System supporting speaker identification in emergency call center

Jakub Gałka, Joanna Grzybowska, Magdalena Igras, Paweł Jaciów, Kamil Wajda, Marcin Witkowski and Mariusz Ziółko

AGH University of Science and Technology
Faculty of Computer Science, Electronics and Telecommunications, Kraków PL–30059
{jgalka|gjoanna|migras|witkow|ziolko}@agh.edu.pl; {jaciow|kamylus}@student.agh.edu.pl

Abstract

A supporting system of voice analysis for emergency call centers is being developed at AGH University of Science and Technology in Krakow. The aim of our work is to provide an innovative supporting tool for rapid and accurate assessment of caller profile. The project covers not only speaker identification (when speaker's speech sample is known), but also speaker's gender and age detection, recognition of emotions, recognition of acoustic background. The system consists of: speech signal analysis, voiceprints learning, adaptation and classification.
Index Terms: speaker recognition, emotion detection, age detection, acoustic background detection, emergency call center

1. Introduction

To get help in emergency situations, we usually call a public-safety answering point (PSAP) in the first place. This essential service has to be very fast and as reliable as possible given current technology capabilities. A huge effort have been already made to improve quality of public safety with speech technology (see [1, 2], e.g. systems like [3]).

Experience and abilities of operators are crucial during emergency notification. This mentally exhaustive work leads to loss of concentration. During a single call, emergency responders have to process a lot of information quickly, including creating a description of a notification. Therefore any system that effectively supports such a process, reduces call time and increases efficiency of operators. Consequently they may concentrate on the conversation and provision efficient help to citizens in emergency situations. For example, multiple calls from one person during handling an emergency situation can be automatically recognized and the data form can be initially fulfilled. What is more, the system can help in early detection of speakers that have frequently abused PSAP services.

2. Description of the system

The proposed system is able to identify caller or describe his/her profile. Additionally, it registers and analyzes acoustic effects that people may not notice in the recorded voice and in the acoustic background. The following caller information is extracted:

- identity;
- gender and age;
- emotional state;
- substance intoxication;
- acoustic background.

The system operates on an acoustic signal obtained from telecommunication channel. This requirement has direct influence on a quality of analyzed signal (acoustic band limited to $300 - 3400\ Hz$, minimal bitrate 12 $kbps$ and reduced amount of information due to lossy compression). Secondly, a speech signal is mostly distorted with environmental background noise.

Based on the preliminary tests, we concluded that a 20 sec long speech sample is sufficient to create an accurate voiceprint of a calling person and 4 sec long is enough to identify one.

The software can be integrated with existing software or work as a separate tool. An automatically recognized profile of a caller is presented directly to a responder via designed interface and stored in a dedicated database.

2.1. Graphical User Interface

Since the system is assumed to facilitate the work of a PSAP responder, Graphical User Interface (GUI) is very clear, ergonomic, responsive and quickly provides information that can be read intuitively with a minimal effort. To satisfy those assumptions, GUI (Fig. 1) includes more graphical elements like icons and colorful images than text and numeric values. Diversified quality of acquired signal has a strong impact on recognition accuracy. Hence, it is necessary to include confidence indicators for each recognized characteristic. After signal acquisition, the results of recognition are displayed with about 2 sec delay.

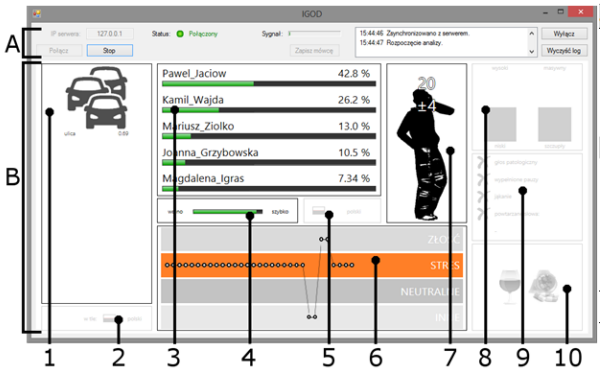

Figure 1: *System GUI: A – communication panel, B – results panel (1 – acoustic background, 2 – language in the background, 3 – identity, 4 – speech rate, 5 – language, 6 – emotional state, 7 - gender and age, 8 – physical features, 9 – voice characteristics, 10 – degree of intoxication).*

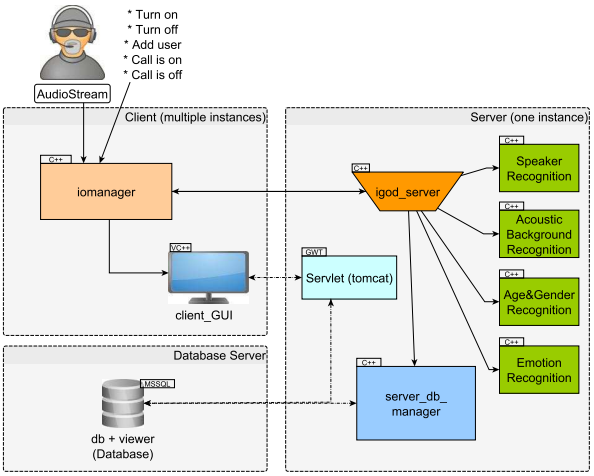

Figure 2: *System architecture*

2.2. Architecture

The system is designed in a star topology, which assumes one server instance and multiple client instances connected inside of a local network (Fig. 2). Server, as a heart of the system, is responsible for analysis of the signal obtained from each client and storing results into a database. During a call, client application gathers audio samples directly from an audio device or from VoIP gateway using Session Initiation Protocol (SIP), transmits it to the server and presents results of analysis in GUI. The visual representation of results is generated by servlet using data from the database and transmitted to each client over HTTP. This architecture was designed in cooperation with end-users to best suit their needs.

2.3. Speech processing

Automatic speaker recognition component provides speaker enrollment and identification functionalities. The enrollment process creates compact speaker model (voiceprint) that can be used to discriminate one speaker from another or to identify or verify the speaker. Voiceprints are created in three major steps: pre-processing, parameterization, stochastic and scattered modelling.

In the pre-processing phase, voice activity detection (VAD) is performed using 4 Hz frequency energy modulation and energy variance analysis in different acoustic bands. Speaker-dependent features are extracted in parameterization step. Our system uses Mel frequency cepstral coefficients (MFCC) and features obtained from Psychoacoustic Wavelet Fourier Transform (PWFT) [4].

In the phase of stochastic modelling we use Gaussian Mixture Models (GMM) with various number of Gaussian components for different modules; we also use i-vectors. Voice modelling is based on two approaches. First, a GMM-based universal background model (UBM) with Maximum A-Posteriori (MAP) voiceprint adaptation is used. UBM is created using expectation maximization (EM) algorithm from over 100 hours of speech recordings. The second modelling approach utilizes i-vectors using the UBM and total-variability (TV) matrix, created using mentioned recordings.

In the identification process, based on calculated similarity scores, the system decides which user from the database is most likely to generate given unknown voice sample. The process of identification is divided into three steps: pre-processing and parameterization, multiple verification, calculation of the final scores for the group of most probable speakers. Extracted features are used to calculate likelihoods, which determines how similar is a feature vector to voiceprints in the database. Assuming that N voiceprints were analyzed from the database, N likelihoods are calculated in this step. Those likelihoods are then sorted and converted into *0-1* score range.

Algorithms of emotions, gender and age recognition follow the same scenario, but voiceprints represent respectively emotional states (i.e. neutral, anger, stress), gender, and age classes (children, young adults, adults, seniors). The algorithm of background recognition allows to extract acoustic background segments from a signal without a speech. The algorithm is based on feature value thresholding. At the beginning, normalization and the pre-emphasis is applied to entire acquired signal. For every frame of the signal two features are calculated: the spectral centroid and the short-time energy. Vectors of feature values are smoothed with median filtering to avoid random noise. Next, local maxima of feature value histograms are located and the adaptive threshold is set as a weighted average of these maxima. It is assumed that frames with feature values less than threshold include acoustic background. Finally, speech and acoustic background parts of the signal are being separated. Speech-free segments are then identified using GMM-based maximum-likelihood classifier with standard MFCC processing front-end.

The efficiency of the implemented algorithms matches state-of-art systems [5].

3. Conclusions

In this paper we present system that supports PSAP operators. It identifies caller characteristics basing on a voice signal and helps in gathering information which may be unnoticed by responders during an emergency call in real-time scenario. We present methods that allow automatic recognition of identity, emotional state, age and gender, background noise type. The GUI prototype was designed to satisfy the assumptions of high performance, ergonomy and ease of use. In the next version, the system will also be able to recognize psychical characteristics (i.e. height and weight), language used by the speaker and degree of his potential drug or alcohol intoxication.

4. Acknowledgements

The project was supported by the National Research and Development Center granted by decision 0072/R/ID1/2013/03.

5. References

[1] A. Drygajlo, “Automatic Speaker Recognition for Forensic Case Assessment and Interpretation,” In.: A. Neustein, H. A. Patil (eds.), *Forensic Speaker Recognition*, 2012.

[2] S. Mathur, S. K. Choudhary and J. M. Vyas, “Speaker Recognition System and its Forensic Implications,” 2: 723, 2013. doi: 10.4172/scientificreports.723

[3] http://speechpro.com/product/biometric

[4] B. Ziółko, W. Kozłowski, M. Ziółko, R. Samborski, D. Sierra and J. Gałka, “Hybrid wavelet-fourier-HMM speaker recognition”, *International Journal of Hybrid Information Technology* vol. 4, no. 4, pp. 25–42, 2011.

[5] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors”, *Speech Commun.* 52, 1, pp. 12–40, 2010.