SPEAKER RECOGNITION BASED ON SHORT POLISH SEQUENCES

Tomasz Marciniak, Radosław Weychan, Agnieszka Krzykowska, Szymon Drgas, Adam Dąbrowski

Poznań University of Technology, Chair of Control and System Engineering,
Division of Signal Processing and Electronic Systems,
ul. Piotrowo 3a, 60-965 Poznań, Poland,
e-mail:Tomasz.Marciniak@put.poznan.pl,

Abstract: *This paper presents results of speaker recognition research carried out using a short polish sentences. We analyze proper selection of vector quantization representation in order to maximize of identification effectiveness. We compare vector quantization algorithm with GMM (Gaussian mixture model) technique. During our research we use a special prepared database which consist of short speech sequences.*

1 Introduction

Proper selection of acquisition parameters and the representation of the input signal is a key element in the effectiveness and speed of biometric identification systems. These systems are mainly based on image analysis such as fingerprint, face, iris, ear, hand geometry [1, 2].

Techniques of identification based on the acoustic signal (voice) are less popular and they hold about 3% share in commercial biometrics market [3]. It should be noted, however, that the speaker identification has a number of advantages and can be used to authorization access during access to multiple services and systems such as voice dialing options, telephone banking, shopping by phone, database access, voicemail, information services, access to restricted zones access to computers, etc. In contrast to systems based on image recognition, the speaker recognition allows for the detection of sex or nationality. It may also be part of a multimodal biometric system, examining the many features of the biometric, thus allowing to obtain more effective identification.

Speaker recognition techniques based on the individual characteristics of the speech signal are divided into two types: verification and identification [4, 5].

The first type of speaker recognition namely verification consists in an acceptance or rejecting of the speaker. Speech input after parametrization process is compared with the reference model. Depending on certain threshold diagnosis, the speaker is accepted or rejected. The verification process is a simpler task than identification.

The identification process recognizes which person speaks from the set of the registered people. The parameters of the input signal (speech signal) are compared with the base parameters of the reference N-models. Then a maximum selector shows the greatest similarity to the reference model and gives the appropriate speaker ID.

Speaker recognition methods can be divided into two main categories [6]:

- text-dependent - speaker recognition is performed on the basis of notice specified word / phrase, such as passwords, PIN numbers
- text-independent – recognition process is performed based on the characteristics of speech regardless of what is spoken
- text-prompted - method in which password sentences are completely changed every time.

Text-dependent methods based on DTW (dynamic time warping) or HMM (hidden Markov model).

In the case of text-independent methods two methods are the most popular:

- vector quantization VQ technique – a small number of representative feature vectors in codebooks represents speaker-specific features
- GMM (Gaussian mixture model) representation – popular method based on maximum likelihood estimation [7].

Both above methods use as an input the mel frequency cepstral coefficients MFCC [8] . Typically it is assumed that the length of the input signal for recognition process is should not exceed 10 seconds.

Problems with the implementation of speaker recognition systems are associated with the variability of the input speech signal. The change in the voice (thereby changing certain characteristics of the speech signal) affect different factors, including: time (change of voice at the time), illness e.g. a cold, speed of talking, and the external factors: environmental conditions and background noise [9].

2 Elements of software implementation

2.1. Feature extraction

Figure 1 shows simplified block diagram of speaker recognition system. Speaker reference models are calculated during the training phase, however an identification process realizes the test/operational phase.

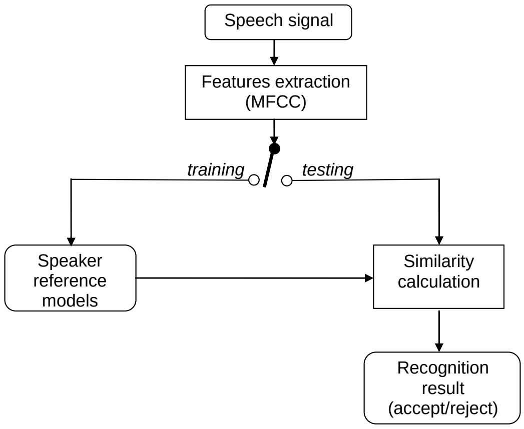

Fig. 1. Simplified block diagram of speaker recognition

Extraction of speech features from a particular speech sequence consists of the standard steps: division of sampled signal into blocks of length equivalent to 20-30 ms, multiplication of the blocks by a window function (typically Hamming window), calculation of DFT transform (typically using FFT), mel-scaling, and finally calculation of MFCC.

2.2. Parametrization using Vector Quantization

Vector quantization is a method in which the analyzed data are modeled using a small set of vectors - a cluster centroids of the feature space. In contrast to the GMM method [7], vector quantization is not the probability distribution models of the input data and offers high efficiency in both speaker recognition categories (dependent and independent) in the case of the relatively shortly expressions. Another advantage of VQ methods are very small memory requirements [8].

Implementation of VQ in the Matlab environment uses following functions [10, 11]: *mfcc* – calculation of the MFCC vectors, *vqlbg* - containing LBG (Linde, Buzo, Gray) algorithm [12] of vector quantization to create a code book and *disteu* - to calculating the Euclidean distance. Genuine software components are supplemented by the possibility of batch processing, generation of statistics and the accurate EPD (end points detection) of words [13].

2.3. Recognition using GMM

As mentioned earlier the speaker modeling by Gaussian mixture models is currently one of the most frequently used techniques in the text independent automatic speaker recognition systems. As in the case of vector quantization speaker voice is modeled with mel-frequency cepstral coefficients (MFCCs), then the determined feature vectors are used for GMM model training. This process was performed by expectation-maximization (EM) algorithm [14]. The speaker classification is provided with calculation of a conditional likelihood.

An implementation of GMM based speaker recognition can be found at [15]. Feature extraction from wave file is realized by *melcepst* m-function, which calculates mel cepstrum with 12 coefficients from 256 sample frames [16]. Statistical modeling performs function *gmm_estimate*, which determines means, covariances and weights of the models created during training phase, while *lmultigauss* computes multigaussian log-likelihood during test phase. As in the case of VQ, we have supplemented GMM software with possibility of batch processing and generation of statistics.

Original GMM based project assumes that the feature extraction and modeling during training should be taken from 30-60 second speech recordings. For testing phase 10 second of speech is needed [16]. It should be noted that our research was carried out with several times shorter recordings.

3 Experimental results

3.1. Database of short sentences

In order to test the accuracy of the speaker identification, we have recorded a group of 25 speakers of both sexes, aged from 22 to 55 years who spoke at three sessions the following short phrases:

- „Dzień dobry” („Good morning!”)
- „Dobry wieczór” („Good evening!”)
- „Do widzenia” („Goodbye”)
- „Moje nazwisko” („My name”)
- „Chciałbym/chciałabym zgłosić wypadek” („I would like to report an accident”).

These phrases were selected based on statistical analysis of expressions spoken by phone during emergency calls. Additionally, was also spoken by the speakers one longer sentence „Czas nadziei nie trwa wiecznie” („Time of hope doesn’t last forever”). It may be noted that all above statements don’t have the characteristics of isolated words (typically names or numbers) and each time they are pronounced differently.

Speech recording takes place at the sampling rate 22050 samples/second and 16-bits resolution. Individual speaker spokes each phrase 30 times (10 times during each session). This database consists of 4500 wave files, which were recorded in the range of 5 months. Time interval between successive sessions ranged from 1 to 4 weeks.

3.2. Results of recognition

During our research we compare speaker recognition accuracy using VQ and GMM methods. There has also been limited attempts made using the DTW algorithm. The disadvantage of this algorithm is relatively high time-consuming. In comparison with the GMM, the DTW algorithm is about 30 times longer.

Vector quantization algorithm works with following parameters: 32 centroids, 30 filters in filter bank, 0 Hz low end of the lowest filter, 4000 Hz high end of highest filter. Training is realized with 1 file of each sentence, and therest of files (29) for test phase.

While studies using GMM set the following parameters: 12 cepstral coefficients excluding 0'th coefficient, 30 filters in filter bank, 0 Hz low end of the lowest filter, 4000 Hz high end of highest filter. For training phase we have used 5 files of each sentence of each speaker and for test – 25 files.

Figure 2 presents most important results, namely an overall recognition accuracy depending on number of centroids for VQ and number of Gaussians for GMM. It can be observed that text-dependent recognition accuracy in both cases is greater than 81,4%. Vector quantization algorithm tests were done for 5 different numbers of centroids – 16, 24, 32, 48 and 64. As we can see in some cases, numbers of centroids has not influence of recognition accuracy, but overall it has growing trend. Too many numbers of centroids increase time of computing and can lead to algorithm overlearned, so number of centroids has been set to 32. In case of GMM, tests were done for 4 different numbers of Gaussians – 5, 10, 20 and 30. Number of Gaussians has been set to 10.

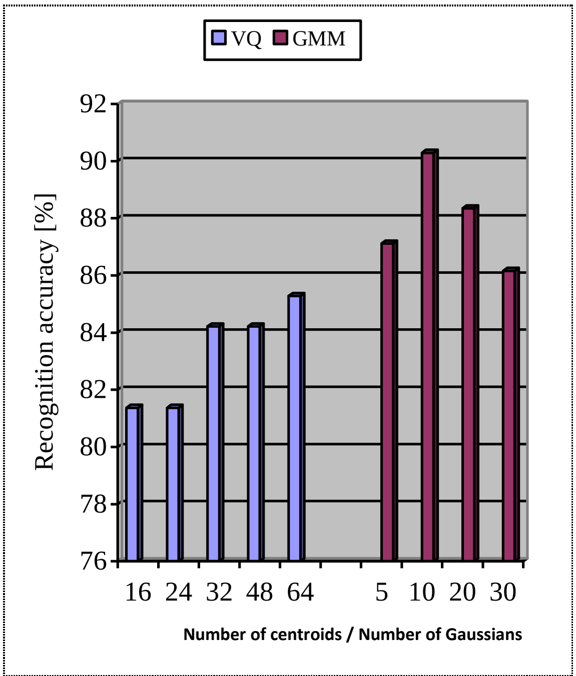

Fig. 2. Speaker recognition accuracy

Figure 3 shows speaker dependent results for particular sentences. We see that GMM computing guarantee in most cases better speaker recognition. The difference is 7-10 %, but in case "Moje nazwisko" phrase disparity is in favor of VQ (2,4 %).

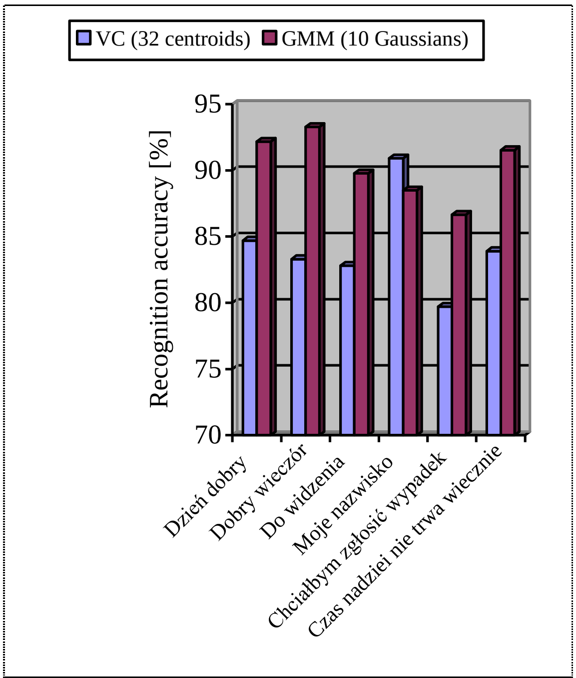

Fig. 3. Speaker recognition accuracy for text-dependent

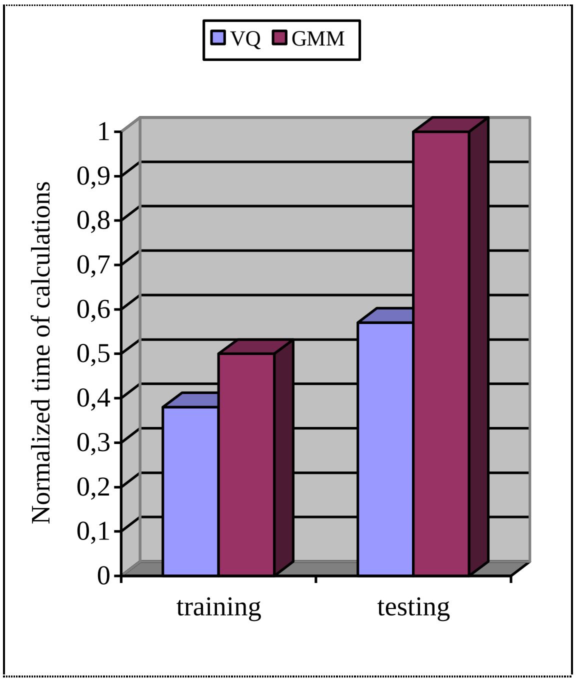

Fig. 4. Time of calculations

Figure 4 shows comparison of normalized time of training and test phase. We can observe that in the case of GMM test phase is 2 times greater than training phase and in both cases greater than VQ computing. In case of different numbers of testing files for each speaker (29 for

VQ and 25 for GMM), real testing time ratio GMM to VQ came to 2. Time of parametrization and Euclidian distance computing take about 14 ms for VQ and 28 ms for GMM algorithm (Matlab environment v.7.8.9, computer's efficiency by Matlab Bench Relative Speed=20).

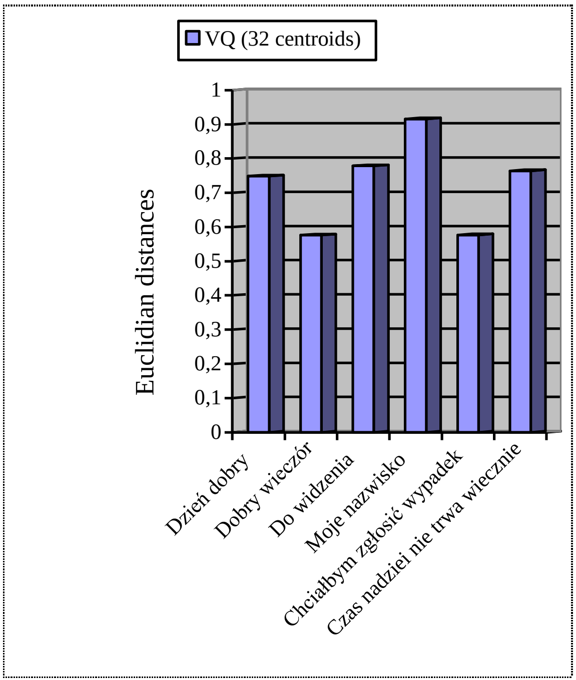

Fig. 5. Comparison of euclidian distances

Figure 5 shows Euclidian distances between the nearest model and next model. This numbers have been set for each sentence using VQ algorithm. The sooner this value is great, the better recognition accuracy we can get. Besides, by using end point detection algorithm, this values and overall recognition accuracy have been improved.

4 Conclusions

Our aim was analysis of speaker recognition methods in the case where voice signal is relatively very short. Our next step is an implementation selected method in an embedded system with DSP processor. Modern development environments such as Matlab make it easy to convert the software from the Matlab / Simulink environment to for example Code Composer Studio. [17]

The resultant efficiency of the identification of more than 90% for GMM (10 Gaussians) and 84 % for VQ (32 centroids) shows that a relatively simple method of vector quantization works well for very short duration of expression to less than 3 seconds. Of course, still needs to refine and improve efficiency. Some ideas are to better end-point-detection algorithm and to devise methods of training models choosing.

References

[1] V. Govindaraju, *Advances in Biometrics - Sensors, Algorithms and Systems*, Springer-Verlag London Limited 2008.

[2] A. Dąbrowski, T. Marciniak, Sz. Drgas, P. Pawłowski, Ekstrakcja informacji z obrazów, wideo i mowy w systemach ochrony i bezpieczeństwa, rozdział w monografii *Ergonomia - Technika i Technologia - Zarządzanie*, red. Marek Fertsch, ss.151-167, Wydawnictwo Politechniki Poznańskiej, Poznań 2009.

[3] Biometrics Market and Industry Report 2009-2014, http://www.biometricgroup.com/reports/public/market_report.php .

[4] D. O'Shaughnessy**,** *Speech Communications: Human and Machine*, Wiley-IEEE Press, 2000.

[5] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, Second Edition, Revised and Expanded*, Marcel Dekker, Inc., New York, 2001.

[6] S. Furui, Speaker recognition, Scholarpedia (2008), http://www.scholarpedia.org/wiki/index.php?title=Speaker_recognition&printable=yes

[7] D. Reynolds, Robust text-independent speaker identification using Gaussian Mixture Speaker Models, *IEEE Trans. Speech Audio Proc.*, Vol. 3, No. 1, 1995.

[8] A.V. Oppenheim, R.W. Shafer, From Frequency to Quefrency: A History of the Cepstrum, *IEEE Signal Processing Mag.*, pp. 95-99, Sep. 2000.

[9] J. Keshet, S. Bengio, *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, John Wiley & Sons, 2009.

[10] DSP Mini-Project: An Automatic Speaker Recognition System http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition .

[11] R. Chassaing, Digital Signal Processing and Applications with the TMS320C6713 and TMS320C6416 DSK Second Edition, John Wiley & Sons, Inc., 2008.

[12] Y. Linde,A. Buzo, R.M. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84--95, Jan. 1980.

[13] Marciniak, T., Dąbrowski, A., *Influence of subband signal denoising for voice activity detection*, Elektronika – konstrukcje, technologie, zastosowania, nr 3/2009, ss. 67-70.

[14] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.

[15] A. Alexander, A. Drygajlo Speaker identification: A demonstration using Matlab, http://scgwww.epfl.ch/matlab/student_labs/2005/labs/ .

[16] VOICEBOX: Speech Processing Toolbox for Matlab http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html .

[17] R. Weychan, T. Marciniak, A. Dąbrowski, "Akwizycja i parametryzacja sygnału mowy w czasie rzeczywistym z zastosowaniem pakietu Target Support Package TC6", VII Sympozjum MiS, materiały konferencyjne, ss. 185-188.

This work was partly supported by INDECT, PPBW and DS. projects.