J. Bobulski, Text independent Speaker recognition. Computing, Multimedia and Intelligent Techniques, 2007, 3, 1, August, 101-107.

TEXT-INDEPENDENT SPEAKER RECOGNITION

Janusz Bobulski

Czestochowa University of Technology, Institute of Computer and Information Science, Dabrowskiego Street 73, 42-200 Czestochowa, Poland
januszb@icis.pcz.pl

In this paper, text-independent speaker recognition method based on Wavelet Transform and mel-cepstrum is presented. The results of experiments point the best parameters of Wavelet Transform for speaker identification, and can be useful for design speaker identification systems. This kind method of person identification is useful in services such as banking by telephone, access authorization to resources and for forensic purpose

Keywords: speaker identification, voice recognition

Introduction

Speaker recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking base on speech signal. This method of persons identification use unique information included in voice of speaker, and allows verify their identity and control access to services such as voice dialling, banking by telephone, telephone shopping, database access services, voice mail, access authorization to resources and for forensic purpose.

Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification [1].

Speaker recognition methods are divided into text-dependent and text-independent methods. In case of text-dependent systems the speaker says key words or sentences having the same text for both training and recognition mode. Whereas in second case, the words don't matter.

Voice analysis

Fourier Transform is basic technique of voice analysis used for a long time. It is suitable only for stationary signals, because the Fourier transform gives no respect to where in time each frequency exists. It assumes that each frequency exists over all time. Practical signals are never stationary over their whole domain, but we can suppose that every signal is stationary on some interval. If a signal is break into

short intervals of time and take the Fourier transform over those intervals, this will give the frequencies of this signal and the location in time. The Short Term Fourier Transform solves the problem of knowing when the frequencies occur, but there appears a new problem. How to choose a good window for speech recognition? Taking into consideration frequency of speech the minimum is 20 Hz. If you take widow size 50ms (period of 20 Hz), FT will show not only frequency 20 Hz, but wider band. If you take window size 1s, the result of the Short Term Fourier Transform will no better than the Fourier Transform [2].

Another method of voice analysis is Wavelet Transform (WT). One major advantage afforded by wavelets is the ability to perform local analysis of a large signal. The input signal S passes through two complementary filters and we obtain two signal, *A* and *D* (Fig. 1). In wavelet analysis, we often speak about approximations and details. The approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components [3]. The process of decomposition of a signal can be repeated by recurrence. The result of this is more detailed data about process information. After first level wavelet decomposition, the output signals become input signals of second level decomposition (Fig. 2). Multilevel wavelet decomposition gives convenient choice of parameters depending on processing signal.

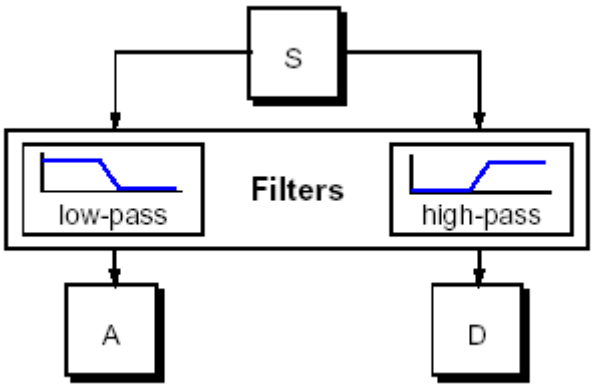

Fig.1 The filtering process in WT [10]

The frequency resolution increases during wavelet decomposition with each level, but decreases time resolution. We lost information about the time, but we get very precise energy distribution in individual range of the frequency. So, there is smooth transition from time domain to frequency one [2, 4].

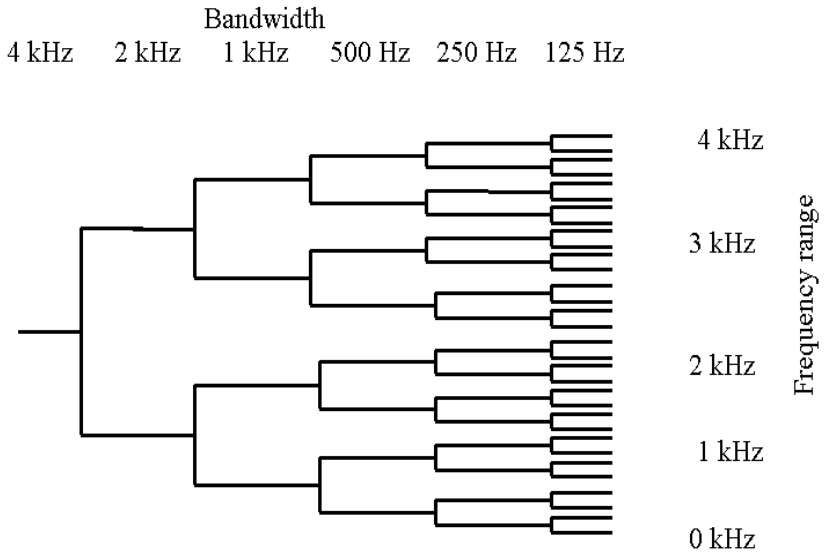

Fig.2 Multilevel decomposition of signal with WT

Cepstral analysis

The cepstrum of signal is defined as the inverse Fourier transform of the logarithm of the absolute value of spectrum. The absolute value of spectrum is a combination of cosine basis functions with varying frequencies. The cepstral coefficients are the magnitudes of the basis functions. The cepstrum coefficients are the Fourier series coefficients of the log-spectrum and that the Fourier series presentation reduces to cosine series. It means, log spectrum determine infinite summation of cosines of different frequencies, and the cepstral coefficients are the modulus of the basis functions. The lower cepstral coefficients represent the slow changes of the spectrum. The higher coefficients represent the rapidly varying components of the spectrum. In voiced speech sounds, there is a periodic component in the modulus spectrum, the harmonic fine structure that results from the vocal chord vibration [5].

Mel-cepstrum

Mel-cepstrum is one of the most commonly used parameters in speaker recognition. It consists in mel-scale based on experimental connection between the frequency of pure harmonic tone and frequency perceived by human. Basis on mel-

scale is made the filters bank, which does non-linear frequency analysis of signal. Mel filters bank is applied in the frequency domain before the logarithm and inverse DFT. The purpose of the mel-bank is to simulate the critical band filters of the hearing mechanism. The filters are evenly spaced on the mel-scale, and usually they are triangular shaped [5, 6].

Identification system

The structure of proposed method is shown on Fig.3. This system works in two modes, training and testing. These modes are different from each other. The algorithm of this method consists of four main parts:

1. Pre-processing.
2. Signal analysis with WT.
3. Features extraction with mel-cepstrum.
4. Training: saving to database.
   Testing: making a decision – distance measurement.

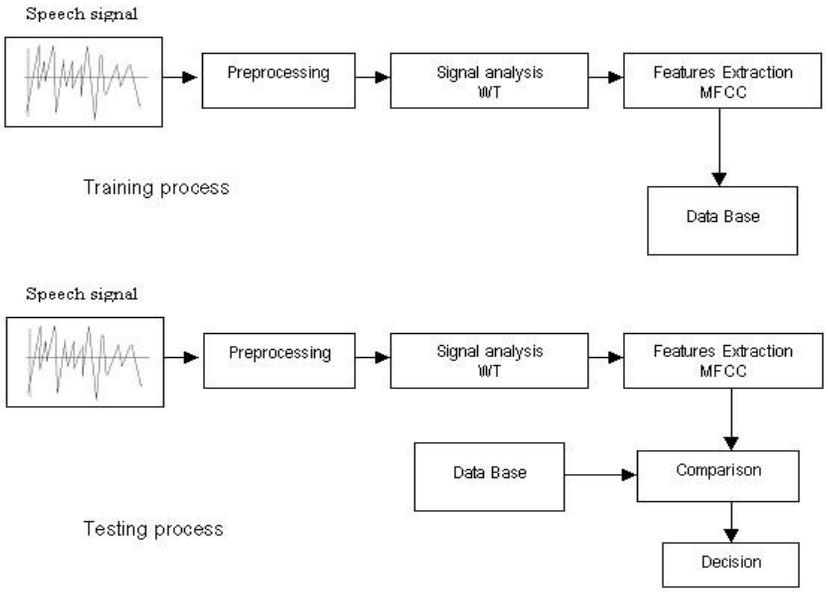

Fig. 3 Diagram of speaker identification system

The normalization is made in pre-processing. Normalization consists of the samples number verification in the input signal, and then the signal is filtered. There is remove the blow effect course of proximity of microphone.

WT is used to time-frequency analysis. There is possibility to choice the decomposition level from 4 to 16. Another option is choice spectrum bandwidth of analysis. All spectrum coefficients are calculating with WT and store in ASCII code.

Set of cepstral coefficients calculate in mel-scale is used as features of signal. The number of cepstrum coefficients has been made experimentally and is equal the number of WT coefficients.

The comparison of cepstral coefficients is made with distance measure and is made for each voice signal. The decision about identification is made on the base measure of distance between signals.

Experimenting

The experiments had made on voice signal base consist of 60 records of 30 persons, two signal per person, one for training one for testing. Signals are records of random independent text and have length 20s, sample frequency 48 kHz, and resolution 8 bits per sample.

The aim of research was determine the best parameters of WT for recognition. The parameters which was tested are as follows:

- Wavelet function
- Level of WT decomposition
- Spectrum size

Results

The aim of first experiment was to point the best wavelet function for speaker recognition. The best results get for db10*, db16, coif 2, sym10* (Tab.1), but with respect to calculate time the quickest wavelet is *db10*. The calculate time with some wavelet function was two time longer with the same recognition rate. The results were worse for simple function like *haar, db4-db8, roob, vaid*. Better results we can get for wavelet function with long impulse response.

The depth of WT decomposition was aim of second experiments. There are made analysis efficiency from level 5 to 16. The highest recognition rate is with 9 and 10 level of decomposition (Fig.4). The levels under 9 have worse recognition rate, and levels above 10 have the same recognition rate, but the calculate time was longer. To sum up: the best wavelet function is *coif2* at 9th level of decomposition.

Third research aspect was determining the optimal size of spectrum. The previous experiments were carried on with the full spectrum size. Optimal spectrum size is about 11,5 KHz. Shorter band doesn't give sufficient information for an identification, because there is the most useful information about voice and

speaker in this band. Longer band is no needed because there are few useful information above this band but there are high-frequency noise.

Tab.1 Recognition rate for 9 and 10 level decomposition

Wavelet function	Recognition rate [%]	
	Level 9	Level 10
Haar	88,89	83,89
db4	90,56	89,44
db6	91,11	88,33
db8	90,28	88,33
db10	91,94	90,28
db12	90,28	90,00
db14	89,72	89,17
db16	90,56	90,28
db18	90,83	89,44
db20	90,28	89,17
db50	89,44	89,44
rob16	77,50	71,67
Vaid	89,44	89,17
coif1	90,56	88,89
coif2	92,50	90,56
coif3	70,28	66,11
coif4	90,28	90,00
coif5	90,56	89,72
sym4	91,39	89,72
sym5	91,39	88,61
sym6	91,39	90,00
sym7	89,17	88,61
sym8	90,00	90,28
sym9	88,89	89,72
sym10	91,11	90,28

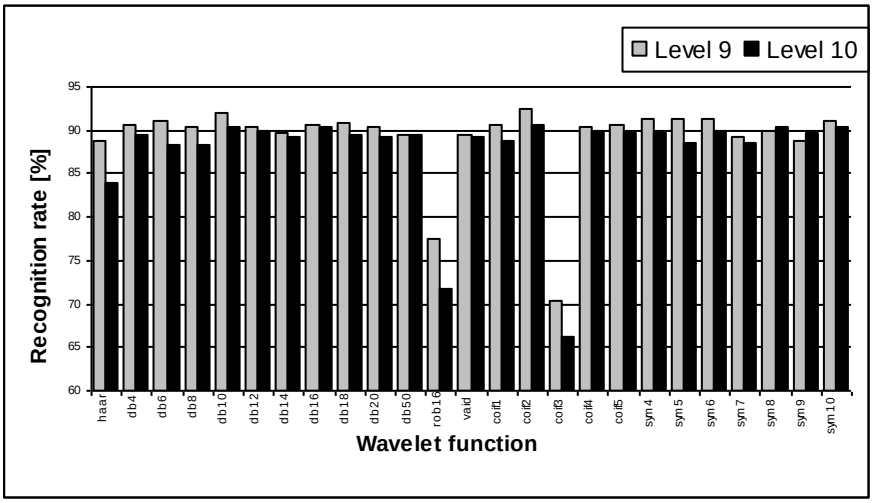

Fig. 4 Recognition rate for wavelet functions at 9th and 10th decomposition level

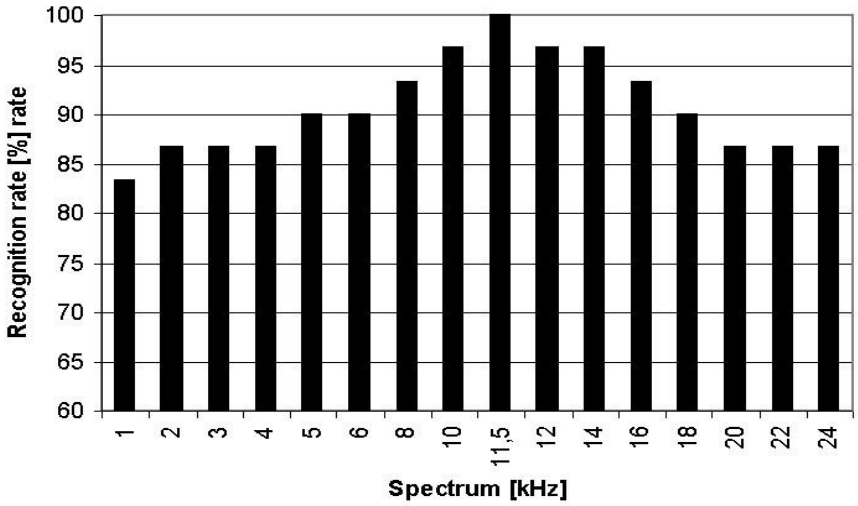

Fig. 5 Recognition rate depending on spectrum size analysis

Conclusion

The results of experiments show that text-independed system of speaker recognition based on WT and mel-cepsum can be built. The system worked correctly, had good recognition rate and proved that premises were correct. The present result can be useful for design speaker identification systems.

References

[1] Sadaoki Furui, *Survey of the State of the Art in Human Language Technology*, Tokyo, Japan, http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html, 1996.
[2] Aboufadel E., *A Wavelets Approach to Voice Recognition*, http://www.gvsu.edu/math/wavelets/student_work/Hoekstra/voice_recognition.htm - 29 November, 2001.
[3] Misiti M., Misiti Y., Oppenheim G., Poggi J.-M.: *Wavelet Toolbox, MathWorks,* 1998.
[4] Augustyniak P., *Transformacje falkowe w zastosowaniach elektrodiagnostycznych*, Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków 2003 (in Polish).
[5] Kinnunen T., *Features for Automatic Text-Independent Speaker Recognition,* http://www.cs.joensuu.fi/pages/pums/public_results/2004_PhLic_Kinnunen_Tomi.pdf - 21st December 2003.
[6] Kubanek M., *The Method of Audio-Visual Polish Speech Recognition Based on Hidden Markov Models*, PhD Thesis, Czestochowa University of Technology, Czestochowa, Poland, 2005 (in Polish).

Janusz Bobulski, PhD Eng. Currently is Assistant Professor at Czestochowa University of Technology (Poland), Faculty of Mechanical Engineering and Computer Science, Institute of Computer and Information Science, Department of Multimedia and Biometrics Techniques. His research interests are biometrics, image processing, and person's identification.