Online caller profiling solution for a call centre

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

AGH University of Science and Technology
Department of Electronics, Kraków PL–30059
{*witkow*|*jgalka*|*gjoanna*|*migras*|*jaciow*|*ziolko*}@agh.edu.pl

Abstract

The aim of the described system is to provide an online solution that profiles customers of a call centre. As an auxiliary module it might enhance functionality of modern call centre systems by active voice analysis. Integrated with existing databases, our system allows for analysis of constant and temporal caller characteristics during a call — respectively identity, age, gender, emotional state, speech rate and an acoustic background. The specifically developed tool both shortens call time and enhances the amount of information gathered, and consequently reduces cost and workload of call centre responders.

Index Terms: speaker recognition, emotion detection, age detection, gender detection, acoustic background detection, call centre support, voice biometrics.

1. Introduction

One of the problems encountered by companies or organisations that provide their services to large groups of people is communication with their clients. While the Internet provides an efficient and cost-effective way of interaction, it is still burdened with a lack of reliability in a public domain. Even if a service fails due to conditions independent of a company (like failure of a network or a web-interface at a client's device), it decreases user's satisfaction and may eventually lead to loss of a customer. Secondly, web-based interfaces provide a finite number of options to clients so they cannot satisfy all their extraordinary requests. Usually the solution for those problems is using a direct method of communication — visiting an agency or phone contact. Furthermore the interaction with human being is more natural and puts the responsibility of potential mistakes on a company representative.

As a consequence, call centres still have not been superseded by Internet-based solutions. A telephone interaction between an agent and a customer heavily depends on a human factor. Each call requires a high level of attention from an agent. Therefore, they need frequent pauses. Reduction of this cost by application of machine learning technology facilitates minimizing this factor and increases performance of call centre employees. Usually the first task of an agent during a call is to verify the identity of a customer. When utilizing speaker recognition technology, this task is unnecessary since identification or verification may be performed passively (without customer effort) with voice recorded during a call. Such a solution reduces the time of a single conversation and improves customer experience. Secondly speech analysis permits objective monitoring of caller state during a conversation, which might support agents that are burdened with a subjective method of assessment.

To address these problems we developed a solution that utilizes machine learning techniques to facilitate the work of a call centre agent. The proposed prototype of the system is designed to both identify callers and describe their temporal characteristics as presented in Fig. 1 based on voice analysis. Additionally, it registers and investigates acoustic effects that people may not notice in the recorded voice and in the acoustic background. It is assumed to work as an online solution i.e. to provide information about a customer to an agent during a call. The following characteristics are extracted from a signal acquired from a telephone line:

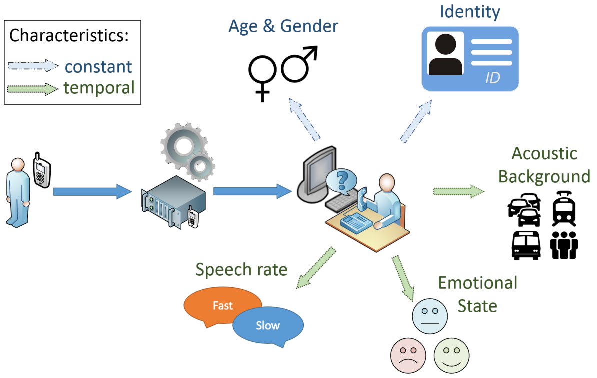

Figure 1: *System use case*

- identity;
- gender and age;
- emotional state;
- acoustic background;
- speech rate.

To the best knowledge of the authors there is no such system that provides an online speech signal analysis for multiple clients working simultaneously. Secondly, the developed system incorporates implementation of state-of-the-art methods with graphical interfaces that work in client-server scalable architecture.

The composition and relation of system components are further explained in section 2. Brief description of implemented methods of signal analysis are presented in section 3.

2. The architecture of the system

The system was developed in a tree topology, which establishes one database server as a root (with possible backup instance) and multiple leaves composed of the computational server and multiple clients. All of the system components should be inside a logical local network. Fig. 2 presents the root and one leaf of

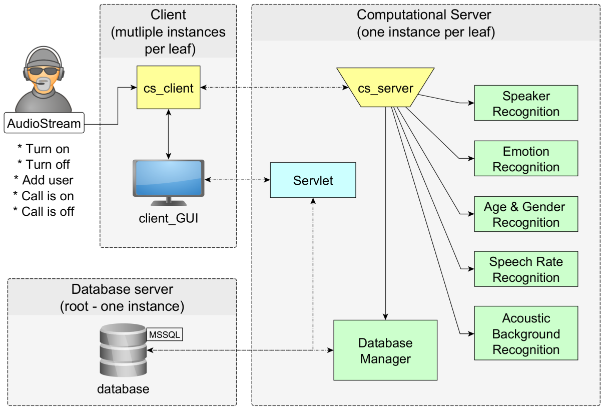

Figure 2: *System architecture with the root (database) and one leaf (the server and multiple instances of client)*

a tree of a system including all of its elements. The scalability of the system is achieved by multiplication of leaves.

Client applications are responsible for gathering signals from VoIP gateways using Session Initiation Protocol (SIP) or directly from an audio input, transmitting signals to the server and displaying call analysis results. Servers receive and analyse audio signals from all clients and send results to the root of the system — the database server. The graphical result of the analysis is generated by a servlet using data from the database and transmitted to each client over HTTP. Each servlet works on the same machine as the computational server.

The proposed solution was designed to be integrated with an existing call centre system or work as a standalone solution.

2.1. Client instance

The aim of the system is to support a call centre's agents, so we assumed that the interface of the system should be ergonomic, clear and require minimal attention from an agent. Therefore, it provides information mostly with graphics, minimizing the amount of text messages. Secondly, the user interface is animated, which allows it to present the characteristics that change in time. Since the system analyses signals from the real telephone line, the results provided to a user have to include the indicators of confidence for each characteristic. Thirdly, according to our empirical tests, the system should present the visual effect of analysis—considered as a consequence of an event that occurred in an acoustic signal—in less than 3 seconds.

2.2. Computational server instance

The server instance has three major tasks — simultaneous analysis of a caller signal from multiple clients, storing its results in a database and providing visual content for each connected client. Of consequence to the first task, the interaction with each client is performed in separate threads. The maximal number of clients connected to the server depends on physical computational resources of the machine that it operates on. Each module presented in Fig. 2 performs the analysis of specific characteristics. Methods and algorithms that have been implemented are described in section 3. The second task is realized by the module that transforms binary data obtained by computational modules into text queries for a database. The visual content for each client is generated using a servlet that retrieves analysis results back from the database, converts it into the web content, which is then provided for each client over HTTP. This solution allows a separation of the presentation layer of the application from the computational engine.

2.3. Database

As a central point of the system, the database sever should base on a stable and reliable database engine. In this particular application, Microsoft SQL has been used. The database comprises three main types of data — viewer data, speaker data and system data. The viewer content consists of tables with data obtained from a call analysis and a call description like time-stamp, caller phone number and description of an agent. This content is retrieved by servlets to generate a visual representation of the results and provide it to clients. The speaker content includes data necessary to perform speaker identification or verification — unique caller identifiers, caller models and caller auxiliary data. The third type of data—system data—includes stochastic models for other modules. When a computational server is initialized, data of the third type is downloaded and stored in a temporary memory.

3. Acoustic signal processing

The system is designed to operate on an acoustic signal transmitted via a telephone line. Consequently, it was adjusted to work on a low quality signal with an acoustic band limited to 300 – 3400Hz, minimal bitrate 12 kbps and degraded by lossy compression applied during a transmission.

3.1. Speaker recognition

The Automatic speaker recognition component provides speaker enrollment and identification functionalities. The enrollment process creates compact speaker model (voiceprint) that is used afterwards in identification. 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 4Hz frequency energy modulation and energy variance analysis in different acoustic bands. Speaker-dependent features are extracted in the parameterization step. Our system uses Mel-Frequency Cepstral Coefficients (MFCC) as features.

In the phase of stochastic modelling we use Gaussian Mixture Models (GMM) and 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 the expectation maximization (EM) algorithm from recordings gathered from multiple databases - RSR2015 [1], TIMIT [2], SITW [3] and some privately collected corpora. The second modelling approach utilizes i-vectors using the UBM and total-variability (TV) matrix, created using the aforementioned recordings.

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

3.2. Age and gender recognition

Algorithms of emotions, gender and age recognition follow the same scenario, but voiceprints represent emotional states (i.e. neutral, anger, stress), gender, and age classes.

In the first step, we perform gender recognition of a caller. The second step is age classification. In both steps we use MFCCs with their first order derivatives as features and i-vector framework.

I-vectors have been previously used for age recognition, although we use them for age classification and not an age regression task like in [4, 5]. We distinguish four age classes: children (age 7-14), young (age 15-24), adult (age 25-54) and senior (age 55-80). Those age classes are motivated by market aspects stemming from the application of our system.

3.3. Emotional state recognition

The emotion recognition module aims at tracking the dynamics of emotional states throughout a conversation. Four categories are detected: neutral state, sadness, fear and anger. The focus is put on negative emotions because they are more neuralgic in the context of proper customer service in a call centre. It would sensitize the call centre agent to situations in which a client is upset or dissatisfied.

To ensure stability and responsiveness of the module, the last three frames of speech are taken into account to make the final decision. The interface displays the history of the indicated emotions from the beginning of the conversation. Intensity of a current emotion marker reflects the confidence level of the recognized category. In the cases of insufficient confidence of categorization, neutral state is displayed.

Models of emotions categories were trained using various corpora of emotional speech, including acted and spontaneous emotions, telephone or studio quality, in different languages (e.g. [6, 7, 8]).

3.4. Speech rate evaluation

Estimation of speech rate is based on Mermelstein's algorithm for segmentation into syllables with use of an acoustic criterion of a syllable [9]. The computed number of syllables per second is then normalized to obtain a value in a scale of 0-1 which is displayed in the graphical interface. The calibration is based on the distribution of speech rate values obtained using over three thousand telephone conversations.

3.5. Acoustic background recognition

This is the only module that does not analyse the speech signal itself — its aim is to detect and classify the acoustic background of a caller. The most frequent background noises occuring in calls, according to their source are: bus, train, crowd of people and cars. The method described below was used to classify these sounds.

At first the background recognition algorithm extracts acoustic background segments from a signal without speech. Extraction is performed using feature value thresholding. At the beginning, normalization and the pre-emphasis is applied to the entire acquired signal, which is then split into non-overlaping frames of 20ms. For each frame two features are calculated: the spectral centroid and the short-time energy. The two feature vectors are smoothed with median filtering in order to reduce 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 where feature values are less than the adaptive 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.

4. Conclusions

In this paper we present the prototype of the system that enables profiling of customers of call centres by online analysis of the acoustic signal. Our solution provides automatic recognition of the individual caller characteristics like identity, gender, age and assessment of behavioural ones, like emotional state and speech rate, based on state-of-the-art methods of speech technology. In addition, it allows for recognition of the background noise type. Described solution has a scalable architecture that includes multiple client applications used by agents, dedicated computational servers and a database sever. Our solution may lead to work optimization, a reduction in the workload of call centre agents and, therefore, reduction of costs for an institution.

5. Acknowledgements

This work was supported by the AGH University of Science and Technology under Deans Grant №15.11.230.252.

6. References

[1] A. Larcher, K.A. Lee, B. Ma, H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56-77, 2014.

[2] J. Garofolo, L. Lamel,W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," DARPA, NIST 1990.

[3] M. McLaren, A. Lawson, L. Ferrer, D. Castan, M. Graciarena, "The Speakers In The Wild Speaker Recognition Challenge Plan,", http://www.speech.sri.com/projects/sitw/, 2015.

[4] M. H. Bahari, M. McLaren, H. Van hamme, D. A. van Leeuwen, "Speaker Age Estimation Using I-vectors", *Eng. Appl. Artif. Intell.*, vol. 34, no. C, pp. 99–108, 2014.

[5] A. Silnova, O. Glembek, T. Kinnunen, P. Matějka, "Exploring ANN Back-Ends for i-Vector Based Speaker Age Estimation", *Proceedings of Interspeech 2015*, pp. 3036–3040, 2015.

[6] M. Igras and B. Ziółko, "Baza danych nagrań mowy emocjonalnej (Eng. Database of emotional speech recordings)", *Studia Informatica* 34(2B), pp. 67–77, 2013.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss: A Database of German Emotional Speech, *Proc. Interspeech*, 2005.

[8] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[9] P. Mermelstein, "Automatic segmentation of speech into syllabic units", *J. Acoust. Soc. Am.*, 58(4), pp. 880-883, 1975.