threshold

Fig. 1.3 Basic structure of speaker verification system.

The sub-area speaker selection includes some specific applications as selection of sex, age, education, geographical provenience and other demographic factors. The age estimation of unknown speakers' 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 using speech samples (mood state identification) and to detect any pathologies using speech samples (health state identification).

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

1.3 Ideal Voice 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 systems 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. 1.4 or by using physical measuring devices, e.g. X-ray.

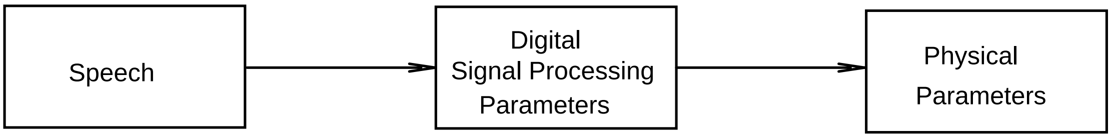

Fig. 1.4 Ideal parameter extraction.

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.

As an example, Fig. 1.5 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. 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 [18]

$$F_0 = 1700 / L_m \quad , \tag{1.1}$$

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

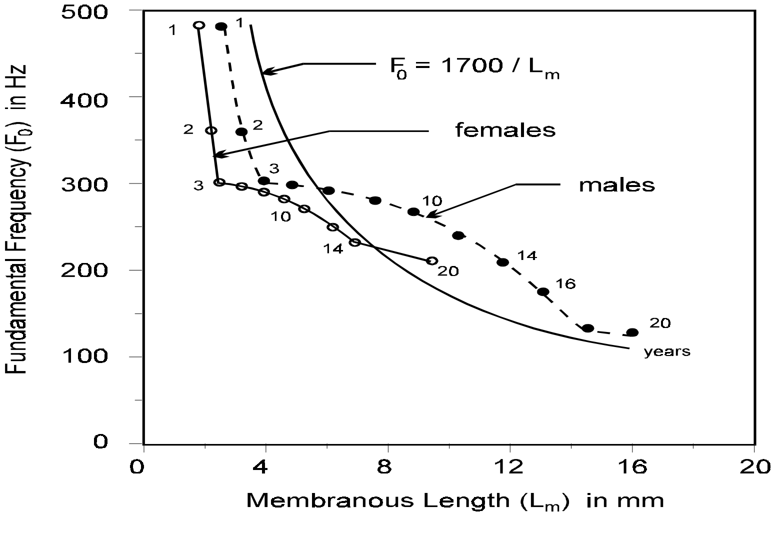

Fig. 1.5 Mean speaking fundamental frequency F_0 as a function of membranous length L_m.

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 modeled 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.

2 Feature Parameters

2.1 Parameters for Speaker Recognition

Speaker identity is correlated with the physiological and behavioural characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. Although it is often impossible to separate these kinds of characteristics, and many voice characteristics are difficult to measure explicitly, many characteristics are captured implicitly by various signal measurements. The most basic type of parameters used for voice recognition [13] are either quantifiable by a human listener or have been borrowed from systems for speech coding, recognition or synthesis.

The first type of machine speakers recognition using spectrograms of their voices, called voiceprint analysis or visible speech [6], was begun in the 1960s. The term voiceprint was derived from the more familiar term fingerprint. Voiceprint analysis was only a semiautomatic process. First, a graphical representation of each speaker’s voice was created. Then, human experts manually determined whether two graphs represented utterances spoken by the same person. The graphical representations took one of two forms: a speech spectrogram or a contour voiceprint [1].

Figure 2.1 illustrates a comparison of spectrogram variation within one speaker and between two speakers. The word „alarm“ was normally spoken twice by one male speaker and once by another male speaker.

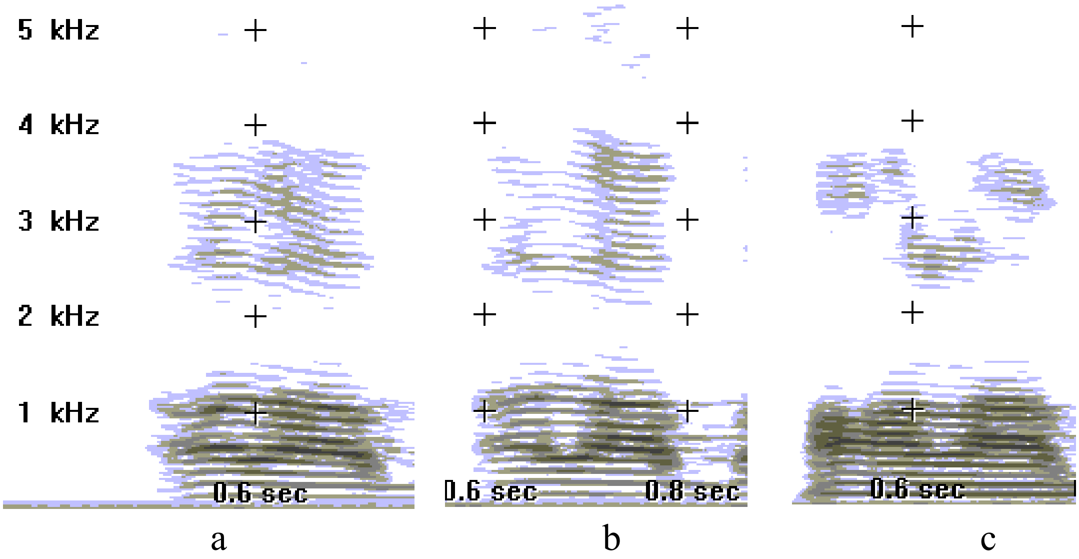

Fig. 2.1 An example of spectrogram variations of the „ala“ sequence (cut out from the word „alarm“) twice for the same speaker a), b) and once for another speaker c).

Both speech and speaker recognition rely primarily on spectral features, but speaker recognition makes more use of prosodics (F_0 in particular) than speech recognition does. Mean F_0 averaged over all test data from an unknown speaker is frequently used as a simple feature to classify speakers coarsely into broad groups (e.g., male, female, children).

Following are the most significant acoustic parameters used to characterize and control different types of voices:

1) Fundamental frequency F_0 (definition and details can be found in [4])
2) Fundamental frequency changes

Variation of fundamental frequency during an analysis frame is a source of modulation aperiodicity. This situation is normal in natural speech due to intonation. The effect of these variations can be described using a set of synthetic signals, in which the slope of fundamental frequency is varied in the range 0 - 24 semitones per second in seven steps: 0, 1.5, 3, 6, 12,18, 24.

3) Harmonic-to-noise ratio *HNR*

The *HNR* is defined as the periodic components to aperiodic components energy ratio level.

4) Noise burst duration

The noise burst is modeled as a gated white Gaussian noise signal. This choice for the noise source is based on synthesis and perception experiments [5]. Glottal turbulence noise is commonly assumed to result from a combination of high air flow velocity and imperfect glottis closure, and can be more or less modulated.

5) Jitter

The jitter is defined as the maximum perturbation of fundamental frequency. Jitter values are expressed as a percentage of the duration of the pitch period. Large values for jitter variation may be encountered in pathological voices. However, jitter in normal voices is generally less than 1% of the pitch period. Jitter appears a very significant source of aperiodicity in the speech signal. It is generally known that the effect of jitter on the spectra of voiced speech is to widen the harmonic peaks [5].

6) Shimmer

   The shimmer represents the maximum range of peak amplitude change in the signal and thus the maximum variation in peak amplitudes of successive pitch periods. Large values for shimmer variation may be encountered in pathological voices. However, shimmer in normal voices is generally less than about 0.7 dB. The effect of shimmer appears less important than the effect of jitter on the spectrum and on the perceived aperiodicity.

Spectral features in specific sounds tend to be very useful for voice recognition, e.g., formants F_2 - F_4 in vowels and nasals. Vowels, nasals and fricatives (in decreasing order) are commonly recommended for voice recognition because they are relatively easy to identify in speech signals and their spectra contain features that reliably distinguish speakers. Nasals have been of patricular interest because the nasal cavities of different speakers are distinctive and not easily modified (except via colds). One study found nasal coarticulation between „m“ and an ensuing vowel to be more useful than spectra during nasals themselves [17].

The current most commonly used short-term spectral measurements are LPC-derived cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed form LPC coefficients and therefore provides a stabler 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 are extracted at every frame period to represent spectral dynamics (the delta- and delta-delta-cepstral coefficients).

2.2 Evaluation of Parameters

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 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 intraspeaker variances:

$$F = \frac{\textit{Variance of speaker means}}{\textit{Mean intraspeaker variance}} \tag{2.1}$$

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

$$F_n = \frac{\frac{J}{I-1}\sum_{i=1}^{I}\left(S_{i,n} - U_n\right)^2}{\frac{1}{J-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{i,j,n} - S_{i,n}\right)^2} \quad , \tag{2.2}$$

where $x_{i,j,n}$ 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.

$$S_{i,n} = \frac{1}{J}\sum_{j=1}^{J} x_{i,j,n} \tag{2.3}$$

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

$$U_n = \frac{1}{I}\sum_{i=1}^{I} S_{i,n} \tag{2.4}$$

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 Long-Time Spectrum of Vocal Tract

As text-independent features, long-term sample statistics of various spectral features, such as the mean and variance of spectral features over a series of utterances, are used. However, long-term spectral averages are extreme condensations of the spectral characteristics of a speaker's utterances and, as such, lack the discriminating power included in the sequences of short-term spectral features used as models in text-dependent methods.

3.1 Estimation of Vocal Tract Long-Time Spectrum

The procedure for determining the speaker-specific vocal tract spectrum is based on the LPC approach [14]. First, we compute the autocorrelation coefficients $R_j(k)$ for the j-th frame of speech signal and the average autocorrelation coefficients

$$\overline{R}(k) = \frac{1}{J}\sum_{j=1}^{J} R_j(k) \tag{3.1}$$

corresponding to the whole vocabulary formed by J frames. Thus, from the average autocorrelation coefficients, we get the predictor coefficients $\overline{a}_m$ using the Durbin's recursive procedure [8] and then the average LPC-based spectrum using

$$S(f) \;=\; \left|\frac{1}{1-\sum_m \overline{a}_m z^{-m}}\right|^2_{z=\exp(j2\pi f/f_s)} \tag{3.2}$$

where $m = 1, 2, ..., M$ is limited by the order M of the predictor and f_s denotes sampling frequency.

The speech data used in the experiment described below were recorded with an electret microphone. The speech signal was sampled at 22 kHz using a 16-bit A/D converter under laboratory conditions over a period of five months. A group of 26 speakers (19 male, 7 female) aged 20 to 25 years took part in the research, the speaker's nationalities were Czech and Hungarian.

First, the order of LPC spectrum needed for vocal tract description was investigated. Using a set of sixteen LPC coefficients, $\overline{a}_1$ through $\overline{a}_{16}$, the accuracy of long-time LPC spectrum was measured. The LPC orders of 6, 8 or 12 seems to be more appropriate orders of the LPC model, considering the accuracy of represented spectrum and the computational volume needed to obtain the spectrum.

An important factor for the accuracy of vocal tract spectrum estimation is the needed speech duration. Duration refers to how much of the training/test data must be used to eliminate the text-dependent effect on the variation of the average spectrum. As an example, we present spectrum accuracy as a function of speech duration in Fig. 3.1. The solid and dotted curves correspond to the Czech and the Hungarian text spoken by the same speaker (native Hungarian living in the Czech Republic). Both curves differ in details but tend to the same contour.

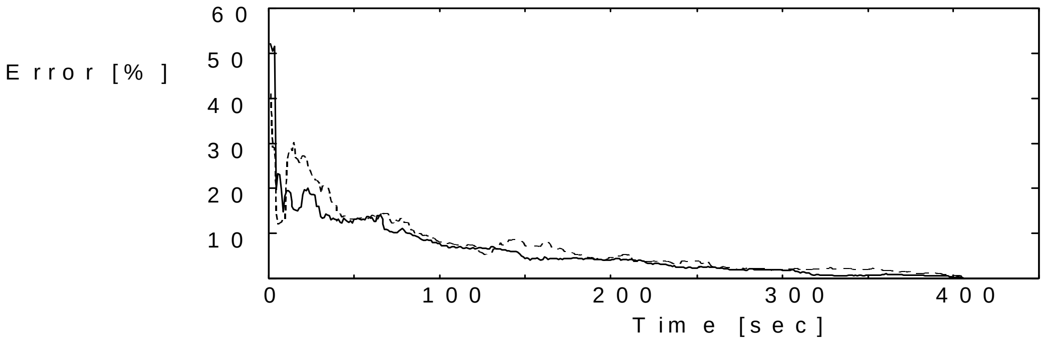

Fig. 3.1 Long-time spectrum accuracy as a function of speech duration (LPC, *M*=6).

A comparison between intra- and inter-speaker variability in long-time spectrum is shown in Figures 3.2 and 3.3. Fig. 3.2 illustrates two vocal tract spectra of the same speaker corresponding to two different texts. The difference between both curves is 12%.

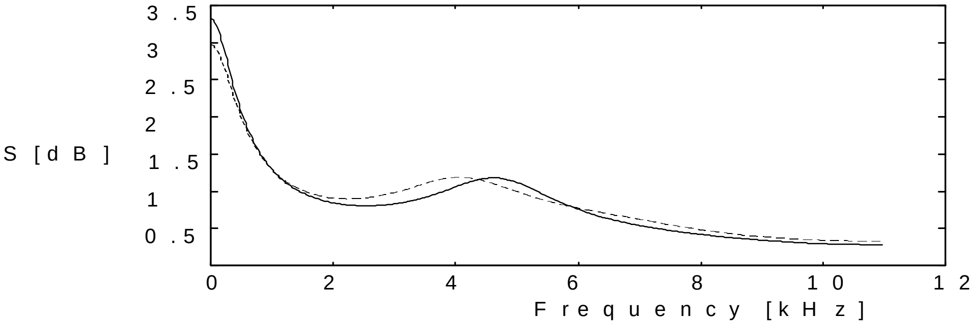

Fig. 3.2 Long-time spectrum difference of one and the same speaker (LPC, *M*=6, 100 sec).

Vocal tract spectra obtained from two different speakers saying the same text is shown in Fig. 4.8. The difference between both curves increased to 22% in this case. The average intra-speaker difference over all speakers was 12.6%, while the average inter-speaker difference (gender-specific) reached 23.4%. In accordance with the inter-gender differences, the estimated difference between the two groups of speakers (male and female) was more apparent (29.6%) than within the groups.

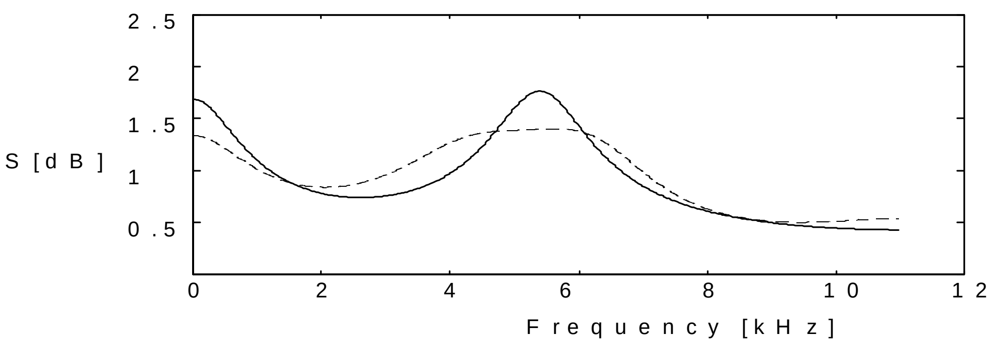

Fig. 3.3 Long-time spectrum variability between speakers (LPC, *M*=6, 100 sec).

3.2 Speech Normalization by Long-Time Spectrum

An important aspect of the described long-time spectrum is that it also offers a potential tool for speech normalization applicable to speaker-independent speech recognition [11]. To normalize the speech signal by LPC long-time spectrum, we can transform the autocorrelation coefficients $R_j(k)$ on each frame into the form

$$R_j^n(k) = R_a(0)R_j(0) + \sum_{m=1}^{M} R_a(m)\left[R_j(|k-m|) + R_j(|k+m|)\right] \; , \tag{3.3}$$

where

$$R_a(m) \; = \sum_{i=0}^{M-m} \bar{a}_i \, \bar{a}_{i+m} \; . \tag{3.4}$$

The normalized autocorrelation coefficients $R_j^n(k)$ can then be used to evaluate various sets of parameters for speech recognition. Figure 3.4 illustrates the effects of the normalization by long-time spectrum for the spectrum of vowel „a“ cut out from continuous speech. Solid line shows the spectrum before normalization, dotted line after normalization.The formant peaks of normalized speech are weighted more heavily and thus represented more accurately.

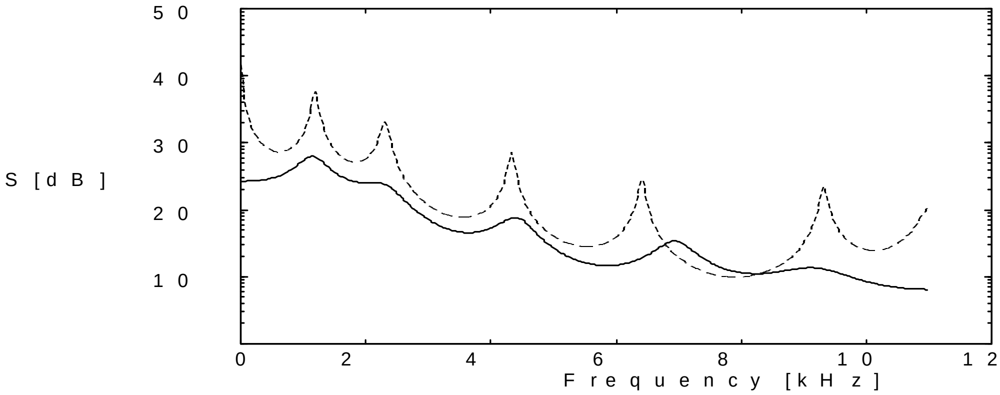

Fig. 3.4 Effect of speaker normalization on the spectral function for phoneme „a“.

3.3 Conclusions

- Long-time spectra can yield high recognition accuracy for normal speech but not for speech spoken under stress and for disguised (impersonated) speech.
- Results show that it is possible to use long-time spectra models across languages for normally spoken speech.
- To estimate relevant long-time spectra with respect to their computational simplicity, a set of 6, 8 or 12 LPC coefficients and speech of about 100 seconds in duration seem to be sufficient.
- Long-time spectra used for speech normalization can bring better formant localisation and increased performance of word recognition systems.

4 Effects of Emotional Stress on Speech

Voice has been shown to be a reliable indicator of speaker's internal state. Mood, emotion, personality and other pragmatic information about the state of the speaker are present in every spoken utterance. At present, interest in this area of research is increasing as the number of potential applications grows and vocal emotions have also tended to be studied as a separate topic.

4.1 Stressed Speech Data

It is really difficult to obtain realistic voice samples of speakers in various stressed states, recorded in real situations. There are not many corpora designed to allow the study of speech under stress [3]. A typical corpus of stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. For the Czech language, no research in emotional speech is known and no appropriate public database exists.

However, for our studies conducted within the research of speech processing in noise and stress we used our own database consisting of data collected during oral final examinations at our Institute of Radio Electronics [15]. The training data in the experiments were extracted from ca. 12 hours of raw conversational male speech (mostly answers). The recorded utterances were manually examined. This material contains stressful phases (improvisations relating to unknown technical problems) and other phases with lower stress (during discussions relating to known technical problems mainly in the final stages of the examination). The recording platform is set up to store the speech signals „live“ in 16-bit coded samples at a sampling rate of 22 kHz.

All the 14 speakers in the database are students finishing their university studies aged about 25 years, non-smokers, have no speech, language or hearing difficulties, and are Czech natives speaking with standard Moravian accent. 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. Further, short portions of 1 to 2 minutes of fluent stressed speech were selected, cut out and written down. A few days later, the same speakers read this written text.

4.2 Detection of Stressed Speech

To get the quantitative changes of speech parameters, we applied in this study some simple features that had not been specifically designed for the detection of stressed speech, such as vowel duration, formants and fundamental frequency. These features were measured in normal and stressed speech, and obtained values were then compared [16].

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 vowel „e“ seems to be irrelevant for the detection of emotional stress (Fig. 4.1).

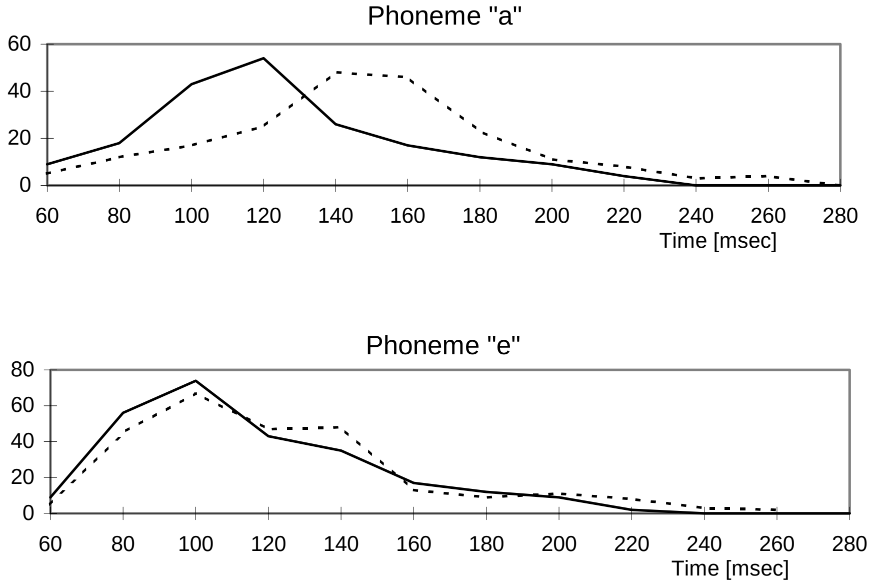

Fig. 4.1 Distribution of duration for vowels „a“ and „e“ (the solid lines are for normal speech, the dotted lines for speech under stress).

In general, more significant results are given by formants. 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 4.1 illustrates 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

Tab. 4.1 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. 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. Figure 4.2 illustrates the F_0 distribution. The curves are comparable because they were obtained from speaking/reading the same text.

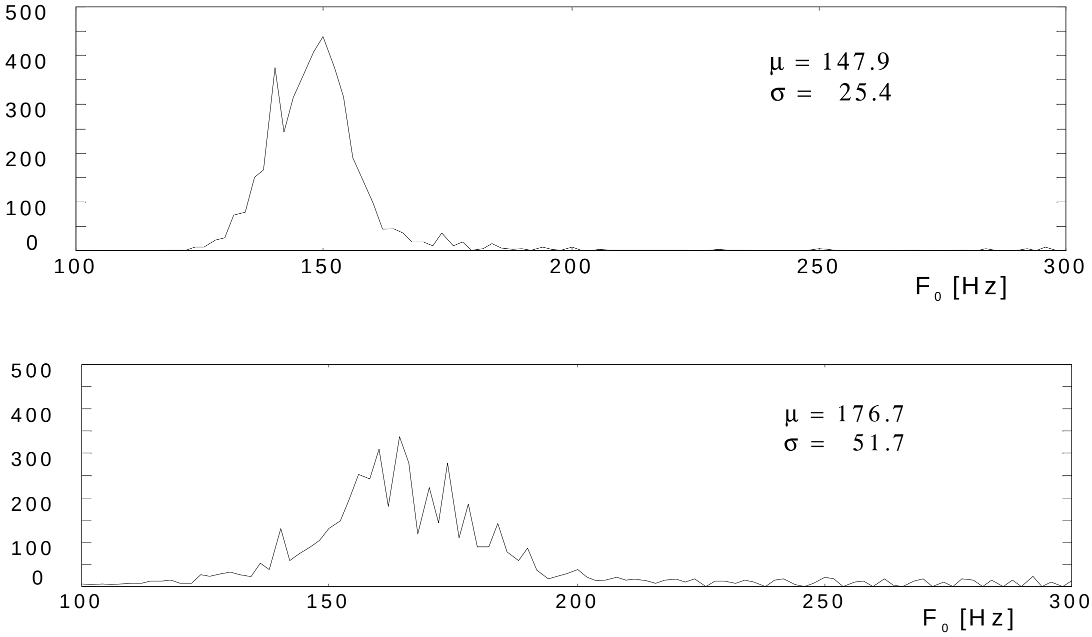

Fig. 4.2 Pitch distribution for a male speaker (top graph is for normal speech, bottom graph is for speech under stress).

The effect of changes in speech due to the emotional state of speaker on long-time spectrum can be observed in Fig. 4.3. 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).

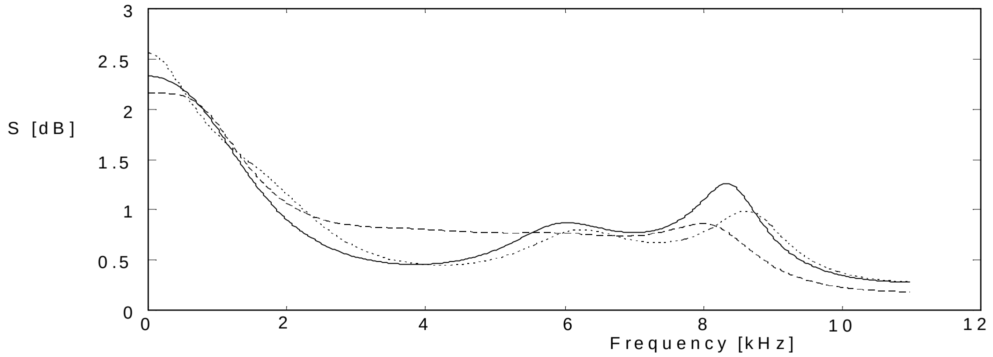

Fig. 4.3 Long-time spectrum variability within speaker for normal and emotional speech (LPC, M=8, speech duration 114 sec).

4.3 Conclusions

- Emotional stress is essentially characterized by an increase in the first and second formant frequencies of the vowels in stressed speech.
- Fundamental frequency F_0 may be used as significant stress indicator, both its mean value and variance increase when the speaker is involved in a stressful activity.

5 Conclusion

The thesis submitted deals with speaker recognition by his/her voice trying to comprehend as far as possible all the important aspects regarding this theme. A lot of the thesis can be seen as a report about the state-of-the-art. In the new portion, our research experiments are described and the results obtained are presented. Because of differences among national languages and various „speaking behaviour“ relating to each nation, it is necessary to perform research for each language separately.

In summary, the following general conclusions can be drawn from the experiments and data mentioned in the habilitation thesis: In broad groupings the vowels, liquids and nasals are found to provide the best speaker recognition performance, followed by the fricatives and affricates, with the plosives providing the worst performance of all. Experiment results show the effectiveness of using the information in the chosen individual phonemes for specific tasks, e.g. „a“ for sex identification, „r“ for analysis of alcoholic speech. The log area ratio coefficients, mel-cepstra and line spectral-pair frequencies appear to be best suited for discriminating speakers. The speech signal could by used as possible indicator for stress and alcohol consumption. In our future research we will try to assess objective measures of these factors. We show that all of the speaker-related problems can be effectively handled by the use of phoneme-based acoustic analysis in the field of LPC-derived features. Finally, it should be noted that automatic speaker recognition is extremely sensitive to noise and channel effects.

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. It is very important to investigate feature parameters that are stable over a long period, insensitive to the variation of speaking manner, including speaking rate and level, and robust against the variation of voice quality such as those due to voice disguise or colds. It is also important to develop a method to cope with problems of distortion due to telephone sets and channels, and background and channel noises.

As part of fundamental research, it is important to pursue a method for extracting and representing the speaker characteristics that are commonly included in all the phonemes irrespective of the speech text. From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. It is expected, that computer power will continue to grow exponentially for at least the near and foreseeable future and that leading researchers will use it productively.

Alphabetical List of Principal References

[1] Boner,A.: Spracherkennung mit Computer. Aarau, AT-Verlag, 1992.

[2] Godfrey,J.-Graff,D.-Martin,A.: Public Databases for Speaker Recognition and Verification. ESCA Workshop, Martigny, Switzerland, 1994, pp. 39-42.

[3] Hansen,J.H.-Ghazale,S.E.: Getting Started with SUSAS. Eurospeech’97, Rhodes, pp. 1743-1746.

[4] Hess,W.: Pitch Determination of Speech Signals. New York, Springer-Verlag, 1983.

[5] Childers,D.G.-Lee,C.K.: Vocal Quality Factors: Analysis, Synthesis and Perception. J. Acoust. Soc. Am., Vol. 90, 1991, No. 8, pp.2394-2410.

[6] Juang,B.H.-Rabiner,L.R.: Fundamentals of Speech Recognition. Englewood Cliffs, NJ, Prentice Hall, 1993.

[7] Lancker,D.-Kreiman,J.-Emmorey,K.: Familiar Voice Recognition: Patterns and Parameters - Recognition of Backward Voices. J. Phonetics, Vol. 13, 1985, No. 1, pp. 19-38.

[8] Psutka, J.: Komunikace s poèítaèem mluvenou øeèí. Academia, Praha 1995.

[9] Reich,A.-Duke,J.: Effects of Selected Vocal Disguises upon Speaker Identification by Listening. J. Acoust. Soc. Am., Vol. 66, 1979, No. 4, pp.1023-1028.

[10] Sigmund,M.: Task Oriented Applications of Automatic Speech Communications. Proc. 7th EAEEIE, Oulu, 1996, pp.77-80.

[11] Sigmund,M.: Speaker Normalization by Long-Time Spectrum. Proc. Radioelektronika’96, Brno, 1996, pp.144-147.

[12] Sigmund,M.: Use of Computers for Innovative Tests in Speech Signal Processing. Proc. 8th EAEEIE, Edinburgh, 1997, pp.F1.7-F1.11.

[13] Sigmund,M.: Speech Analysis for Speaker Identification. Proc. DSP’97. Herlany (Slovakia), 1997, pp. 20-22.

[14] Sigmund,M.-Menšík,R.: Estimation of Vocal Tract Long-Time Spectrum. Proc. Elektronische Sprachsignalverarbeitung, Dresden, Vol. 9, 1998, pp. 69-71.

[15] Sigmund,M.: Parameters of Speech under Stress. Proc. Radioelektronika’99, Brno, 1999, pp. 190-192.

[16] Sigmund,M.: Detection of Stressed Speech. . Proc. Telecommunications and signal processing – TSP’99, Brno, 1999, pp. 121-124.

[17] Su,L.-Li,K.-Fu,K.: Identification of Speakers by Use of Nasal Coarticulation. J. Acoust. Soc. Am., Vol. 56, 1974, No. 5, pp.1876-1882.

[18] Titze,I.R.: Physiologic and Acoustic Differences between Male and Female Voices. J. Acoust. Soc. Am., Vol. 85, 1989, No. 4, pp.1699-1707.

Souhrn

Habilitační práce „*Speaker Recognition - Identifying People by their Voices*“ se pokouší poskytnout ucelený pohled na problematiku rozpoznávání mluvčích podle hlasu zahrnující popis nejčastěji používaných metod, aktuální stav problematiky a možné směry dalšího rozvoje oboru. Vzhledem k rozdílům mezi národními jazyky nejsou všechny metody a výsledky zcela přenositelné a pro aplikační účely je žádoucí provádět výzkum v každém jazyce zvlášť. Předkládaná habilitační práce obsahuje výsledky vývoje získané v českém jazykovém prostředí (i když je práce psaná anglicky).

Pro získání základního přehledu o vlivu různých řečových signálů na efektivnost rozpoznávání a o účinnosti různých metod zpracování bylo provedeno několik statistických měření. Nejdříve byly všechny české fonémy rozděleny do skupin podle příbuznosti parametrů, jednotlivé skupiny pak použity samostatně na rozpoznávání mluvčích a vyhodnocena úspěšnost rozpoznávání. Na fonémech z nejúspěšnější skupiny bylo potom prováděno porovnávací měření jednotlivých metod a opět vyhodnocena úspěšnost rozpoznávání. Pozornost byla věnována také vývoji algoritmů na automatickou fonémovou segmentaci řečového signálu s ohledem na budoucí automatický výběr předem zvolených fonémů z řečového signálu plynulé řeči.

V oblasti analýzy řeči nezávisle na textu bylo vytvořeno dlouhodobé spektrum hlasového traktu mluvčích a zkoumány různé parametry ovlivňující tvorbu a použitelnost dlouhodobého spektra. Získané poznatky byly použity pro odlišení imitátorů hlasu od originálních mluvčích na zkušební množině nahrávek.

V oblasti analýzy řeči v závislosti na textu bylo řešeno několik specielních témat zabývajících se automatickým určováním pohlaví mluvčích z krátkého úseku řečového signálu, určováním vlivu stresu mluvčího na řečový signál a určováním vlivu alkoholu v malém množství na řečový signál.

Některé získané poznatky již byly zahrnuty do výuky zpracování řečových signálů, další výsledky pro rozšíření resp. aktualizaci výuky lze očekávat v oblastech, ve kterých dále pokračuje výzkum, zejména v oblasti určování psychického stavu mluvčích. Otevřeným tématem zůstává zatím možnost použití některých algoritmů vyvinutých pro rozpoznávání mluvčích a jejich stavů také v dalších oblastech zpracování signálu, zejména pro aplikace na jiné biologické signály.

I když byly dosaženy měřitelné výsledky slibující možnost praktického využití, zůstává celá řada nedořešených problémů, které vyplývají ze složitosti procesu tvorby a vnímání řeči. Výraznější pokrok v automatickém rozpoznávání mluvčích lze zřejmě očekávat s uplatněním nových poznatků také z jiných vědních oborů a s dalším rozvojem výpočetní techniky.