Real time speaker recognition from Internet radio

Radoslaw Weychan, Tomasz Marciniak, Agnieszka Stankiewicz, Adam Dabrowski
Poznan University of Technology
Faculty of Computing Science
Chair of Control and Systems Engineering
Division of Signal Processing and Electronic Systems
E-mail: tomasz.marciniak@put.poznan.pl

***Abstract*—The paper presents an analysis of speaker activity in online recordings from Internet radio. Proposed system has been developed in Matlab environment. Our research is based on four 1-hour length public debates acquired from the Internet radio. 7-8 speakers participate in the recordings (including one presenter). The speaker recognition was performed on short utterances to facilitate real time processing. The time of speech for each politician has been calculated with the use of gaussian mixture model (GMM) algorithm. Influence of MPEG layer 3 compression algorithm on mel frequency cepstral coefficients (MFCC) has been described. Analysis of neighborhood of speaker models have been done with the use of ISOMAP algorithm.**

***Keywords*—Speaker recognition, GMM, Internet radio, ISOMAP**

I. Introduction

Speaker identification is an interesting type of biometric identification system, since the speech signal can be used as authorization technique to access many services and systems such as: banks, voicemail, information services, even computers or restricted areas. The process of identification consists of analysis of person's voice and comparison to a set of the previously registered speakers. Parameters of the input signal are compared to the reference models to select the most similar one, giving the appropriate speaker ID.

In this paper we present an on-line speaker recognition system that uses the recordings of public debates from the Internet radio. Related research, described in [1], [2], propose system based on estimating the direction of arrival of the signal to identify the current speaker. Such solution requires access to the recording studio and advanced acquisition techniques, available only on the broadcaster side. Additionally, static environment is sensitive to the movement of speakers and acquisition devices. In [3] authors focus on speaker segmentation techniques as preprocessing of speaker identification from spoken documents.

The real time working application requires very fast and accurate comparison between available models, thus it is not an easy task. For the application to work in real time, only very short speech signal can be used for the analysis, e.g. 1 second. Our previous research concerning fast speaker recognition systems [4] confirmed the possibility of accurate recognition form such short input signal with high efficiency.

It also has to be noticed, that in the case of application to stand-alone embedded systems [5], not only short sequences should be processed, but also algorithms used to extract features and modelling shouldn't be complex to provide real time processing. Thus our goal is to use methods can be easily moved to hardware solution.

II. Speaker recognition fundamentals

Speaker recognition systems base on individual features extracted from voice. The system works in two main stages:

1) Training stage - in this step a database of speakers models is created
2) Testing stage - the main step, where input signal is processed to find best match while comparing to database models

In both stages, the following steps are proceeded:

1) Feature extraction - typically mel-frequency cepstral coefficients (MFCC) [6], realized in following steps for every input signal frame:
   - Multiply by window function (typically Hamming window)
   - Computing fast fourier transform (FFT) of the input signal
   - Mel-scaling of FFT coefficients with the use of mel-bank filters
   - Computing logarithm of previously calculated coefficients
   - Computing discrete cosine transform from values obtained in the previous step
2) Modeling - for MFCC modelling typically Gaussian Mixture Models (GMM) [7] is used, which is sum of weighted gaussian distributions describing set of cepstral coefficients

In the case of testing stage the distance between computed models or probability is being calculated with the use of following algorithms:

- Euclidian distance
- Mahalanobis distance [8]
- Kulback-Leibler Divergence [9]
- Logarithm probability

Described steps are presented in fig. 1, which is commonly used model-based schema for speaker recognition systems [4].

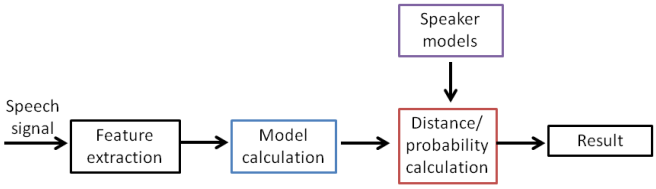

Fig. 1. Automatic speaker recognition system schema

Presented project uses VOICEBOX [10] speech processing toolbox for Matlab environment, which includes most of commonly used functions in speech processing related to analysis, synthesis, processing, modeling and coding:

1) MFCC calculation – „Melcepst" function with inputs of signal frame and sampling rate: 12 coefficients are calculated for every signal frame
2) GMM modeling – „Gaussmix" function with inputs of MFCC and number of gaussians
3) GMM modeling and log probability calculation – „Gaussmixp" function with inputs of MFCC and previously computed speaker model to compare, which includes set of means, variances and weights.

Figure 2 presents example distribution of 1st MFCC coefficient and computed mixture of 16 gaussians which fit the distribution of input data.

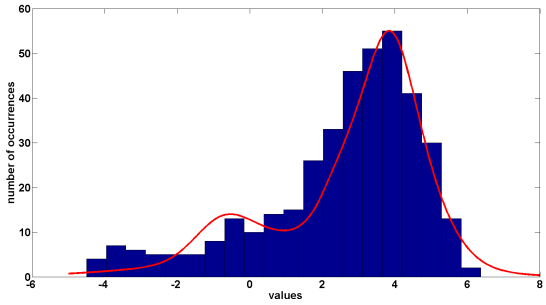

Fig. 2. Adaptation of gaussian weighted mixtures to input data distributions

The acquisition of MPEG layer 3 (MP3) [11] compressed audio stream is done by Matlab DSP class. One of its objects is AudioFileReader, which methods read audio samples from declared input file or stream. This object can call method „step" which constantly acquire declared number of samples being processed in the background. It is also possible to interrupt by timer overflow, but authors have chosen first method because of their legibility.

III. System description

The idea of the system is to acquire and process sound signal in real time from Internet stream. Presented system was developed under Matlab environment. It uses new methods to process the sound - the DSP class and it's methods and properties related to specified objects, which was described in previous section. The software contains two main threads:

1) Play audio stream
2) Process buffered stream

Program flow is presented in fig. 3. The input signal is sampled 44100 times per second and acquired into buffer of length being equivalent to 1 second of recording. It is passed through to constant play and at the same time copied to provide speaker recognition issue. To recognize the speaker „on the fly", input stream is downsampled 4 times to 11025 Hz, which is over minimum required value (8000 samples per second) for speech processing. The sampling rate has been chosen because of downsample factor equal to 4. In this case there is no need to upsample and downsample by high factors, which have be used in the case of 8000 samples per second. In next step the signal is cut into frames and the mel-frequency cepstral coefficients are calculated for each of them. In the next step they are modeled with the use of Gaussian Mixture Model algorithm and compared to the set of previously defined speakers. The best match is presented in GUI updated in real time (fig. 4).

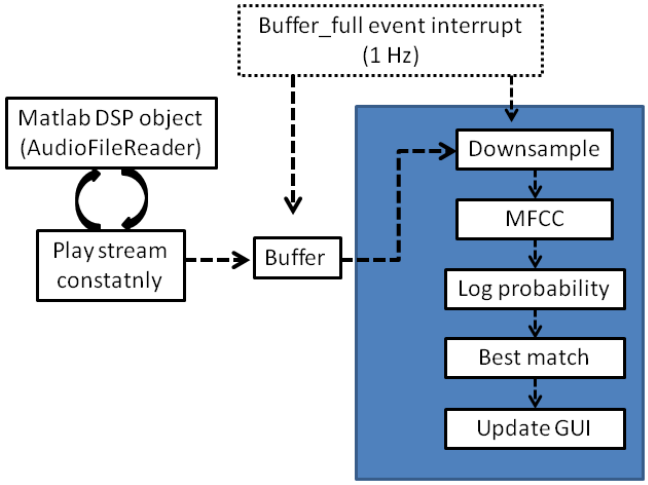

Fig. 3. Online speaker recognition software schema

The graphical user interface (GUI) of the described system is presented in fig. 4.

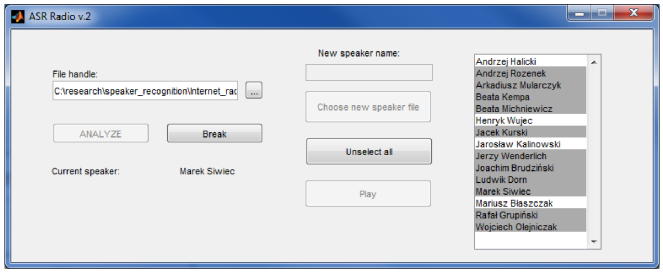

Fig. 4. User Interface of the system

The software allows to choose speakers models for comparison, while the input stream is processed. Input stream can be also replaced with previously recorded stream. This option have been used to analyse the recordings described in section IV.

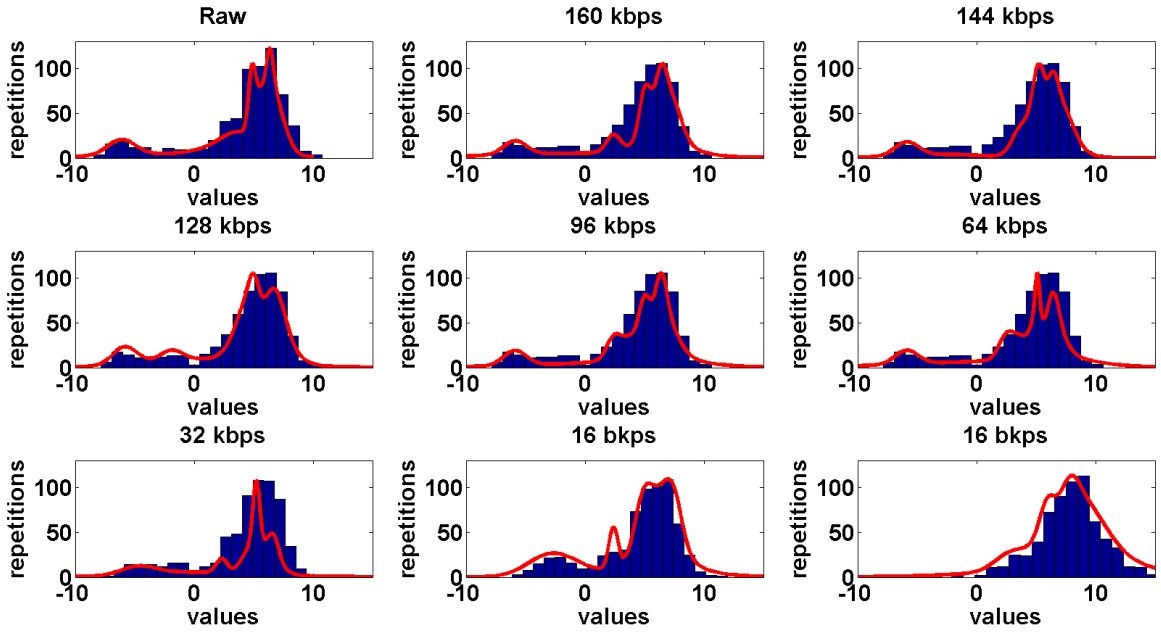

Fig. 5. Influence of MP3 compression (8 bitrates) on distribution of first MFCC

IV. Experimental results

It has to be noticed that the presented system uses compressed speech signal. Unlike in studies described in [12], [13], the algorithm used is MPEG layer 3, which is commonly used method in music and movie soundtrack compression. According to [14], signal transcoding decreases speaker recognition accuracy. To visualize how MP3 algorithm influences the signal, an illustrative distribution of first MFCC (calculated from 5-second length recording) have been transcoded with the use of 8 various bitrates from 16 to 160 kbps. It is presented in fig. 5. It can be observed, that while the bitstream lowers, the distribution of MFCC is more smoothed and some important features can be removed. In the case of the prepared recordings, the bitstream used is 128 kbps. In relation to fig. 5, the dominant value of MFCC has been shifted with reduction of theirs value of occurrences.

Presented algorithm of speaker identification have been described in [4], [15] in the case of efficiency by FAR/FRR plots (false acceptance rate/false rejection rate). Studies prepared in this article focus on application of the algorithm to analyse the public debates on the Internet radio. Four about 1-hour length recordings of „Breakfast with the third program of Polish Radio" („Śniadanie z trójka") have been prepared. The acquired utterances are in Polish language. In the discussion participated 8 or 9 speakers, including 7 to 8 politicians and one presenter (Ms Beata Michniewicz). Every speaker had finite time to present his/her own position in discussed topic. Our goal is to check the time used of every speaker. In the case of manually obtained juxtaposition, the time needed is much more greater than the length of analysed recording. In the case of real-time calculations, the time is almost equal to length of the program. To prepare a database of speakers for the training stage, 5-seconds length recording of each speaker have been used to extract cepstral coefficients. In testing step, which means analysis of the program, 1-second frames have been used. The usage of such short input signals is determined by the character of the recorded conversations that included lot of short pauses between words. These moments of silence and background noise can cause incorrect speaker identification.

Figures 6 to 9 presents analysis of recordings, while table I includes a detailed statistical statement.

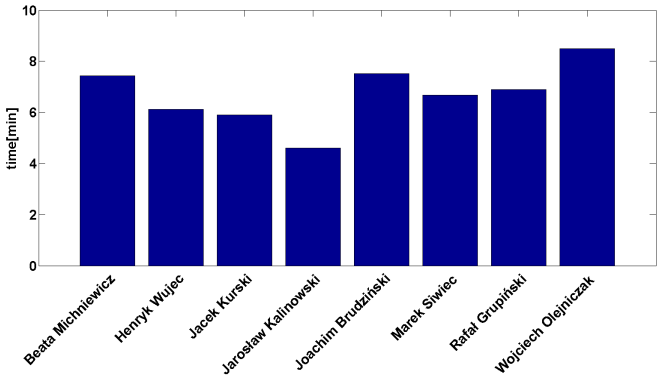

Fig. 6. Speakers activity in program no 1

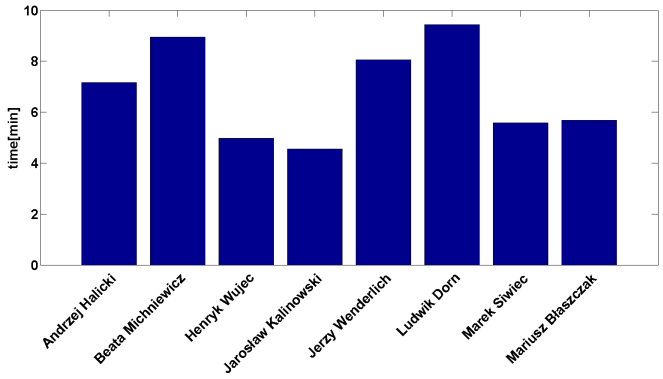

Fig. 7. Speakers activity in program no 2

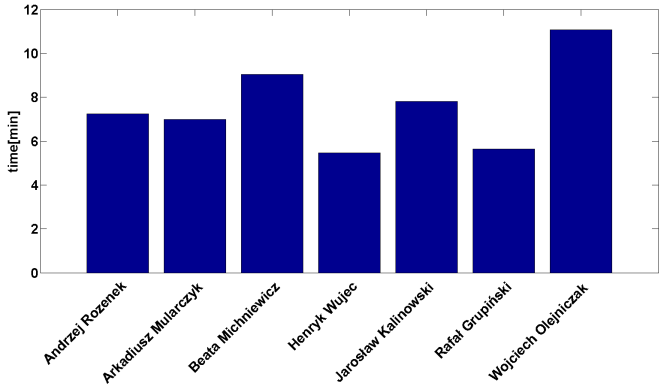

Fig. 8. Speakers activity in program no 3

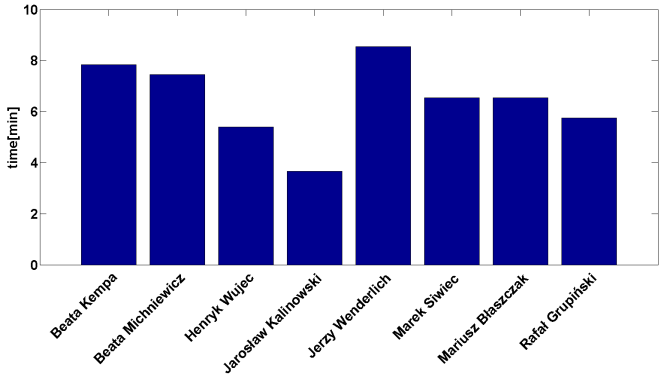

Fig. 9. Speakers activity in program no 4

In order to illustrate the distances between speaker models and their neighbors the ISOMAP [16], [17] algorithm have been used. The algorithm is typically used to reduce dimensionality in the geodesic space of the nonlinear data manifold. To provide the neighborhood graph, models of MFCC have been computed for all of 15 speakers. Then the logarithm probability between each of models have been calculated to obtain 2-dimensional probability matrix $[15 \times 15]$. In next step the matrix (of the same dimensions) of the Euclidian distances have been computed, which matrix is required as input of ISOMAP algorithm. The calculation flow is presented in fig. 10.

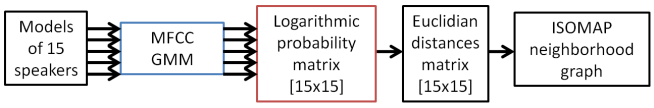

Fig. 10. ISOMAP neighbour graph preparation schema

Figure 11 presents obtained neighborhood graph of speakers participating in the debates.

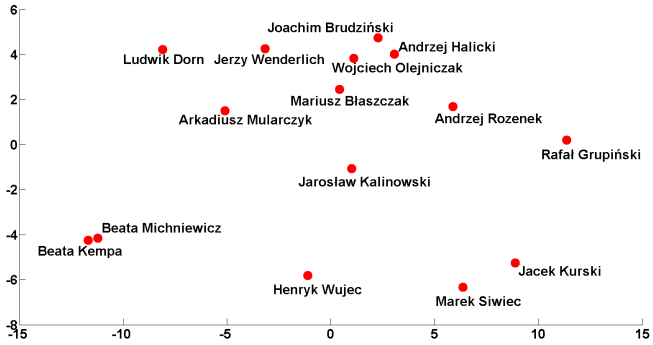

Fig. 11. 2-dimensional ISOMAP space of neighbourhoods

TABLE I. ANALYSIS OF SPEAKERS ACTIVITY

Program	Mean [min sec]	Standard deviation [min sec]	max – min [min sec]
1	6:42	1:11	3:54
2	6:47	1:52	4:53
3	7:36	1:57	5:36
4	6:52	1:49	5:10

V. FUTURE WORK

According to our previous researches [15], [18], the presented system can be improved with the use of voice activity detection (VAD) algorithms, which maximize information content in signal and improve overall accuracy. Because of signal compression, which lowers the effectiveness, another speaker models in database can be used. Additionally, the system has to be tested against speakers that do not exist in database. To avoid invalid speaker detection, the thresholding operation can be added. The ISOMAP algorithm proposed to visualize neighborhood graph can be also used to find closest match between obtained models.

VI. CONCLUSION

In this paper we presented automatic speaker recognition from encoded Internet radio signal. Described system was developed with the use of Matlab environment. Actually Matlab community dose not include solution regarding such software. We plan to share our application in Matlab Cetral File Exchange to make it publicly available. Results obtained from the analysis of 4 hours of recordings show, that speaker recognition issue can be done in real time and can significantly improve analysis of speaking time. In the case of influence of MPEG layer 3 compression algorithm to speech signal, even the best chosen quality flatten the distribution of MFCC. As it has been proved in [13], proper model selection can improve the system effectiveness.

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*, Nov 2010, pp. 1697–1701.

[2] 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*, Sept 2013, pp. 1–5.

[3] K. Park, J.-S. Park, and Y.-H. Oh, "GMM adaptation based online speaker segmentation for spoken document retrieval," *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 2, pp. 1123–1129, 2010.

[4] 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*, pp. 95–98, 2010.

[5] Z. Piotrowski, J. Wojtun, and K. Kaminski, "Subscriber authentication using GMM and tms320c6713dsp," *Przeglad Elektrotechniczny*, no. 12a/2012, pp. 127–130, 2012.

[6] 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*, vol. 1, 2001, pp. 73–76.

[7] D. Reynolds, "Gaussian mixture models," *Encyclopedia of Biometrics*, pp. 659–663, 2009.

[8] R. D. Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1 – 18, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743999000477

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

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

[11] 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*, vol. 45, no. 10, pp. 789–814, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10271

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

[13] R. Weychan, A. Stankiewicz, T. Marciniak, and A. Dabrowski, "Improving of speaker identification from mobile telephone calls," in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science, 2014, vol. 429, pp. 254–264.

[14] A. Dabrowski, S. Drgas, and T. Marciniak, "Detection of GSM speech coding for telephone call classification and automatic speaker recognition," *ICSES*, pp. 415–418, 2008.

[15] 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*, pp. 159–163, 2011.

[16] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[17] G. Wen, L. Jiang, and J. Wen, "Using locally estimated geodesic distance to optimize neighborhood graph for isometric data embedding," *Pattern Recognition*, vol. 41, no. 7, pp. 2226 – 2236, 2008.

[18] T. Marciniak, R. Weychan, and A. Krzykowska, "Speaker recognition based on telephone quality short Polish sequences with removed silence," *Przeglad Elektrotechniczny*, pp. 42–46, 2012.

We would like to thanks our students Albert Malina and Pawel Dymarkowski for preparation the basic software and Graphical User Interface.

This work was partly supported by the project „Scholarship support for PH.D. students specializing in majors strategic for Wielkopolska's development", Sub-measure 8.2.2 Human Capital Operational Programme, co-financed by European Union under the European Social Fund.