Tomasz MARCINIAK, Agnieszka KRZYKOWSKA, Radosław WEYCHAN

Poznań University of Technology

Speaker recognition based on telephone quality short Polish sequences with removed silence

Abstract: *This paper presents the effectiveness of speaker identification based on short Polish sequences. An impact of automatic removal of silence on the speaker recognition accuracy is considered. Several methods to detect the beginnings and ends of the voice signal have been used. Experimental research was carried out in Matlab environment with the use of a specially prepared database of short speech sequences in Polish. The construction of speaker models was realized with two techniques: Vector Quantization (VQ) and Gaussian Mixture Models (GMM). We also tested the influence of the sampling rate reduction on the speaker recognition performance.*

Streszczenie: *Artykuł przedstawia badania efektywności rozpoznawania mówcy opartego na krótkich wypowiedziach w języku polskim. Sprawdzono wpływ automatycznego wykrywania i usuwania ciszy na jakość rozpoznawania mówcy. Przebadano kilka różnych metod wykrywania początku i końca fragmentów mowy w wypowiadanych sekwencjach. Eksperymenty zostały przeprowadzone z użyciem środowiska Matlab i specjalnie utworzonej bazy krótkich wypowiedzi w języku polskim. Do budowy modeli mówców wykorzystano kwantyzacja wektorowa (VQ) oraz Gaussian Mixture Models (GMM). Podczas badań sprawdzono także wpływ obniżenia szybkości próbkowania na skuteczność identyfikacji mówcy.*

Keywords: speaker recognition, vector quantization, GMM, end-point detection.
Słowa kluczowe: rozpoznawanie mówcy, kwantyzacja wektorowa, GMM, detekcja głosu.

Introduction

Techniques based on acoustic signals are an interesting solution in numerous biometry applications [1,2]. In our present study we focus on experiments with short speech sentences [3, 4]. This paper examines an influence of the voice activity detection techniques on efficiency of the speaker identification. This task has been realized with the use of the GMM (Gaussian mixture models) as well as VQ (vector quantization) algorithms. Tests were performed for two sample rates: original – 22050 samples/second and its downsampled versions (8000 samples/second).

The paper is organized as follows: Section "EPD methods" presents algorithms of precise end-point detection (EPD), which gives the possibility of removing silence parts in speech signal. Section "Speaker recognition algorithm" briefly describes the speaker identification techniques. Section "Experimental results" includes the results in a form of the FAR / FRR (false accept rate / false reject rate) graphs, while section "Conclusions" summarizes our work.

EPD methods

Precise detection of voice activity endpoints is crucial in many speech processing procedures like: speech coding in telephone communication, speech enhancement, automatic speech or speaker recognition.

In the case of automatic speech / speaker recognition, the precise detection of word boundaries is an important step, which significantly improves the recognition effectiveness. Voice activity detection can be realized in time and/or in frequency domain with the use of, e.g., the TF (time-frequency) speech parameters [5, 6, 7, 8].

In our experiments we tested four algorithms: energy analysis algorithm and three EPD algorithms prepared by Roger Jang [9]. Table 1 summarizes the applied methods described shortly in the next sections.

Detection and removal of silence was made for the two solutions. In the first part of the experiment only silence at the beginning and end of the speech sequence was removed. In the second part of our experiments individual words were detected in the analyzed sentence.

Sample effects of particular methods are presented in Fig. 1 that uses an illustrative wave sequence „Chciałbym zgłosić wypadek" ("I would like to report an accident"). This wave file was recorded in an anechoic chamber with the sampling rate of 22 050 samples/second.

Table 1. Applied EPD algorithms

Method	Short Description
Energy analysis	Calculation of energy value
Jang (v2) [9]	Application of volume threshold
Jang HOD [9]	Use of volume and high-order differences
Jang ZCR [9]	Silence detection based on volume and zero-crossing rate

Energy analysis

The simplest method of EPD is the analysis of signal energy:

$$E_i = \sum_{n=k_{ip}}^{k_{ik}} x^2(n) \qquad (1)$$

where i stands for the number of the frames of the signal $x(n)$, k_{ip} is the first sample of the signal, and k_{ik} is the last.

The frame length for the energy count is equal to 0,01 [s]. Number of samples in one frame is 220, and the offset of the frame is equal to 0,001 [s].

Jang algorithm (v2)

The algorithm finds the beginning and the end of the speech samples based on the volume threshold V_t defined as:

$$V_t = \frac{V_{\max} - V_{\min}}{V_r} + V_{\min} \qquad (2)$$

where coefficient V_r is 10, $V_{\min}$ – minimal value of volume vector, and $V_{\max}$ is maximal value of volume vector. Every element of $V_{\min}$ and $V_{\max}$ is computed from the equation:

$$V_i = \sum_{n=k_{ip}}^{k_{ik}} |x(n)| \qquad (3)$$

In equation (3) i means number of frames used to count the volume, and k_{ip} and k_{ik} are the first and the last samples. The length of the frame is: 0,016 [s]. Number of samples: 352. Frame offset: 0 [s].

Evident enough is the property of above algorithms where the use of a constant volume threshold gives bad results when the volume of the signal varies [3].

Jang HOD algorithm

This algorithm is based on high-order differences of a given signal as characteristics in time domain. Subsequent steps of the algorithm are as follows:

1. Determination of volume (V) as in Jang (v2) algorithm with the use of equation (3) and absolute value of the sum of j-order difference (H).

(4) $$H_i = \sum_{n=k_{ip}}^{k_{ik}} \left| \frac{\Delta^j x(n)}{\Delta n^j} \right|$$

where *i* stands for number of the time frame used to compute H_i. Parameters k_{ip} and k_{ik} are the first and the last sample of a/the? given frame. Normalization of values *V* and *H* is performed before moving to next step of the algorithm.

2. Selection of weight *w* from interval [0, 1] to compute new curve *VH*:

(5) $$VH = w \times V + (1 - w) \times H$$

3. Calculation of coefficient r to compute threshold t for VH in order to determine end-points. The threshold is defined as:

(6) $$t = VH_{\min} + (VH_{\max} - VH_{\min}) \times r$$

Default values of parameters are: j = 4, w = 0.5, and r = 0.125. Length of the time frame to compute volume is: 0,016 [s]. Number of samples in one frame is: 352. Overlapping of frames is 0 [s].

Jang ZCR algorithm

This method defines the beginning and the end of words on the basis of the volume threshold and zero crossing rate. Steps of the algorithm:

1. Selection of initial beginning and ending points based on energy value.

2. Calculation of Z (ZCR) coefficient and further expanding borders to cross threshold τ_{zc} defined as:

(7) $$\tau_{zc} = \max(Z) \times r_z$$

where r_z = 0.1, and Z (ZCR) is vector of elements, in which each one is defined as:

(8) $$Z_i = \sum_{n=k_{ip}}^{k_{ik}-1} I\{x(n)x(n-1) < 0\}$$

Length of the time frame used to compute the volume is 0,016 [s]. Number of samples of the time frame is 352, where overlapping: 0 [s].

Modification of Jang functions

Algorithms discussed above are suitable only for the detection of silence at the beginning and end of recorded speech. During the experiment we also tested the second approach. Its main goal was to analyze what impact would cutting silence have not only on the beginning and end of the sentence but also in the middle i.e. between the words (cf., Fig. 2). Since tested signals are very short, there are not many words in the sequence. Therefore not much of silence can be cut off. Nevertheless in some cases this approach showed improvement in relation to the previous one.

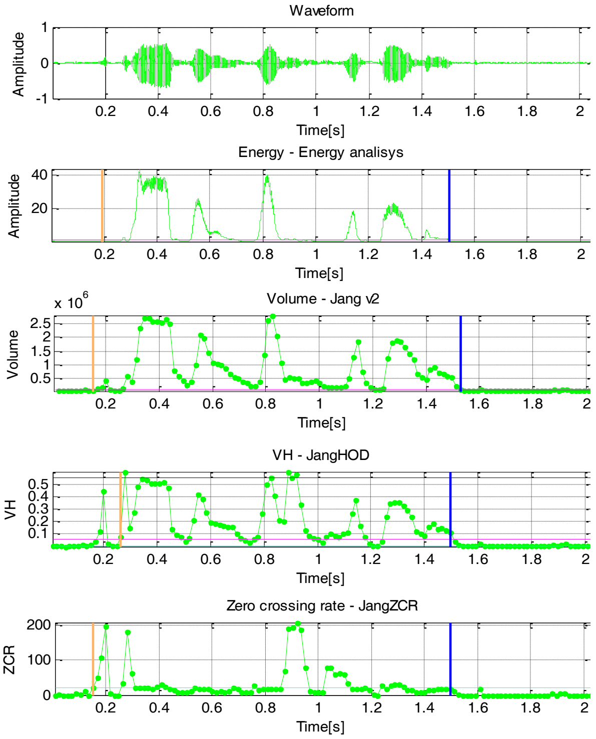

Fig. 1. Detection of silence at the beginning and at the end of sentence using Energy analysis, Jang (v2), Jang HOD and Jang ZCR algorithms. Orange vertical line represents detected beginning of the speech sentence, while the blue line shows detected end of the sequence.

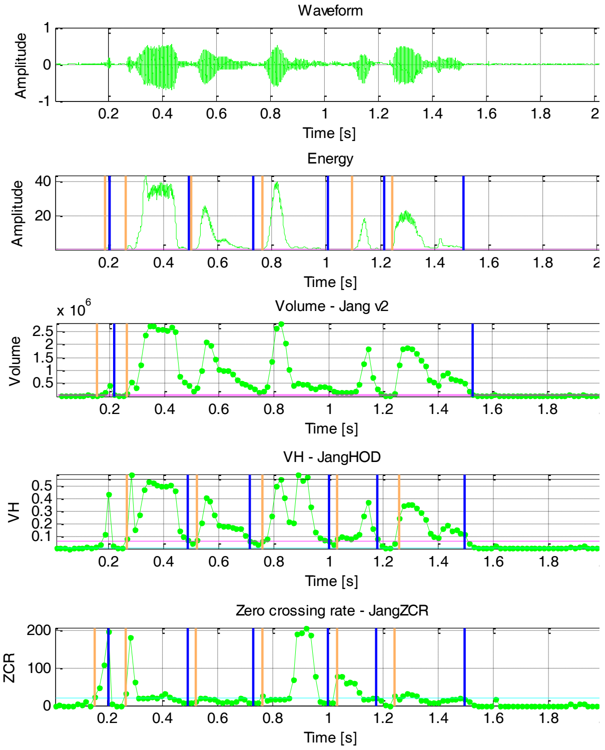

Fig. 2. Detection of silence at the beginning, in the middle and at the end of sentence using Energy analysis, Jang (v2), Jang HOD and Jang ZCR algorithms. Orange vertical line represents detected beginning of the speech, while the blue line shows detected end of the word.

Speaker recognition algorithms

Feature extraction

Speech signal was parameterized with MFCC (*mel frequency cepstral coefficients*) which uses the fact of logarithmic sound perception by human's ears. Speech signal is divided into frames being equivalent to approximately 23,2 ms in time domain. These frames are multiplied by Hamming window function. From every frame DFT coefficients are computed and translated into Mel scale. Finally, MFCC are calculated.

Vector quantization

Vector quantization (VQ) allows for the modeling of probability density functions by the distribution of prototype vectors. Large set of data are grouped into representative feature vectors. The main parameter controlling speed of algorithm is the number of centroids, which was set in our research to 32.

Quantized MFCC coefficients are generated with the use of the filter bank containing 30 filters. Low cutoff - frequency of the first filter is 0 Hz, and high cutoff for the last filter is 8000 Hz.

During our research Matlab functions from the project [10] and VOICEBOX [11] were used. Software was supplemented for a batch processing, non-deterministic model selection and statistics generation.

GMM

Gaussian Mixture Models is one of the most popular algorithm to create speaker model in speaker recognition system. First, speaker voice is modelled by MFCC. Then obtained feature vectors are used to train the model with GMM. Number of Gaussians for these experiments was set to 10. This step is realized with the Expectation-maximization (EM) algorithm. As described in our previous paper [3] our software is based on the approach presented at web page [12].

Experimental results

Experiments were performed in few stages. First, speech database was processed with described silence removal algorithms giving 18 independent bases (9 for each sampling rate). Next VQ and GMM algorithms create speaker models. The experiment resulted in multi-dimensional matrix containing coefficients of similarity between processed samples and models, for every sentence. This matrix enabled the determination of FAR / FRR curves with DETware software [13].

Since in our recordings the average time of speech is 1 second we combined features from 5 random speech sequences to obtain a sufficient speaker model.

Database of short Polish sentences

During experiments, a database containing sentences of 40 speakers of both genders and aged from 20 to 55 was prepared. Each speaker repeated 30 times 6 sentences:

- Dzień dobry (Good morning)
- Do widzenia (Good bye)
- Dobry wieczór (Good evening)
- Moje nazwisko (My name)
- Chciałbym zgłosić wypadek (I would like to report an accident)
- Czas nadziei nie trwa wiecznie (Time of hope doesn't last forever)

Recordings were realized in three stages. Every speaker repeated each sentence 10 times at once. Time period between sessions was 1 to 6 weeks.

Each of 7200 speech files was recorded in anechoic chamber with the use of an omnidirectional characteristic condenser microphone. Sampling rate of recorded samples was set to 22050 samples/second and 16 bit resolution. Next, the database was downsampled to the rate of 8000 samples/second for testing speaker recognition in a common telephone sampling conditions.

Results after detection start and end of the sentence

Vector quantization

Figure 3 presents FAR / FRR plots for unprocessed (raw) and processed with silence removal algorithms for sampling rate 8000 and 22050 samples/second. In this case, only silence at the beginning and at the end is removed. It can be observed that silence removal algorithms improve speaker recognition accuracy. Three of all presented algorithms i.e. Energy, Jang HOD and Jang ZCR give similar results. EER (equal error rate) is within the range of appr. 7-7,5 % in the case of speech sampled with 22050 samples/second. In the case of unprocessed speech, this coefficient is about 10 % higher for the same sampling rate.

An interesting result of definitely better EER can be observed for the raw speech file sampled at the rate of 8000 samples per second without removing the silence from the recording. The reason for this is a bandwidth limitation dedicated for speech (much less noise components), and constant parameters for MFCC calculation – in this case coefficients are computed for a frame of the same size, but greater equivalent in time domain. In this situation VQ algorithm works better than GMM. The same effect can also be noticed at enhanced silence removal algorithms. EER coefficient computed for fs = 8 kSps is similar to results for higher quality, making VQ dedicated for GSM standard.

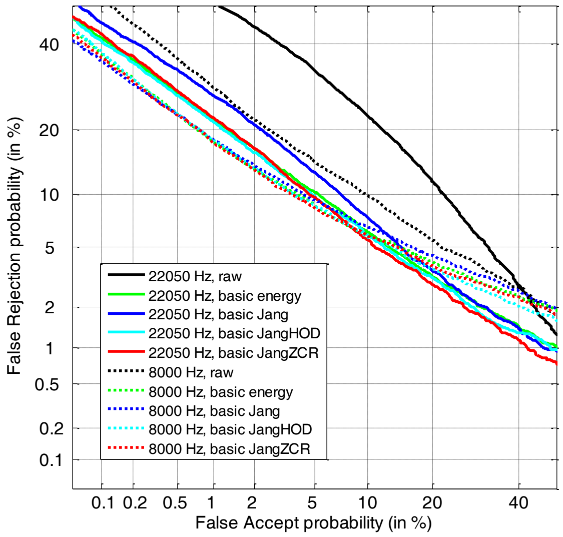

Fig. 3. FAR / FRR plots for basic algorithms of silence removal for VQ processing

GMM

Figure 4 shows FAR / FRR plots of speaker recognition based on 10 databases containing raw (unprocessed) speech files and files with removed silence at the beginning and at the end of sequence. It can be inferred that speaker recognition can be improved by silence removal algorithms regardless of the sampling rate of sequences. Each algorithm gives similar effect. EERs for files with 22050 samples/second are about 6-7 %. For unprocessed speech with the same sampling rate EER is about 11 % for GMM processing. The same sequences but with sampling rate

8000 samples/second give worse results. EER is very high, about 18-19% for all algorithms of removing silence and about 25% for unprocessed speech.

Results after removing silence between words

Vector quantization

In this case, silence was removed from all parts of recorded speech, e.g. between spoken words. Figure 5 presents FAR / FRR plots for raw and processed speech sampled with both sampling rates. It can be noticed, that enhanced Jang ZCR algorithm does not tend to improve for sampling rate equal to 22050 samples/second. Little correction can be observed for enhanced Jang algorithm. The greatest improvement is achieved by enhanced energy and JangHOD algorithm. In this case, EER lowered to about 5 %. In opposite to higher sampling rate, fs = 8000 samples/second gives in this case better results for unprocessed data, and similar results for enhanced silence removal algorithms, however, algorithms named Energy and Jang give the best results, at a sampling rate equal to 22050 samples/second.

GMM

Figure 6 shows a comparison of FAR / FRR plots for GMM processing. Of course, removal of every silence part in speech files improved efficiency of speaker identification, but every algorithm gives similar effect. EER for databases with removed silence is about 6-7 % in case of files with sampling rate of 22050 samples/second. Downsampled sequences with removed silent parts gives EER at 18-19%. For unprocessed speech EER is about 11 % and 25% for sampling rates of 22050 and 8000 samples/second, respectively.

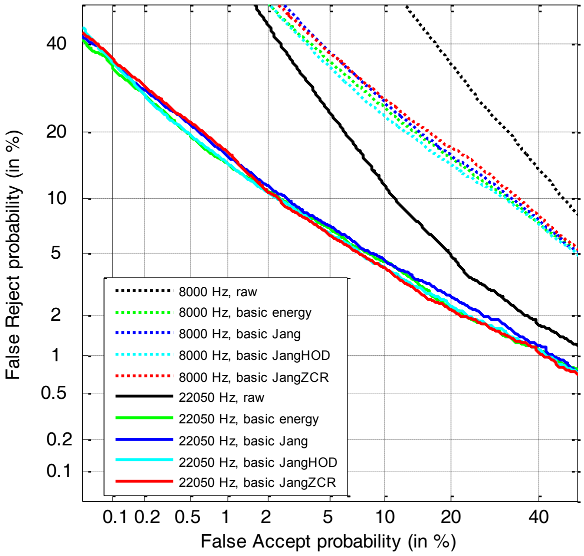

Fig. 4. FAR / FRR plots for basic algorithms of silence removal for GMM processing

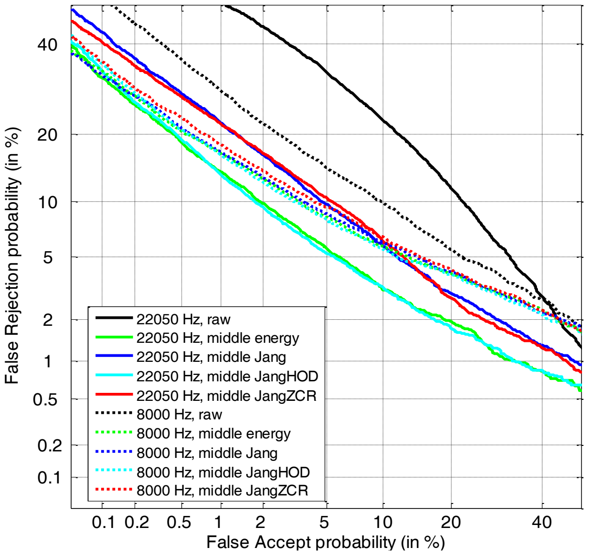

Fig. 5. FAR / FRR plots for enhanced algorithms of silence removal for VQ processing

Conclusions

As expected, the removal of unnecessary sections of the recording (i.e. the silence), resulted in an improved speaker identification performance of about 5-8 % for GMM -based speaker recognition and about 2-12 % for the VQ-based speaker recognition. The best results (i.e. 6-7% EER) of speaker identification were obtained for VQ at 22050 samples/second sampling rate, silence removing method: energy analysis or JangHOD (removing the silence from the whole sentence). In the case of VQ, better results for low sampling rate can be noticed. Influence of particular silence removal methods are quite similar, but the Jang HOD method and energy analysis can be distinguished. It should be noticed that a significantly noisy signal (e.g. during the telephone transmissions) may be more sensitive to the choice of the EPD method.

It seems that further improvement should be sought in the adaptive selection of the length of the analyzed blocks of samples. The authors also plan to investigate other VAD algorithms in time and frequency domain with the generalized autoregressive conditional heteroscedasticity (GARCH) [14].

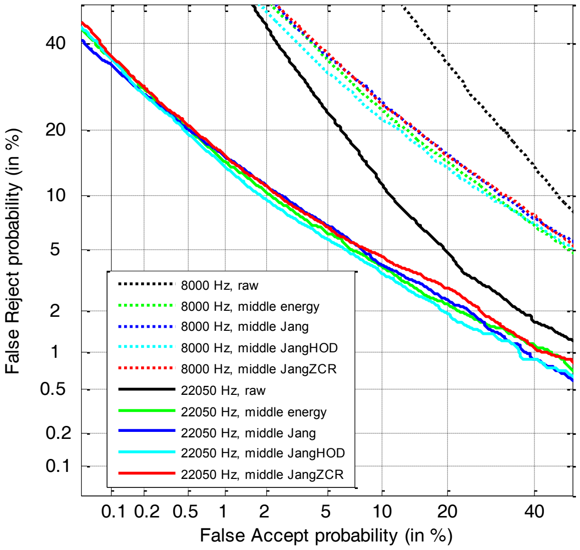

Fig. 6. FAR / FRR plots for enhanced algorithms of silence removal using GMM processing

References

[1] Keshet J., Bengio S., *Automatic Speech and Speaker Recognition*, John Wiley & Sons Ltd, (2009)

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

[3] Marciniak T., Weychan R., Drgas Sz., Dąbrowski A., Krzykowska A., Speaker recognition based on short Polish sequences, *Proc. of SIGNAL PROCESSING SPA'2010*, *Poland Section, Chapter Circuits and Systems IEEE*, Poznań, Poland, (2010), 95-98.

[4] Dąbrowski A. Marciniak T., Krzykowska A. Weychan R., , Influence of silence removal on speaker recognition based on short Polish sequences, *Proc. of SIGNAL PROCESSING SPA'2011, Poland Section, Chapter Circuits and Systems IEEE*, Poznań, Poland, (2011), 159-163

[5] Qi Li, Jinsong Zheng, Tsai A., Qiru Zhou, Robust endpoint detection and energy normalization for real-time speech and speaker recognition*, IEEE Transactions on Speech and Audio Processing*, Volume : 10 , Issue:3 (2002), 146 – 157

[6] Varela, O., San-Segundo R.,Hernandez L.A., Robust speech detection for noisy environments, *IEEE Aerospace and Electronic Systems Magazine*, Volume: 26, Issue:11, (2011), 16 – 23

[7] Kudinov M., Comparison of some algorithms for endpoint detection for speech recognition device used in cars*, Proc. of International Siberian Conference on Control and Communications (SIBCON)*, (2011), 230 - 233

[8] Marciniak T., Dąbrowski A., Rochówniak R., Subband wavelet signal denoising for voice activity detection, *Proc. of NTAV/SPA '2008*, Poznań, Poland, (2008), 93-96.

[9] Jyh-Shing Roger Jang, ASR (Automatic Speech Recognition) Toolbox, available from the link at the author's homepage at http://mirlab.org/jang.

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

[11] Voicebox: speech processing toolbox for Matlab http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

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

[13] DET-Curve Plotting software for use with MATLAB, http://nist.gov/itl/iad/mig/tools.cfm

[14] R. Tahmasbi, S. Rezaei, Change point detection in GARCH models for voice activity detection, IEEE Transaction on Audio, Speech, and Language Processing, Vol. 16, No.5, (2008),1038–1046

This work was partly supported by INDECT, PPBW, and DS 2012 projects.

***Authors**: dr inż. Tomasz Marciniak, mgr inż. Agnieszka Krzykowska, mgr inż. Radosław Weychan, Poznań University of Technology Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, ul. Piotrowo 3a, 60-965 Poznań, Poland, e-mail: tomasz.marciniak@put.poznan.pl*