FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 40 (2015) No. 3
ISSN 0867-6356
DE GRUYTER OPEN
DOI: 10.1515/fcds-2015-0014 e-ISSN 2300-3405

REAL TIME RECOGNITION OF SPEAKERS FROM INTERNET AUDIO STREAM

Radoslaw Weychan, Tomasz Marciniak,
Agnieszka Stankiewicz, and Adam Dabrowski*

Abstract. In this paper we present an automatic speaker recognition technique with the use of the Internet radio lossy (encoded) speech signal streams. We show an influence of the audio encoder (e.g., bitrate) on the speaker model quality. The model of each speaker was calculated with the use of the Gaussian mixture model (GMM) approach. Both the speaker recognition and the further analysis were realized with the use of short utterances to facilitate real time processing. The neighborhoods of the speaker models were analyzed with the use of the ISOMAP algorithm. The experiments were based on four 1-hour public debates with 7–8 speakers (including the moderator), acquired from the Polish radio Internet services. The presented software was developed with the MATLAB environment.

Keywords: Speaker recognition, GMM, Internet radio, ISOMAP

1. Introduction

Speaker recognition is an interesting type of biometric identification, because the speech signal can conveniently be recorded and used as the basis for authorization techniques to access such services as: bank accounts, voicemail, information services, admission to restricted areas, etc.

A process of speaker identification consists of the voice acquisition, analysis, and comparison with a set of voices of the previously registered speakers. Thus, features extracted from the voice of the current speaker form a model, which is compared with all stored models in order to select the most similar model, thus providing a speaker identification.

An approach to simultaneous multi-speaker recognition was already analyzed and described in [1, 15]. However, it is based on estimation of the direction of arrival of the

*Faculty of Computing, Poznan University of Technology, Poznan, Poland, {radoslaw.weychan, agnieszka.stankiewicz, tomasz.marciniak, adam.dabrowski}@put.poznan.pl

speech signal in order to identify the current speaker. Such a solution requires direct access to the recording studio and the use of advanced acquisition techniques, available for the broadcasters only. Additionally, this approach is sensitive to movements of speakers and to acquisition devices. Therefore we cannot use it in our application, as we assume an access to merely the audio streams via Internet.

In our case the speaker recognition methodology should be related to signal encoding techniques. We noticed that by this means the extracted speaker voice features may be less sensitive to the transmission standard, e.g., to the common MPEG 1 Layer 3 (MP3) algorithm [7, 21].

Our goal is also to search for methods that can easily be used in portable hardware devices. It has to be noticed that the autonomous embedded systems [14, 11] should process short audio sequences to get results in real time. The algorithms used to extract voice features and form speaker models should not be computationally demanding in order to make low power consumption feasible.

An example of a such kind of algorithm is Gaussian mixture modeling, GMM. Its use for speaker identification from the spoken documents was already described in [13]. However, such speech differs a lot from the live spontaneous speech in radio. Besides, the authors of [13] focused on the speaker segmentation techniques only.

Nevertheless, the approach presented in [13], although not directly used, was an inspiration to develop a technique described in this paper. This is because the real time application requires an optimized, fast and accurate comparison between the available models. Thus only quite short speech signals could be analyzed, e.g., 1-second long recordings. This is why our research [9] was focused on a possibility of highly efficient (fast and accurate) speaker recognition form short input signals.

The first version of our system was presented in [20]. In the current paper we present an enhanced version of this system and focus on the improvement of speaker recognition accuracy from the speech coded with the use of various encoders typically used for audio data streaming.

2. Speaker recognition fundamentals

Speaker recognition is typically based on individual features extracted from voice. Such systems operate in two main stages. The first one is training and creation of the database of speaker models to be identified. The second and the most important is the testing stage, in which the input signal is processed in order to find the best match among all models stored in the previously prepared database.

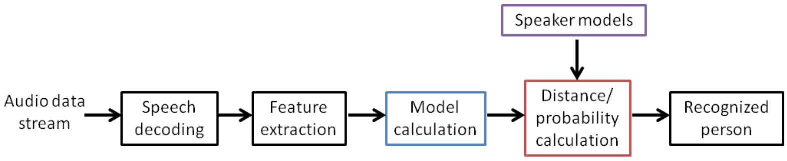

Figure 1: Schema of the automatic speaker recognition system

In both stages many similar internal steps have to be proceeded. At the beginning, specific features are extracted from the input signal. Typically, they are mel-frequency cepstral coefficients (MFCCs) [12]. To extract them, the input signal is blockwise multiplied by a window function (in most cases the Hamming window). Then, the fast Fourier transform is computed and scaled to mel-scale with the use of mel-bank filters. Next, the logarithm of the results of this operation is computed and transformed with the use of the discrete cosine transform (DCT). A set of these (typically 13) coefficients is gathered for few seconds of recordings and modeled, which is the second major operation in both stages. To model $N \times 13$ coefficients, where N stands for the number of frames, most commonly an algorithm based on GMMs is used [16]. The proper model is a sum of weighted Gaussian distributions fitting the histogram of cepstral coefficients.

In the testing stage, the just obtained model should be compared with all other models. This can be done with many metrics, e.g., with: Euclidean distance [5], Mahalanobis distance [8], Kulback-Leibler divergence [6], or log-likelihood metric [2].

The described steps are presented in Fig. 1, which is the commonly used model-based schema for the speaker recognition [9].

During our experiments we used the VOICEBOX speech processing toolbox for the MATLAB environment [4], which includes the most important functions for speech processing related to the speech analysis, synthesis, modeling, and coding. Among them are:

1. MFCC calculation — the ,,Melcepst" function with signal frame and sampling rate at the input and 12 MFCCs at the output

2. GMM formulation — ,,Gaussmix" function with MFCCs and the required number of Gaussians at the input and parameters of the GMMs at the output

3. GMM comparing and log probability calculation — ,,Gaussmixp" function, whose inputs are MFCCs and sets of means, variances, and weights of GMMs to compare.

Figure 2 presents an illustrative distribution of the 1st MFCC and the computed mixture of 16 Gaussians, which fit the distribution of the input data. The data comes from about 8 seconds of speaker voice who participated the public debate used in our experiments. The input signal was divided into frames, 20 ms each. For every frame 13 MFCC was calculeted, and a set of 1st MFCC was modelled in this case.

An acquisition of the MP3 compressed audio stream [3] can be realized with the MATLAB DSP package, i.e., with one of its classes named AudioFileReader. It reads audio samples from the declared input file or a stream. The constructed object can call the "step" procedure, which constantly acquires the declared number of samples being processed in the background. The sample acquisition can also be organized by interrupts with the timer overflow. We have, however, chosen the "step" option because of its simplicity.

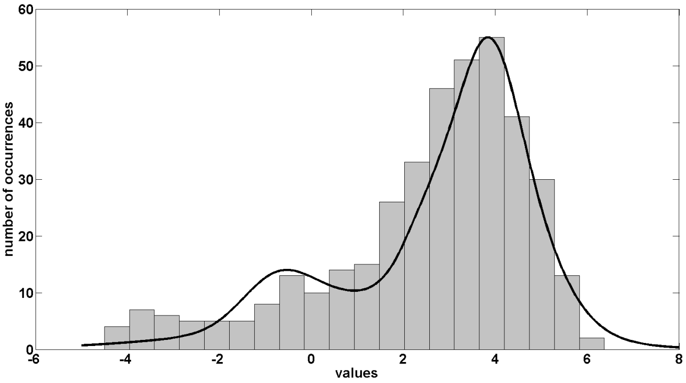

Figure 2: Adjustment of the Gaussian weighted mixture to the input data distribution

3. System description

An idea of the developed system is to acquire and process audio signals in real time from the Internet stream. The prepared software contains two main threads: play the audio stream and process the buffered stream.

The program flow is presented in Fig. 3. The audio signal is decoded from the MP3 data stream and acquired into the buffer of the length being equivalent to 1 second of recording. The audio is constantly played and in the same time it is buffered in order to provide the on-line speaker recognition. To recognize the speaker "on the fly", the original audio stream sampled 44100 times per second is downsampled by factor 4, i.e., to the rate 11025 S/s (samples per second).

In the next step the signal is divided into frames and the MFCCs are calculated for each of them.

Then these MFCCs are expressed with parameters of the Gaussian mixture models and compared with a set of those for the previously defined speakers. The best matches for the current and for the previous speaker are presented in the graphical user interface (GUI), which is updated in real time (Fig. 4).

In order to illustrate distances between the speaker models and their neighbors the ISOMAP [17, 18] visualization has been used. This is a convenient method for nonlinear dimensionality reduction of data lying on a multidimensional manifold. To provide the neighborhood graph, models of MFCC have been computed for all of 15 speakers. Then the logarithm probabilities between each of the models have been calculated to obtain a 2-dimensional probability table (a [15 × 15] matrix). In the next step the same size matrix of the Euclidean distances has been computed, which is just the input for the ISOMAP graph. The calculation scheme is presented in Fig. 5.

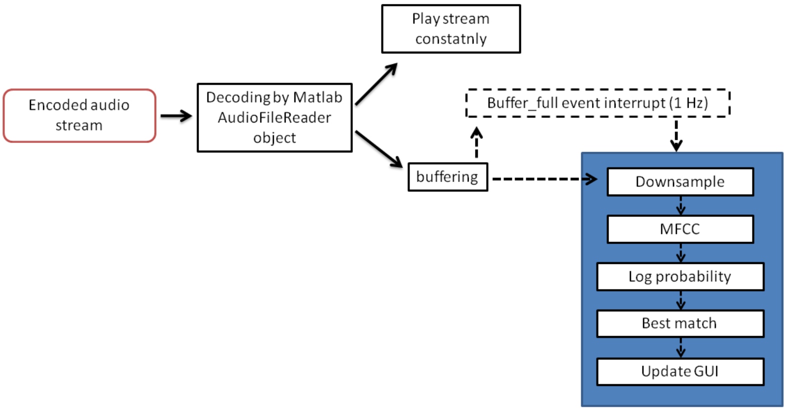

Figure 3: General scheme of the online speaker recognition software

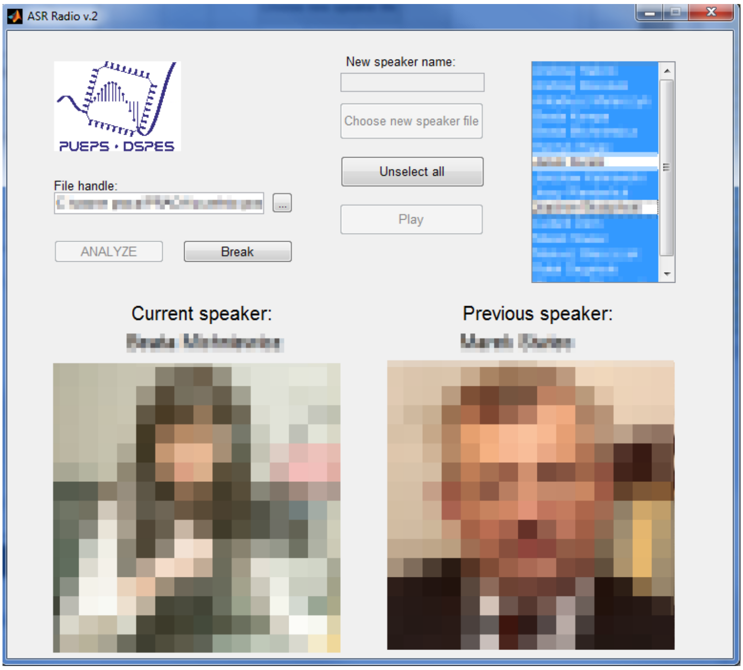

Figure 4: GUI of the prepared software (faces and names are blurred for privacy

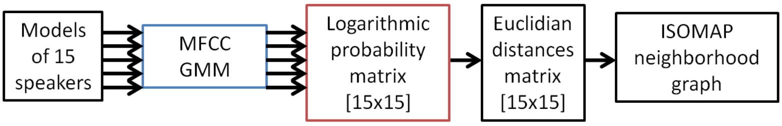
Models of 15 speakers
MFCC GMM
Logarithmic probability matrix [15x15]
Euclidian distances matrix [15x15]
ISOMAP neighborhood graph

Figure 5: Scheme of the ISOMAP neighborhood graph preparation

4. Evaluation of speaker models

As was already stressed, the presented system uses compressed speech signals. Unlike results described in [19, 21], the new approach is adjusted to MP3 streams, which are typical for the compressed Internet audio. Other encoders commonly used for audio streaming over Internet, namely: Vorbis Ogg (the encoder name is Vorbis while Ogg describes the container) and WMA (developed by Microsoft), have also been considered.

To examine how these encoders influence creation of the GMMs for speech signals, distributions of the first MFCC calculated for an illustrative, variously transcoded, 5-second long test speech signal, have been modeled. This signal has been transcoded with the MP3 codec using 5 common bitrates: 192, 128, 96, 48, and 32 kbps, with the Vorbis codecs also using 5 typical bitrates: 160, 120, 96, 48, and 40 kbps, and with the WMA codec using 2 bitrates, namely 48 and 32 kbps.

Results obtained for the MP3 codec are presented in Fig. 6. It can be observed that the lower the bitrate, the smoother is the MFCC distribution. This means that with a too low bitrate, some important features can be lost.

On the top of Fig. 6, the GMMs adjusted with random initial states are shown, while those on the bottom are calculated with the initial states inherited just from the previous models. It should be noticed that the GMMs obtained with the strategy of inherited initial states are much more plausible.

The signal quality can be checked by means of the signal to noise ratio (SNR) calculated between the original and the compressed signal. The obtained results are presented in Fig 7. They show direct relation of the signal quality with the bitrate. However, for the MP3 coder the SNR difference between the most typical bitrates, namely: 128 kbps and 96 kbps is practically unobservable. That is why in a lot of the Internet streaming schemes the latter bitrate is used. However, the tested MP3 recordings were prepared with the larger bitrate equal to 128 kbps.

In Fig 7 it can also be observed that the Vorbis encoder gives better results than the other two tested coders.

5. Experimental results

To perform tests and present the results four ca. 1-hour length recordings of “Breakfast in the studio of the third program of Polish radio” („Śniadanie w trójce” in Polish) have been prepared. These utterances were performed in Polish. In the discussions participated 8 or 9 speakers, including 7 to 8 politicians and one moderator. Overall durations of statements of particular participants should be, by rule, evenly distributed. Our goal is just to check it. Manual check is cumbersome, as the time needed for this task is much greater than the length of the analyzed recording. Our system offers this result in the time almost equal to the duration of the program.

To prepare a database of speakers for the training stage, 5-second long recordings for each speaker have been used to extract cepstral coefficients. In the testing stage (i.e., for the analysis of the programs), 1-second recordings have been used. The

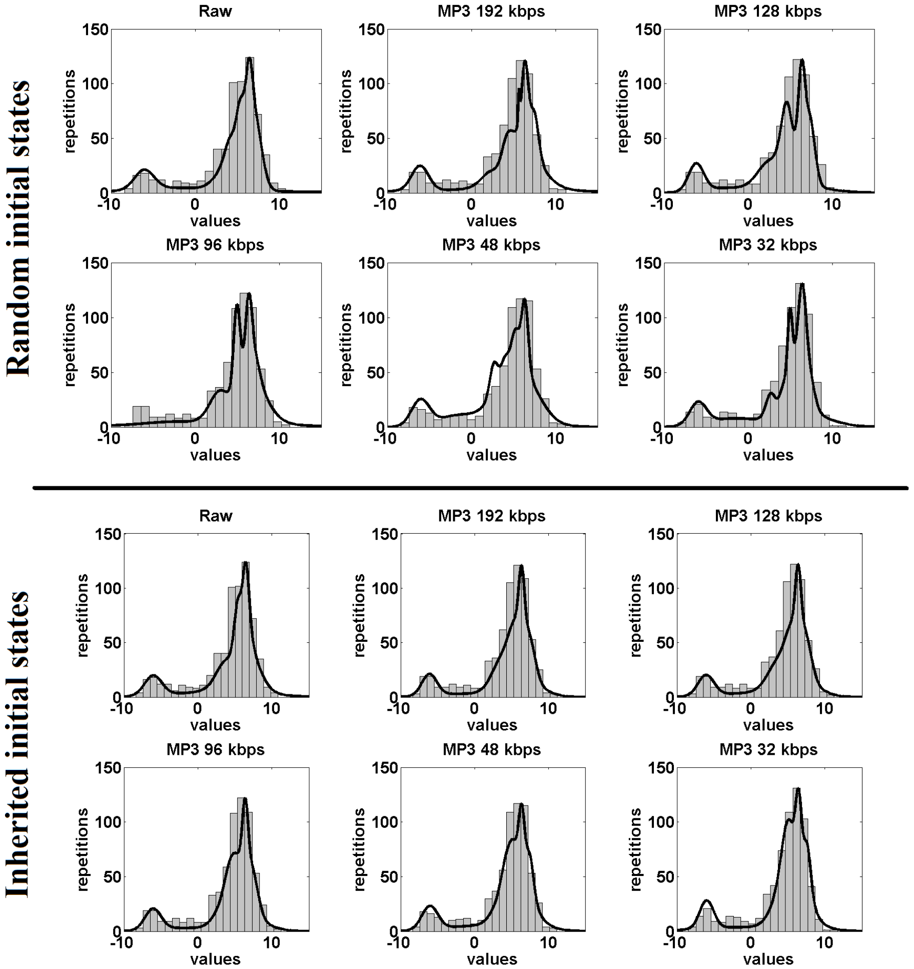

Figure 6: Comparison of GMMs calculated with random initial states (top) and inherited initial states (bottom) for the first MFCC of a 5-second long test speech signal coded with the MP3 standard using 5 common bitrates: 192, 128, 96, 48, and 32 kbps

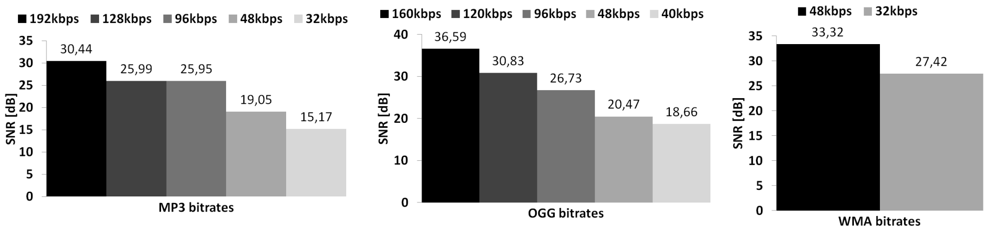

Figure 7: Comparison of SNR values for MP3, OGG, and WMA encoders with various bitrates

usage of such short input signals is determined by the character of the recorded conversations that included a lot of short pauses between words. In longer input signals the instants of silence and the background noise could cause incorrect speaker identification. Moreover, silence can reduce the information content of a long signal. Such aspects are widely analyzed in [10].

In order to illustrate effectiveness of our approach the results obtained for one of the programs (program No. 1) are shown in Figure 8. The obtained results was calculated online and presented just after the end of the debate.

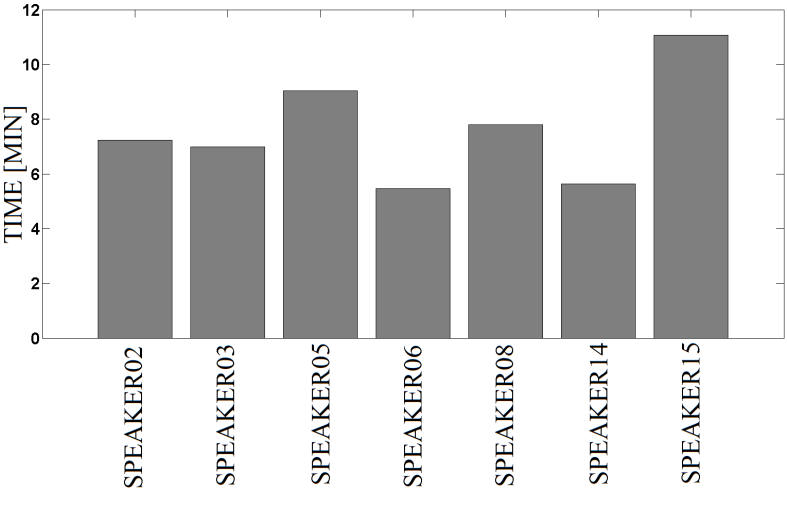

Figure 8: Speakers' activity analysis in program No. 1

A 2-dimensional ISOMAP used to visualize the distances between speech models of all speakers participated in the analyzed debates, is shown in Figure 9.

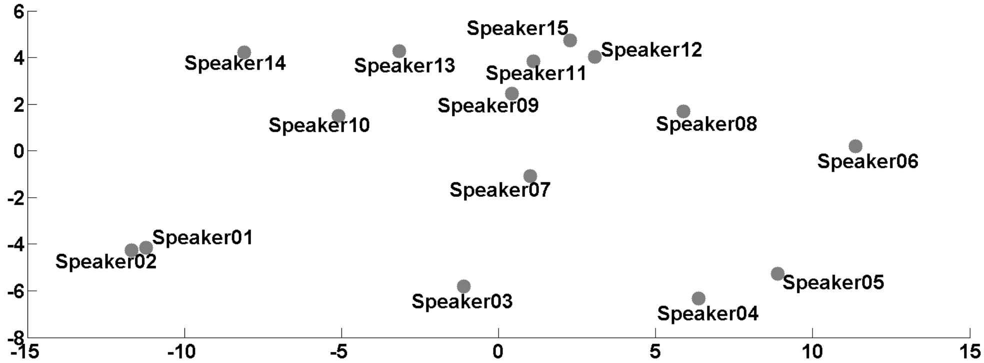

Figure 9: 2-dimensional ISOMAP space of neighborhoods with the normalized scale

6. Conclusion

A proper selection of the speech models is of primary importance for the system effectiveness. In order to obtain repeatable (stable) GMMs, the strategy of inherited initial states should be used.

Although the presented system is not yet equipped with the voice activity detection (VAD) algorithm, it is precise, effective, stable, and the results of the speaker recognition are correct and fully repeatable.

In the end it worth to stress that actually the MATLAB community does not include any solution regarding similar software. We plan to share our application within the MATLAB Central File Exchange to make it publicly available.

Acknowledgment

This work was partly supported by the DS project and partly by Scholarship support for Ph.D. students specializing in major strategic development for Wielkopolska, Submeasure 8.2.2 Human Capital Operational Program, co-financed by the European Union under the European Social Fund.

References

[1] S. Araki, T. Hori, M. Fujimoto, S. Watanabe, T. Yoshioka, T. Nakatani, and A.Nakamura. Online meeting recognizer with multichannel speaker diarization. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 1697–1701, Nov 2010.

[2] D. Blatt and A. Hero. On tests for global maximum of the log-likelihood function. *Information Theory, IEEE Transactions on*, 53(7):2510–2525, July 2007.

[3] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz. ISO/IEC MPEG-2 Advanced Audio Coding. *J. Audio Eng. Soc*, 45(10):789–814, 1997.

[4] M. Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB, 2005.

[5] J. Dattorro. *Convex optimization and Euclidean distance geometry.* Lulu. com, 2008.

[6] J. R. Hershey and R. A. Olsen. Approximating the Kullback Leibler divergence between gaussian mixture models. In *ICASSP (4)*, pages 317–320, 2007.

[7] T. Jiang and J. Han. Map-based audio coding compensation for speaker recognition. *Journal of Signal and Information Processing*, 2:165, 2011.

[8] R. D. Maesschalck, D. Jouan-Rimbaud, and D. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1 – 18, 2000.

[9] T. Marciniak, R. Weychan, A. Dabrowski, and A. Krzykowska. Speaker recognition based on short Polish sequences. *IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings*, pages 95–98, 2010.

[10] T. Marciniak, R. Weychan, A. Dabrowski, and A. Krzykowska. Influence of silence removal on speaker recognition based on short Polish sequences. *IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings*, pages 159–163, 2011.

[11] T. Marciniak, R. Weychan, A. Stankiewicz, and A. Dabrowski. Biometric speech signal processing in a system with digital signal processor. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, Vol. 62, nr 3:589–594, 2014.

[12] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing Mel-frequency cepstral coefficients on the power spectrum. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 1, pages 73–76, 2001.

[13] K. Park, J.-S. Park, and Y.-H. Oh. GMM adaptation based online speaker segmentation for spoken document retrieval. *Consumer Electronics, IEEE Transactions on*, 56(2):1123–1129, 2010.

[14] Z. Piotrowski, J. Wojtun, and K. Kaminski. Subscriber authentication using GMM and tms320c6713dsp. *Przeglad Elektrotechniczny*, (12a/2012):127–130, 2012.

[15] A. Plinge and G. A. Fink. Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5, Sept 2013.

[16] D. Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663, 2009.

[17] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[18] G. Wen, L. Jiang, and J. Wen. Using locally estimated geodesic distance to optimize neighborhood graph for isometric data embedding. *Pattern Recognition*, 41(7):2226 – 2236, 2008.

[19] R. Weychan, T. Marciniak, and A. Dabrowski. Analysis of differences between MFCC after multiple GSM transcodings. *Przeglad Elektrotechniczny*, pages 24–29, 2012.

[20] R. Weychan, T. Marciniak, A. Stankiewicz, and A. Dabrowski. Real time speaker recognition from internet radio. *IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings*, pages 128–132, 2014.

[21] R. Weychan, A. Stankiewicz, T. Marciniak, and A. Dabrowski. Improving of speaker identification from mobile telephone calls. In *Multimedia Communications, Services and Security*, volume 429 of *Communications in Computer and Information Science*, pages 254–264. 2014.

Extended and revised version of work presented at the 18th IEEE SPA Signal Processing Conference, 22-24 September 2014, Poznan, Poland