Fast speaker recognition based on short Polish sequences

Szybkie rozpoznawanie mówcy na podstawie krótkich wypowiedzi w języku polskim

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

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 experiments using short Polish sentences. We developed and analyzed various parameters in speech signal modeling in order to first maximize identification effectiveness and second to compare VQ (vector quantization) and GMM (Gaussian mixture model) approaches. For the research and experiments we created and exploited a database, containing specially prepared short Polish speech sequences typical for emergency phone calls.*

Streszczenie: *Artykuł prezentuje wyniki badań nad rozpoznawaniem mówcy na podstawie krótkich wypowiedzi w języku polskim. Przeanalizowano dobór parametrów modelowania sygnału mowy w celu maksymalizacji skuteczności identyfikacji oraz porównania rozwiązań wykorzystujących kwantyzację wektorową VQ oraz sumę rozkładów normalnych CMM. Do badań eksperymentalnych utworzono i wykorzystano przygotowaną przez autorów bazę nagrań zawierającą specjalnie dobrane krótkie wypowiedzi w języku polskim, typowe dla rozmów telefonicznych na numery alarmowe.*

1 Introduction

Among key issues determining effectiveness and speed of biometric identification systems there is a proper selection of acquisition parameters and representation of the input data. Typically biometric systems are based on the analysis and recognition of various images such as fingerprint, face, iris, ear, hand geometry, etc. [1, 2]. However, nowadays speaker recognition techniques are becoming more and more important.

Identification techniques based on acoustic signals (voice) is yet holding about a 3% share in the commercial biometric market only [3]. However, it should be noted that the speaker identification has a number of advantages and can be used to authorization access for many 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, speaker recognition easier detects sex and nationality. It may also be a part of complex multimodal biometric systems examining many biometric features thus guarantying more effective identification.

Speaker recognition techniques can be divided into two types: verification and identification [4, 5].

Speaker verification consists in acceptance or rejection of the speaker. Speech after parameterization is compared with the reference model. Depending on certain diagnosis threshold, the speaker is accepted or rejected.

Verification is simpler than the second task namely identification, which consists in recognition which person from a set of the registered people speaks. Parameters of the input speech signal are compared with the base parameters of the reference N-models. Then the maximum selector shows the greatest similarity to the reference model and selects the appropriate speaker ID.

Speaker recognition methods can be divided into the following 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 has been spoken
- text-prompted – a method in which password sentences are changed every time.

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

For text-independent recognition two approaches are particularly important:

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

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

Most of problems with the implementation of speaker recognition systems are associated with the variability of the input speech signal. Changes in the voice (thereby changes of certain characteristics of the speech signal) affect various factors including: time (change of voice at the time), illness (e.g., a cold), speed of talking, and external factors such as: environmental conditions as well as 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.

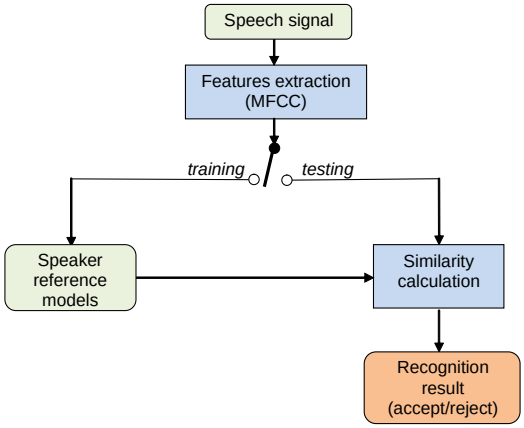

Fig. 1. Simplified block diagram of speaker recognition

Rys. 1. Uproszczony schemat blokowy rozpoznawania mówcy

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

2.2. Parameterization using Vector Quantization

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

Implementation of VQ in the Matlab environment uses the 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* – for 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 already mentioned, speaker modeling by the 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 (MFCC's). Then the determined feature vectors are used for the GMM model training. This process is performed by expectation-maximization (EM) algorithm [14]. The speaker classification is provided with calculation of the conditional likelihood.

Our realization of the GMM based speaker recognition is based on the approach described in [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 uses function *gmm_estimate,* which determines means, covariances, and weights of the models created during the training phase, while *lmultigauss* computes multigaussian log-likelihood during the test phase. We have supplemented the GMM and VQ software with a possibility of batch processing and generation of statistics.

We assumed that the feature extraction and modeling during training should be taken from 30–60 second speech recordings. Typically for the testing phase 10 seconds of speech is needed [16]. It should be noted that in our research several times shorter recordings were used.

3 Experimental results

3.1. Database of short sentences

In order to test 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").

Additionally, one longer sentence was also spoken by the speakers, namely: „Czas nadziei nie trwa wiecznie"

(„Time of hope does not last forever"). These phrases were selected based on the statistical analysis of expressions spoken by phone during the emergency calls. It may be noted that all above statements do not have the characteristics of isolated words (typically names or numbers) and each time they may be pronounced differently.

Speech recordings were made with the sampling rate 22 050 samples/second and 16-bit resolution. Individual speaker spoke each phrase 30 times (10 times during each session). The 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 compared speaker recognition accuracy using VQ and GMM methods. There has also been experiments performed based on the DTW algorithm. A disadvantage of this algorithm is a relatively long analysis time as the DTW algorithm is about 30 times longer in comparison with the GMM approach.

Vector quantization algorithm worked with the following parameters: 32 centroids, 30 filters in the filter bank, 0 Hz lower edge of the lowest filter, 4000 Hz higher edge of the highest filter. Training was realized with 1 file of each sentence, and the rest (29 files) for the test phase.

GMM experiments were carried out with the following set of parameters: 12 cepstral coefficients excluding 0th coefficient, 30 filters in the filter bank, 0 Hz low end of the lowest filter, 4000 Hz high end of the highest filter. For the training phase we used 5 files of each sentence for each speaker and for the test the rest, i.e., 25 files.

Figure 2 presents the most important results, namely the overall recognition accuracy depending on the number of centroids for VQ and the number of Gaussians for GMM. It can be observed that the text-dependent recognition accuracy is in both cases greater than 81,4%. Vector quantization algorithm tests were done for 5 different numbers of centroids, namely for 16, 24, 32, 48, and 64. As we could see in most cases, the growing number of centroids has not significantly influenced the recognition accuracy but in general it had some growing trend. Too many centroids increased the computation time and could lead to overlearning. Thus the optimum number of centroids has been found to be 32. In case of GMM, tests were done for 4 different numbers of Gaussians, namely 5, 10, 20, and 30. The best number of Gaussians has been found to be 10 (Fig. 2).

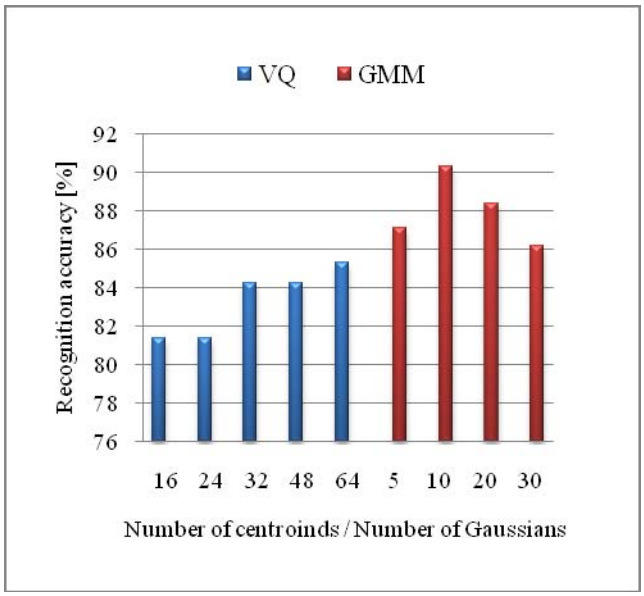

Fig. 2. Speaker recognition accuracy

Rys. 2. Skuteczność rozpoznawania mówcy

Figure 3 shows the text-dependent speaker recognition results. We can see that in most cases GMM guarantees better speaker recognition by 7–10 % than VQ. However in case "Moje nazwisko" the recognition is in favor of VQ by 2,4 %.

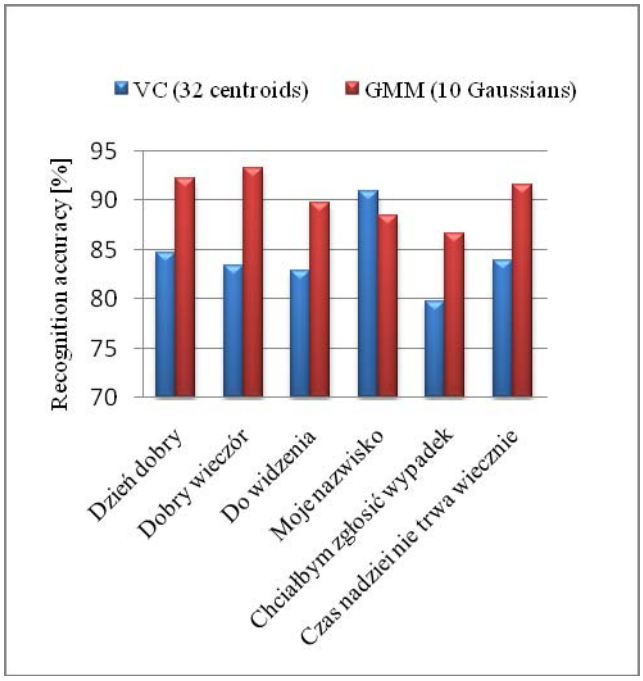

Fig. 3. Text-dependent speaker recognition accuracy

Rys. 3. Skuteczność rozpoznawania mówcy w zależności od wypowiedzi

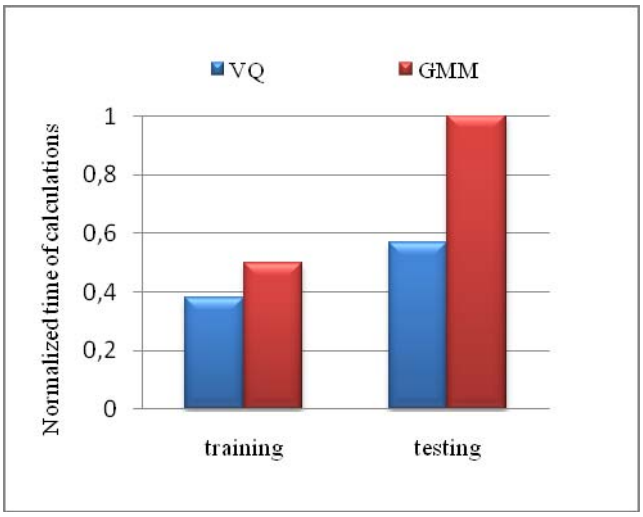

Fig. 4. Normalized time of calculations

Rys. 4. Znormalizowany czas obliczeń

Figure 4 shows comparison of the normalized times of training and test phases. We can observe that in the case of GMM the test phase is 2 times longer than the training phase and in both cases greater than for 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. The time of parameterization and Euclidian distance computing took about 14 ms for VQ and 28 ms for GMM algorithm (Matlab environment v.7.8.9, computer efficiency by Matlab Bench Relative Speed=20).

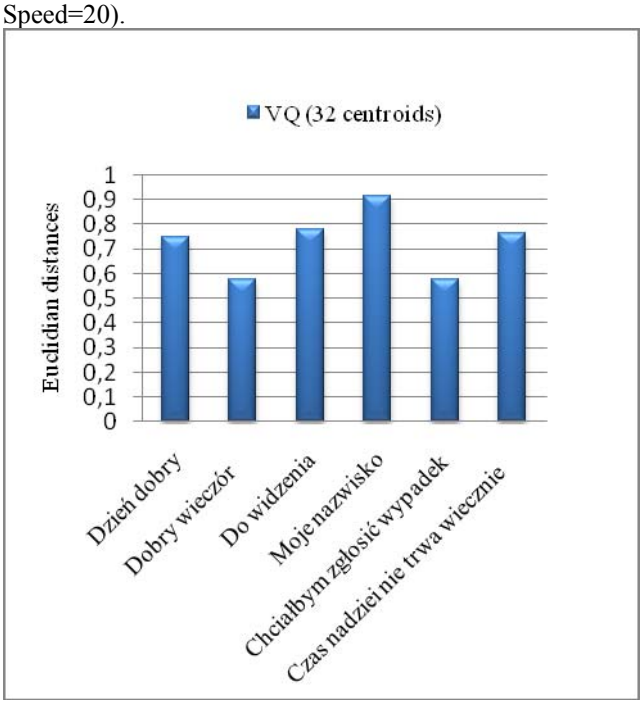

Fig. 5. Comparison of Euclidian distances

Rys. 5. Porównanie odległości euklidesowych

Figure 5 shows Euclidian distances between the nearest model and the next model. These numbers have been set for each sentence using the VQ algorithm. The sooner these values are great the better is the recognition accuracy. Besides, by using an additional end point detection algorithm, these values and the overall recognition accuracy have been improved.

4 Conclusions

Our aims were: first, the analysis of speaker recognition performance for relatively short voice signals and second, implementation of the selected methods in the embedded system with the DSP processor. Modern development environments such as Matlab make it easy to convert the general software, e.g., from the Matlab / Simulink environment to, e.g., the Code Composer Studio for Texas Instruments processors [17].

The resulting efficiency of the identification is more than 90 % for GMM (10 Gaussians) and 84 % for VQ (32 centroids). It shows that a relatively simple method of vector quantization works well for very short duration of expressions (less than 3 seconds).

Some further improvements of efficiency are still possible. The respective ideas are: better end-point-detection and better training models.

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.