CHAPTER I: INTRODUCTION

Modern multimedia systems demand very often specific procedures of people identification and also verification. The most important methods base on iris and finger print analysis. Recently, pattern recognition is also applied in quite new and important domain concerning automatic recognition, understanding and processing of document images. The most representative examples are automatic character recognition, automatic check processing in banking systems, signature verification and direct handwriting recognition.

Multi-layer neural network can be helpful for recognition of handwritten words. Additional analysis shows how to improve proposed solutions using mixed hardware-software co-design. The verification of word recognition can be achieved using fuzzy system for effective comparison process with a given vocabulary. The simplification of recognition process in computer systems gives a lot of redundancy, so fuzzy logic approach seems to be promising solution.

Communication between modern computer systems and users is still a critical problem and needs specific procedures of people verification. A human voice analysis becomes promising solution for people identification in modern telecommunication systems. The automated methods of speaker recognition and emotions verification have potentially different array of application like banking transactions, forensic purpose, shopping using telephone or Internet networks, and other database services. Two important features distinguish voice identification in comparison with other methods. Voice cannot be stolen during verification process and we do not involve our hands.

The main algorithm proposed is based on two different vectors; it means that an emotion will be calculated according to the location of two vectors in the emotion plane. Generally, three kinds of methods can be applied in modern speech processing systems. The first of them takes into account several features of the spectral characteristics. In such a case it is possible to present general attributes of the speaker regardless the individual phoneme features. Proper emotion recognition needs application of statistic methods. For example, it is possible to

apply a Multivariate Auto Regression model (MAR) or Hidden Markov Models (HMM). The second method uses a short-term training feature vectors. Unfortunately, in a real-time emotion recognition the number of training vectors is so large, that the memory and computational time exceed limitations of nowadays computers. Therefore, it is necessary to apply some special solutions like vector quantization (VQ) techniques or HMM-based methods. A VQ codebook consists of small number of very specific but simple feature vectors. Using this codebook, it is possible, clustering these vectors to reflect a given speaker features. It is also necessary to admit that Hidden Markov Models are very promising solution especially using a Gaussian mixture. The third method is based on speech recognition methods. The pronunciation of the same phoneme is different for different speakers; therefore, the emotion recognition can be made using phoneme templates created during training processes.

EMOTION RECOGNITION SYSTEM

Nowadays methods, based on spectrograms of speech sounds, are not faultless. Therefore, further researches leading to improvement of such methods, combining spectral parameters with temporal characteristics are necessary. One of the most important tasks is a proper definition of feature vectors. The following four vectors can be favored as very important and useful: the long-term spectra (LTS) vector, the speaking fundamental frequency (SFF) vector, the time-energy distribution vector (TED) and vowel formant-tracking (VFT) vector.

The LTS vector is one of the most important indicators of voice quality; very helpful for speaker verification. The LTS can be described as high sensitive vector for speaker's identity of low quality voice signals. Therefore, proper results can be achieved using signals containing rather high level of a noise, and limited frequency passband. The application of speaking fundamental frequency can also be a powerful vector. The efficiency of SFF depends however on a processing system, which has to track and extract the fundamental frequencies precisely. As the results, statistical behaviors of these frequencies have to be also included. Fortunately, the vector is very sensitive to different emotions of the same speaker (anger, joy, sadness). The TED vector is also important, because it characterizes a speech prosody, so it is possible to extract an individual

speaker features, like a rate of speaking, pauses, loudness characteristics, etc. The VFT vector depends strongly on the individual size and shape of the vocal tract, it means that you can define this vector by anatomical properties of the tract. The VFT is then stable and independent on almost all types of distortions.

A proper emotion recognition has to avoid of several problems. It is necessary to take into account different sounds becoming from different speakers including noise and channel characteristics. In many cases, especially in banking application, the secure connection is based on speaker recognition. It is the next higher level of speech processing. In such a case we have to identify some specific features of the speaker voice. The identification efficiency depends on the several circumstances. Mainly, it is necessary to take into account the number of potential speakers.

CHAPTER II: LITERATURE REVIEW

2.1 F. Hartung et.al: “**Multimedia watermarking techniques”2004** Multimedia watermarking technology has evolved very quickly during the last few years. A digital watermark is information that is imperceptibly and robustly embedded in the host data such that it cannot be removed. A watermark typically contains information about the origin, status, or recipient of the host data. In this tutorial paper, the requirements and applications for watermarking are reviewed. Applications include copyright protection, data monitoring, and data tracking. The basic concepts of watermarking systems are outlined and illustrated with proposed watermarking methods for images, video, audio, text documents, and other media. Robustness and security aspects are discussed in detail. Finally, a few remarks are made about the state of the art and possible future developments in watermarking technology.

2.2 Enis Cetin et.al: “**Speaker identification and video analysis for hierarchical video shot classification”2014** We present a new video shot classification and clustering technique to support content-based indexing, browsing and retrieval in video databases. The proposed method is based on the analysis of both the audio and visual data tracks. The visual stream is analyzed using a 3-D wavelet transform and segmented into shot units which are matched and clustered by visual content. Simultaneously, speaker changes are detected by tracking voiced phonemes in the audio signal. The clues obtained from the video and speech data are combined to classify and group the isolated video shots. This integrated approach also allows effective indexing of the audio-visual objects in multimedia databases.

2.3 C. Busso ; Chi-Wei Chu ; Soon-il Kwon ; Sung Lee ; P.G. Georgiou ; I. Cohen ; S. Narayanan “**Smart room: participant and speaker localization and identification” 2008** Our long-term objective is to create smart room technologies that are aware of the users presence and their behavior and can become an active, but not an intrusive, part of the interaction. In this work, we present a multimodal approach for estimating and tracking the location and identity of the participants including the active speaker.

Our smart room design contains three user-monitoring systems: four CCD cameras, an omnidirectional camera and a 16 channel microphone array. The various sensory modalities are processed both individually and jointly and it is shown that the multimodal approach results in significantly improved performance in spatial localization, identification and speech activity detection of the participants.

2.4 Hernanz et.al "**Fusion of audio and video based speaker identification for multimedia information access"2004** A method and apparatus are disclosed for identifying a speaker in an audio-video source using both audio and video information. An audio-based speaker identification system identifies one or more potential speakers for a given segment using an enrolled speaker database. A video-based speaker identification system identifies one or more potential speakers for a given segment using a face detector/recognizer and an enrolled face database. An audio-video decision fusion process evaluates the individuals identified by the audio-based and video-based speaker identification systems and determines the speaker of an utterance in accordance with the present invention. A linear variation is imposed on the ranked-lists produced using the audio and video information. The decision fusion scheme of the present invention is based on a linear combination of the audio and the video ranked-lists. The line with the higher slope is assumed to convey more discriminative information. The normalized slopes of the two lines are used as the weight of the respective results when combining the scores from the audio-based and video-based speaker analysis. In this manner, the weights are derived from the data itself.

2.5 M. Weintraub et.al "**Robust text-independent speaker identification over telephone channels"2008** This paper addresses the issue of closed-set text-independent speaker identification from samples of speech recorded over the telephone. It focuses on the effects of acoustic mismatches between training and testing data, and concentrates on two approaches: extracting features that are robust against channel variations and transforming the speaker models to compensate for channel effects. First, an experimental study shows that optimizing the front end processing of the speech signal can significantly improve speaker recognition performance. A new filter bank design is introduced to improve the robustness of the speech spectrum computation in the front-end unit. Next, a new feature based on spectral slopes is described. Its ability to discriminate between speakers is shown to be superior to that of the traditional cestrum. This feature can be used alone or combined with the cestrum. The second part of the paper presents two model transformation methods that further reduce channel effects. These methods make use of a locally collected stereo database to estimate a speaker-independent variance transformation for each speech feature used by the classifier. The transformations constructed on this stereo database can then be applied to speaker models derived from other databases. Combined, the methods developed in this paper resulted in a 38% relative improvement on the closed-set 30-s training 5-s testing condition of the NIST'95 Evaluation task, after cepstral mean removal.

CHAPTER III: AIM AND OBJECTIVE

AIM:

To analyze the emotions of speaker from basic human speech

OBJECTIVE:

1) To describe all the possibilities of modern emotional recognition from speech.
2) To identify the authenticity of the given sample.
3) To identify the speaker.

CHAPTER IV: MATERIALS AND METHODOLOGY

MATERIALS

1. Computer
2. Mic
3. Voice recorder

METHODOLOGY

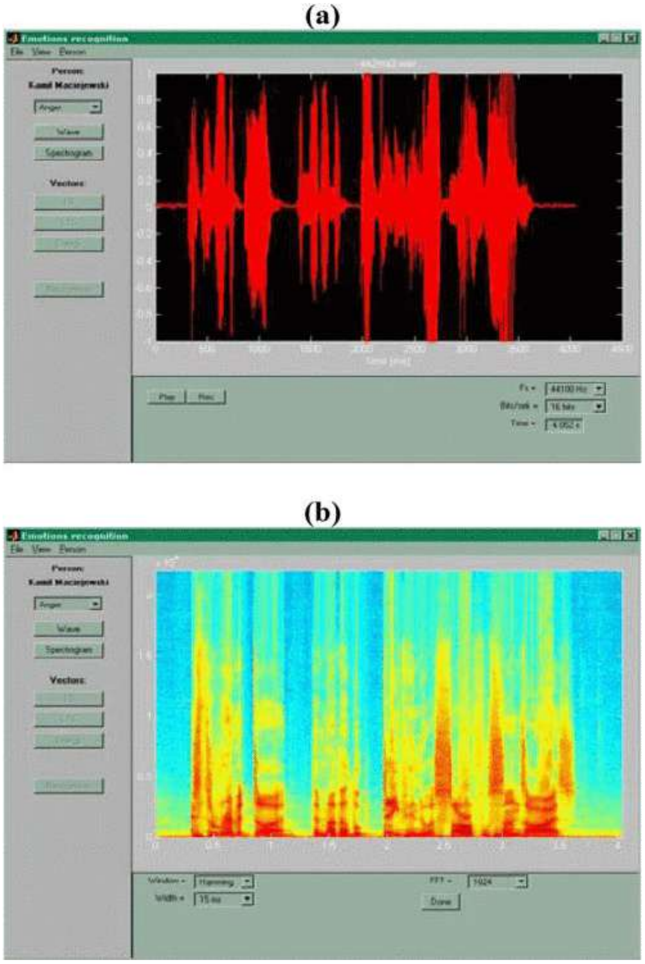
The proposed method permits to compare specific features of the current speaker voice with the computer base containing stored vectors of known speakers. As the result, we obtain the distances between the vectors of current speaking person and the reference vectors for different emotions (neutral, angry, joy and sadness). We can create different feature vectors using time domain speech signal up to 20 seconds. An example of discrete time domain voice signal and corresponding spectrogram.

The program contains the following building blocks:

- registration and playback of the speech signal,
- normalization of the recorded signal,
- spectrogram calculation using fast Fourier transform (FFT),
- calculation of two feature vectors (SFF and TED) and normalization procedures,
- training block, which permits to create the base of referential templates,
- comparison of current vectors with the reference vectors.

The program has been created using MATLAB environment. We applied TED vector and SFF vector for recognition purpose and learning session. Emotions have been divided into four main parts: neutral, angry, joy and sadness. During the training process, the emotion plane has been divided into four regions (corresponding to the above emotions).

CHAPTER V: OBSERVATIONS


(a)

(b)

Fs = 44100 Hz

Bits/sek = 16 bits

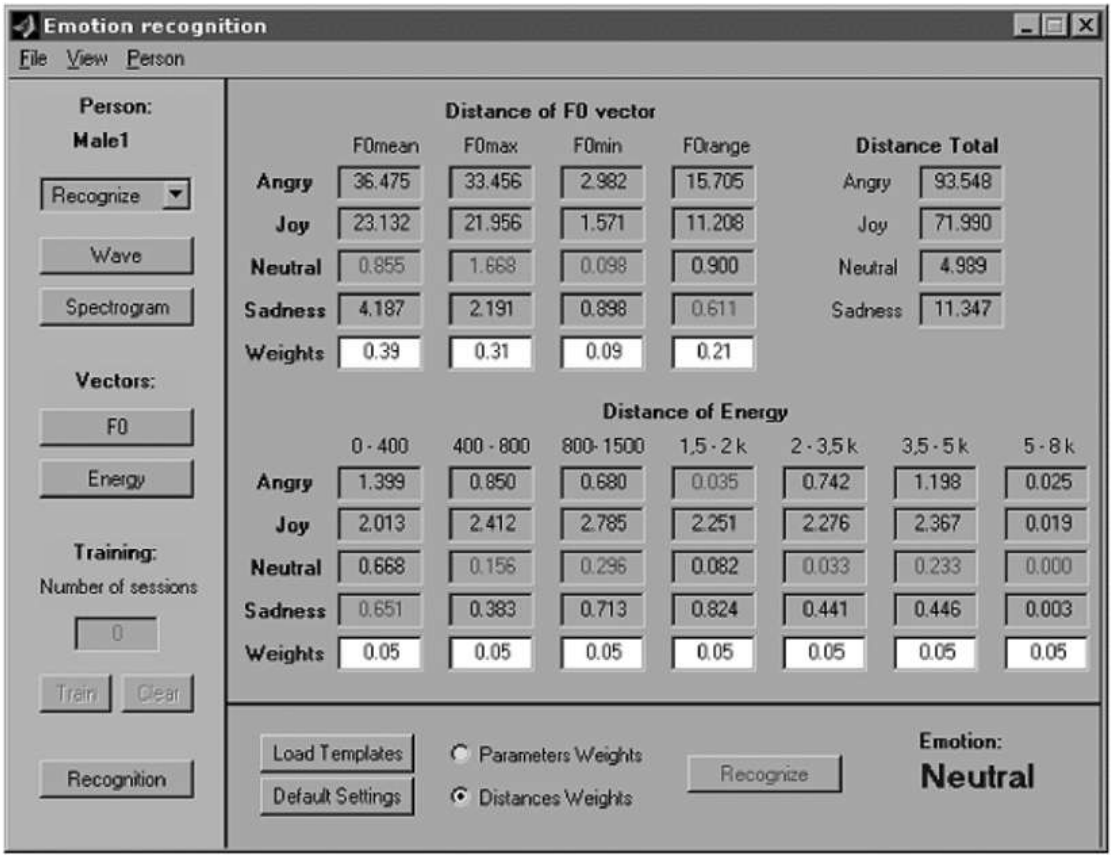
Time = 4.052 s

Window = Hamming

Width = 15 ms

FFT = 1024

Fig. 1. Example of speech processing: (a) - time domain function, and (b) - spectrogram of the above speech



Person:
Male1

Number of sessions: 0

Distance of F0 vector

	F0mean	F0max	F0min	F0range
Angry	36.475	33.456	2.982	15.705
Joy	23.132	21.956	1.571	11.208
Neutral	0.855	1.668	0.098	0.900
Sadness	4.187	2.191	0.898	0.611
Weights	0.39	0.31	0.09	0.21

Distance Total

Angry	93.548
Joy	71.990
Neutral	4.989
Sadness	11.347

Distance of Energy

	0 - 400	400 - 800	800- 1500	1,5 - 2 k	2 - 3,5 k	3,5 - 5 k	5 - 8 k
Angry	1.399	0.850	0.680	0.035	0.742	1.198	0.025
Joy	2.013	2.412	2.785	2.251	2.276	2.367	0.019
Neutral	0.668	0.156	0.296	0.082	0.033	0.233	0.000
Sadness	0.651	0.383	0.713	0.824	0.441	0.446	0.003
Weights	0.05	0.05	0.05	0.05	0.05	0.05	0.05

○ Parameters Weights
● Distances Weights

Emotion:
Neutral

Fig. 2. Emotion recognition – results

CHAPTER VI: RESULT AND CONCLUSION

RESULT

The distance of FO vector for anger is 93.548. The distance of FO vector for joy is 71.990. The distance of FO vector for neutral is 4.989 and for sadness is 11.347. Thus the voice recognition for the given sample is neutral.

CONCLUSION

The results of the emotion recognition are promising. The quality of the program depends on the training processes. Unfortunately, it is difficult to obtain a proper base of voice examples for different emotions. However, we discovered, that our program can recognize false and true emotions as two different states. Moreover, the proposed algorithm can be applied not only for emotion detection but also can be helpful in the process of people identification. The sensitivity of the program for such emotions, like anger or fear is measurable, but the vectors of properties have to be modified. Moreover, a proper distance calculation between vectors of examined person and database is very important task. The implementation of the proposed algorithm as hardware-software system, including mixed analog-digital approach, should improve the speed and also the quality (resolution) of the system.

REFERNCE

1. Z. Ciota, M. Gawinowski, "Multilayer Neural Networks in Handwritten Recognition", TCSET'2004, pp. 24-28, February 2004.

Show Context Google Scholar

2. Z. Ciota, "Improvement of speech processing using fuzzy logic approach", Proc. Joint 9th IFSA Word Congress and 20th NAFIPS International Conference, pp. 727-731, July 25–28, 2001.

Show Context Google Scholar

3. J. Santon et al., Progress in speech synthesis, New York:Springer, 1996.

Show Context Google Scholar

4. Z. Ciota, "Speaker Verification for Multimedia Application", IEEE International Conference on Systems Man and Cybernetics, pp. 2752-2756, October 10–13, 2004.

Show Context Google Scholar

5. Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, Chungyong Lee, "Optimizing Feature Extraction for Speech Recognition", IEEE Trans. Speech and Audio Proc., no. 1, pp. 80-87, January 2003.

Show Context Google Scholar

6. K. Maciejewski, "Emotion recognition using speech signal", 2004.

Show Context Google Scholar

7. M. Jankowski, Z. Ciota, A. Napieralski, "Methodology of CMOS VLSI design using current mode approach", 7th International Conference: “Mixed Design of Integrated Circuit and Systems” MIXDES 2000, pp. 457-461, 15–17 June 2000.

Show Context Google Scholar

8. A. Rominski, Z. Ciota, "Wideband Sound Compression for Multimedia Application", WSEAS Conferences, pp. 391-394, October 2003.

Show Context Google Scholar