Hybrid Wavelet-Fourier-HMM Speaker Recognition

Bartosz Ziółko, Wojciech Kozłowski, Mariusz Ziółko,
Rafał Samborski, David Sierra, Jakub Gałka
Department of Electronics,
AGH University of Science and Technology
Kraków, Poland
www.dsp.agh.edu.pl
{bziolko, ziolko, sambo, jgalka}@agh.edu.pl

July 6, 2011

Abstract

The paper presents successful experiments on combining two speaker recognition methods into a hybrid system. The first branch of recognition is an innovative approach based on discrete wavelet-Fourier transform. The second one is classic, based on HTK and classification into voice and unvoice segments. The hybrid solution outperforms both on a small test set. Further tests of this soultion will be conducted and reported.

1 Introduction

The latest technology improvements in such fields as banking, communications and networking require the latest advances in security systems. In the last years a new kind of biometric identification has risen among the others as it is printed in subjects' bodies, it is impossible to lose, almost impossible to duplicate and univocally designates a person. The biometric keys exploit some human characteristics that are different and unique in each person such as fingerprints, DNA chains, face shape and voice.

Voice is a phenomenon that is highly dependent on the speaker. Many physical aspects of speech such as the timber, tone or intensity vary a lot from a speaker to another one. The same happens with other linguistic aspects as the intonation and range of vocabulary or expressions a speaker normally uses. All these properties make voice a very powerful biometric key to be used in security systems since the physical characteristics of speech are easy to measure in comparison to other biometric keys. In addition, the speech signal has been deeply studied for many years so many powerful algorithms are found to deal with this kind of signal.

A good biometric key has to match certain requirements. It has to be easy to extract, measure, save and compare. Voiceprints match all these requirements since not a very expensive hardware is needed to perform all these operations. In fact, the only infrastructure needed is a microphone and a PC. A medium-quality microphone is relativelly cheap hardware if one compares it to a digital camera, iris or finger scanner, not mentioning a DNA analyser. In addition, many new banking applications rely on the usage of telephone line. Voiceprints are the only suitable biometric technology able to operate in this environment.

There are a number of features in a speech signal which make recognising a speaker possible. The appropriate transformation of speech is an important problem because the representation in the time domain gives little information about the speech signal properties. It is necessary to obtain the optimal spectral representation. Usually, methods which are based on the Fourier or wavelet transform are used. In this way, the frequency properties of speech are analysed. To improve recognition, we tested a hybrid system based on Wavelet-Fourier Transform (WFT) and more traditional HTK based system on Mel Frequency Cepstral Coefficients (MFCC) with voiced/unvoiced classes.

The aim of a speaker recognition system is to efficiently provide accurate and distinguishable individual properties of each speaker. Specific feature in individual speech are always the basis of a speaker recognition system. It is especially important, if a new speaker representation is introduced. It must be cleared out whether the representation carries suitable signal features or not. There are various methods of speaker recognition, where multivariate kernel density [1], gaussian mixture models [2], artificial neural networks [3] or support vector machine [4] were used.

There are numerous advantages in using voiceprints. Voice is a phenomenon that doesn't require the subject to be present. It can be recorded in some place and sent to another almost instantly. In addition, almost every user already has a microphone in PC or telephone whereas not many can afford a camera or a finger scanner. Finger-prints, images or face-scans require to use more complicated devices.

The goal of our research was to analyse, if the WFT can improve traditional speaker recognition methods. The system should be text independent and should be based on the speech characteristics such as accents, speaking styles and disfluencies. Wavelet packets were already tested for speaker verification [5]. Perceptually motivated wavelet based methods were successfully tested for speech recognition [6, 7, 8]. Another method of reduction of a wavelet decomposition tree for a speaker identification task was presented [9], however, described conclusions about their solutions being optimal are questionable.

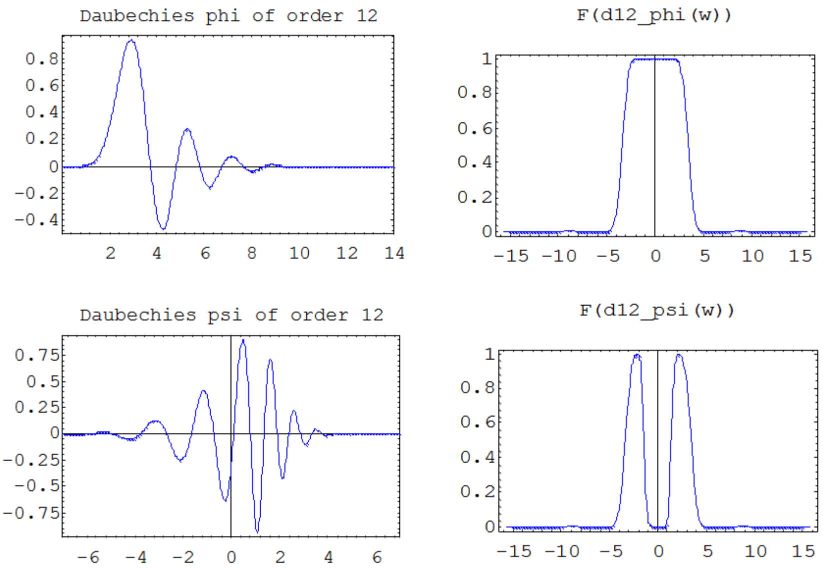

Figure 1: *Daubechies wavelet and scale functions with their amplitude spectra*

2 Models of voice generation

Knowledge of how voice is produced and perceived by a human being playes an important role in speech technology systems [10]. Speech is the result of activity in the various elements of the speaker's respiratory system. All of them contribute somehow to the final speech signal. Every block introduces some speaker-dependent information in the speech signal, however, only some of them can be exploited.

2.1 Lungs

The lungs are used for the vital function of inhalation and exhalation of air. In the speech production model they are the power source that supplies energy to the rest of the blocks in the systems. Inhalation is achieved by reducing the lung air pressure. This is possible thanks to the rib cage and the diaphragm. The rib cage is expanded during this process. The diaphragm, which is placed underneath the lungs, is lowered so the lungs are expanded. This pressure lowering causes air to rush in through the vocal tract and down the trachea into the lungs.

Exhalation is opposite to inhalation. It is caused by an air pressure increase in the lungs. The volume of the chest cavity is reduced by contracting the muscles in the rib cage and lifting the diaphragm. This produces an air flow from the lungs to the larynx through the trachea.

Inhalation and exhalation always rhythmically follow the one to the other when breathing. However, during speaking short spurs of air are taken and steadily released by controlling the muscles around the rib cage. The rhythmic breathing is overridden since expiration takes one sentence or phrase time. During this time the air pressure remains almost constantly above atmospheric pressure. However, the time-varying properties of the larynx and the vocal tract cause this constant pressure to become time-varying.

This airflow produced by the lungs has the shape of white gaussian noise. The only speaker-dependent information that is introduced by the lungs is the energy of this noise. However, this is not discriminative enough and other features have to be found.

2.2 Larynx

The larynx, also called the "voicebox" is a complex system of cartilages, muscles and ligaments. It has different functions such as closing the entrance to the lower respiratory system during swallowing. Since this kind of functions is not important for the speech production models, they will not be analysed. From the voice production point of view, the most important parts of the larynx are the vocal folds and the glottis.

The vocal folds are two twin masses of flesh, ligament and muscle which stretch between the front and the back of the larynx. Their size varies from one person to another and in average it is around 15 [mm] long for men and 13 [mm] long for women. They can remain open to create unvoiced sounds or they can vibrate in order to produce voiced sounds during speech. During breathing, they remain open, allowing the air to flow into the lungs.

Voiced sounds are characterised by the vibration of the vocal folds which open and close the airflow exit very quickly. This process is known as the Bernoulli's Principle in the glottis and consists of three steps:

- Initially, the vocal folds are open. Airflow is produced in the lungs and the vocal folds are immediately closed due to air pressure effects. It causes a vocal folds tension increment to hold a high air pressure as the lungs continue to pump air through the trachea.

- Next is the voicing step. The vocal folds cannot sustain the high pressure generated and have to open. However, the pressure is reduced again so the vocal cords can be closed again.

- The vocal cords go back to the same position, so the whole process can be repeated again. This process is done several time in a very short period causing the vocal cords to vibrate and introduce a quasi-periodic pulse in the airflow. Pitch (F0) is the frequency of vocal folds vibrations.

This process creates a source signal for voiced sounds (such as vowels) which will be lately modified by the vocal tract. F0 is higher for females than for males due to anatomical characteristics such as the length and mass of vocal folds which is lower in the case of women. In the case of children pitch is even higher. Therefore, estimation of the pitch can be a good gender or age discriminator. In addition, pitch does not remain constant during speech. Some systems which use prosodic features take this pitch evolution into account despite the fact that it is relatively easy to imitate by an impostor.

Some models have been created to model the airflow velocity output at the glottis. The vocal folds are opened for a very short period. On the other hand, the breathy voice and the vocal folds remain open for a longer time. Both glottis responses show a pitch of 200 [Hz]. In the frequency domain, it can be seen that the longer the glottis remains open, the higher spectrum roll-off it shows. In addition, the spectrum contains a peak in every multiple of the fundamental frequency. Therefore, voice quality depends on the glottal pulse shape.

Speaker individuality is also present in the quality of voice. This quality is lower in the case of a breathy voice, as the glottis is not almost closed during the vocal folds vibration. This effect makes the glottal pulse spectrum to decrease with frequency in a faster way. However, as this glottal pulse will be later modified by the vocal tract, it is very hard to extract reliable speaker-independent information based on this issue.

2.3 Vocal Tract

Vocal tract refers to the voice production organs above the larynx. They are the pharyngeal, oral and nasal cavities and the velum. They constitutes the main source of speaker-independent features which can be easily extracted. The length and shape of the vocal tract are highly speaker-dependent. Though, they are not fixed constants, however, their contribution to the speech signal spectrum is very important and relatively easy to measure.

The cavities themselves act as resonance boxes and spectrally colour the source airflow coming from the larynx. The frequencies at which the vocal tract resonates are formants. The velum controls the volume of air that flows into the nasal cavity. However, the vocal tract is not still during phonation. Its shape is time-varying as the movement of the tongue, lips, teeth, jaw results in a change in the vocal tract section. Each change is aimed to produce a different phoneme. For example, the tongue may have to move from the top of the oral cavity to the bottom for a new phoneme. This movement is done with some delay. As a result there will be a transition time between both phonemes – coarticulation. Because of this effect, frame overlapping is normally used. In addition, when a change in the position of the articulators is done, they remain in the same position, for a time period of 20-40 [ms]. During this period the voice signal can be considered as a stationary process.

All these properties make the vocal tract a very complex system, since it can be regarded as a time-varying voice-box. The size and a shape of each cavity varies from one speaker to another. Vowels are created when the quasi-periodic impulses generated in the glottis flow without opposition into the oral cavity. In voiced nasals the velum partially closes the conduct that leads to the oral cavity and only allows the air to flow into the nasal cavity. If the oral and nasal cavities could be approximated by a linear filter, the spectrum of a nasal or a vowel would contain much information about the nasal and oral cavities spectrum.

2.4 Source-filter model for speech production

If the vocal tract is considered as a linear system, the final speech signal can be seen as the result of filtering an excitation signal with the vocal tract impulse response. The excitation signal can be white gaussian noise in the case of unvoiced sounds or white noise convoluted with the quasi-periodic impulse train generated in the glottis.

The spectrum of the excitation signal is then coloured by the vocal tract. The vocal tract spectral response is characterised by the presence of certain resonant frequencies or formants. For voiced speech, these formants emphasise the source spectrum in areas close to their location. It is generally accepted that only the first four ones carry relevant information on the speech signal. The central frequency of each formant is numbered and is referred as F1, F2, F3 and so on. Normally F0 is reserved for the pitch frequency.

An important side effect of this theory is that the excitation source is considered independent from the vocal tract response. Some methods try to separate both components and use the information independently. The vocal tract transfer function is the most important for the speaker recognition task. It contains information about the speaker's vocal tract shape which is extremely hard to imitate.

If the articulatory configurations of two speakers are assumed to be the same and the only difference is the length of the vocal tract, then the acoustic theory predicts that the formant frequencies are inversely scaled by the ratio of the speakers' vocal tract lengths. Thus, estimation of the formants location is a good speaker-discriminator.

3 Wavelet-Fourier Transform

The standard transform used for speech signals analysis is the fast Fourier transform (FFT) which gives averaged representation of a signal in the frequency domain. Short Fourier transform is capable of carrying time-frequency changes, however, analysing windows creates artefacts.

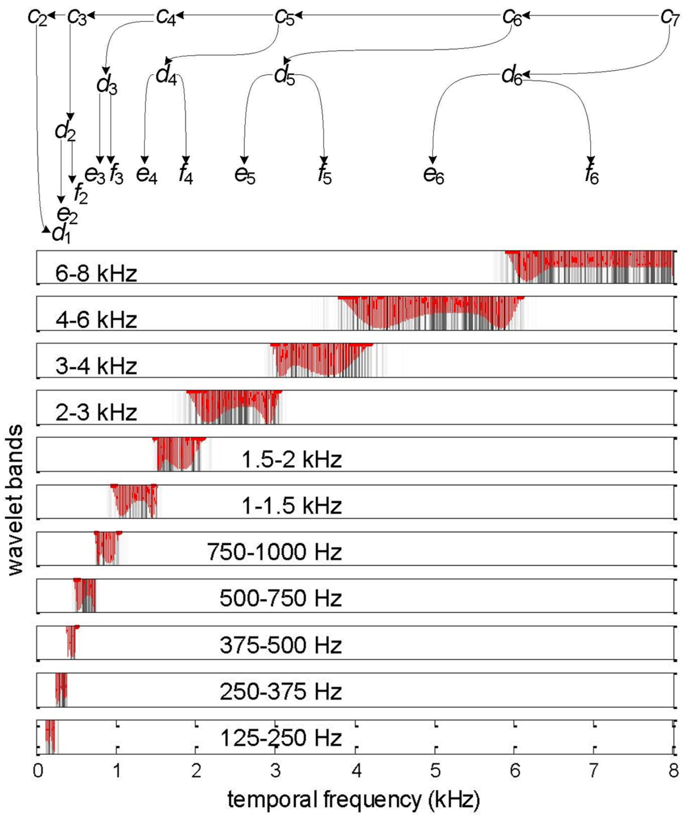

Figure 2: *The frequency bands for speech signal analysis are perceptually motivated. They were chosen in the aim of representing the frequencies most important for a speech signal in several narrow subbands and less important frequencies in wide subbands. The general structure of decomposition tree was found in a process of optimisation, however, the final structure is its slightly smoothed version*

The discrete wavelet transformation (DWT) belongs to the group of frequency transformations and is used to obtain a time-frequency spectrum [11, 12] of signal $\{s(n)\}$. This encourages us to use the DWT as an artificial method of speech analysis. Dyadic frequency division makes the DWT much more compatible with the principles of the operation of human hearing system, equipped with subsystem for frequency analysis (to reveal the important information for the human speech recognition ability), than other methods.

The wavelet transform (WT) is defined by formula

$$\widetilde{s}_{\psi}(a,b) = \frac{1}{\sqrt{a}} \int_{\infty}^{\infty} s(t)\psi\left(\frac{t-b}{a}\right) dt, \tag{1}$$

where $a \in \Re^{+}$ and $b \in \Re$. The two arguments function $\widetilde{s}_{\psi}(a,b)$ represents wavelet spectrum of signal $s(t)$. Parameter a, called scale, reversibly correlates with a frequency. Parameter b is a time translation. Function $\psi(t)$ is an arbitrary chosen wavelet and its example is presented in Fig.1.

Formula (1) shows, that wavelet spectrum carries both, time and frequency representations. The events can be captured precisely, because the analysing wavelet window fits into frequency. WT (1) has a simple physical interpretation: the analysing function $\psi(t)$ is a flexible time-scale window that automatically narrows at high frequencies and widens at low frequencies. A WT depicts information about the signal changeability in the time domain. This kind of analysis provides valuable information about voice irregularity in the time domain, according to frequency variations.

It is an important property that the WT (1) has the form of a correlational operator. It enables us to apply the Fourier Transform (FT) to the WT and define the new method of speech analysis. Let us consider the composition of two transforms. For a speech signal, the wavelet spectrum is calculated first and next the FT is used to obtain

$$\widehat{\widetilde{s}}_{\psi}(a,\omega) = \frac{1}{\sqrt{a}} \int_{\infty}^{\infty} e^{-j\omega b} \int_{\infty}^{\infty} s(t)\psi\left(\frac{t-b}{a}\right) dtdb. \tag{2}$$

FT is calculated with respect to the variable b, and the coefficient a plays the role of constant parameter only. The wavelet-Fourier spectrum has two arguments. The first one describes the frequency band, where its average value is proportional to $1/a$ and ω, the second one, denotes the frequency in which the previous frequency appears in the signal.

Formula (2) plays a role of WFT definitions and has small usefulness due to a large amount of calculations in numerical computing of integrals. To improve the computer calculations, DWT is used instead of (1) and FFT instead of FT in (2).

For each wavelet $\psi(t)$ (see [11]) the scaling function $\varphi(t)$ is defined. These both functions have unique character, in a sense that each wavelet function $\psi(t)$ has only one scale function $\varphi(t)$.

Each function $\varphi(t)$ can be used to build a set of basis functions

$$\varphi_{m,n}(t) = \sqrt{2^m}\varphi(2^m t - n)\ . \tag{3}$$

Let coefficients $c_{6,n}$ of the series

$$s_6(t) = \sum_n c_{6,n}\varphi_{6,n}(t) \tag{4}$$

where

$$\varphi_{6,n}(t) = 2^3\varphi(2^6 t - n) \tag{5}$$

be the values of the DWT for five resolution levels. The coefficients of the lower levels are calculated by applying the well-known [11] formulae

$$c_{m-1,n} = \sum_k h_{k-2n}c_{m,k} \tag{6}$$

$$d_{m-1,n} = \sum_k g_{k-2n}c_{m,k}, \tag{7}$$

where $m = 6$ was taken, h_{k-2n} and g_{k-2n} are the constant coefficients which depend on the assumed wavelet $\psi(t)$ and the scale function $\varphi(t)$. The coefficients of next resolution levels are calculated recursively by applying formulae (6) and (7) for $m = 5, 4, \ldots$. In this way values

$$DWT = \{d_6, \ldots, d_1, c_1\} \tag{8}$$

of the DWT for seven levels are obtained where vectors d_m consists of elements d_m, n and vector $c_1 = [c_{1,n}]$.

Classic discrete decomposition schemes are dyadic and do not provide sufficient number of frequency bands for effective speech analysis. Wavelet packets provide more frequency bands [13]. A wavelet decomposition structure which provides a perceptual frequency analysis is suggested. It was obtained by removing decomposition tree nodes to the best possible approximation of the perceptual frequency division for the given number of decomposition levels and desired frequency bands. Our case is presented in Fig.2. Spectra for the 10 required frequency subbands are computed by applying procedures

$$e_{m,n} = \sum_k h_{k-2n}d_{m,k}\ , \tag{9}$$

$$f_{m,n} = \sum_k g_{k-2n}d_{m,k}\ , \tag{10}$$

for $m = 2, 3, \ldots, 6$. Finally, the spread discrete wavelet transform

$$SDWT = \{f_6, e_6, \ldots, f_2, e_2, d_1\} \tag{11}$$

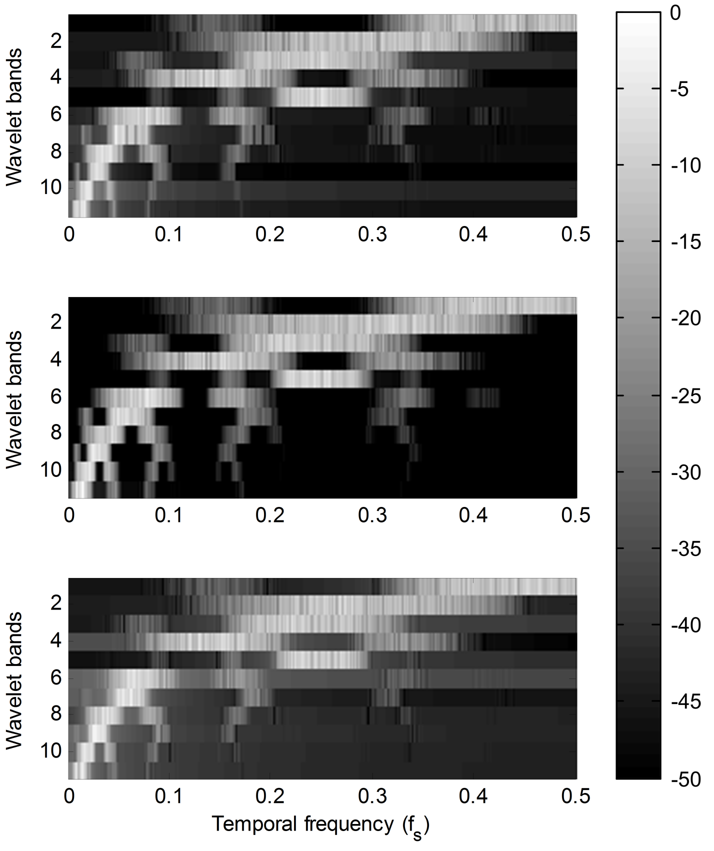

Figure 3: *Example of DWFT spectra for three different speakers and 11 resolution levels in a [dB] scale. DWFT allows easy and detailed analysis, similarly to traditional spectrograms but are more efficient*

is obtained.

The information about the lowest frequency band, from 0 to 125 Hz, is represented by a vector c_1. This part of DWT was skipped in the spectral representation because it carries a relatively strong noise and little information about speech.

Next, the Discrete Wavelet-Fourier Transform (DWFT)

$$DWFT = \left\{ \hat{f}_6, \hat{e}_6, \dots, \hat{f}_2, \hat{e}_2, \hat{d}_1, \right\} \tag{12}$$

is computed by applying the FFT to each level separately.

The DWFT spectrum gives the specific and individual frequency characteristics for voices of each speaker. The sampling frequency have been set to 16 [kHz]. The DWFT spectra were calculated for eleven resolution levels. Their absolute values in [dB] scale are presented in Fig.3. The high-resolution levels are presented in the upper part of this plot.

4 System architecture

Our system is based on two sepearate branches (Fig. 4). One of them is based on DWFT analysis. The second branch analyses the signal using clasification into voice and unvoiced speech segments and HTK [14]. The whole speech frequency band (125 [Hz] - 8 [kHz]) was considered in the experimental analysis. In case of telephone applications, the band would be limited. Further experiments will be conducted to test the method efficiency in telecommunication systems.

The final decision is taken after weighting of scores of both branches. Experiments were conducted to find optimal weight for both branches. Preliminary, we can conclude from them that the weight should be around 0.5.

The normalised amplitude spectra

$$u_{m,n(i)} = \frac{|\widehat{f}_{m,n(i)}|}{\sqrt{\sum_m |\widehat{f}_{m,n(i)}|^2}} \tag{13}$$

$$v_{m,n(i)} = \frac{|\widehat{e}_{m,n(i)}|}{\sqrt{\sum_m |\widehat{e}_{m,n(i)}|^2}} \tag{14}$$

were computed for all resolution levels $m = 2, \dots, 6$, where $n = 1, \dots, M$ is the number of the speaker and $i = 1, \dots, N$ is the number of his utterance.

The average value of normalised amplitude spectra

$$a_n = N^{-1} \sum_{i=1}^{N} |\widehat{d}_{1,n(i)}| \tag{15}$$

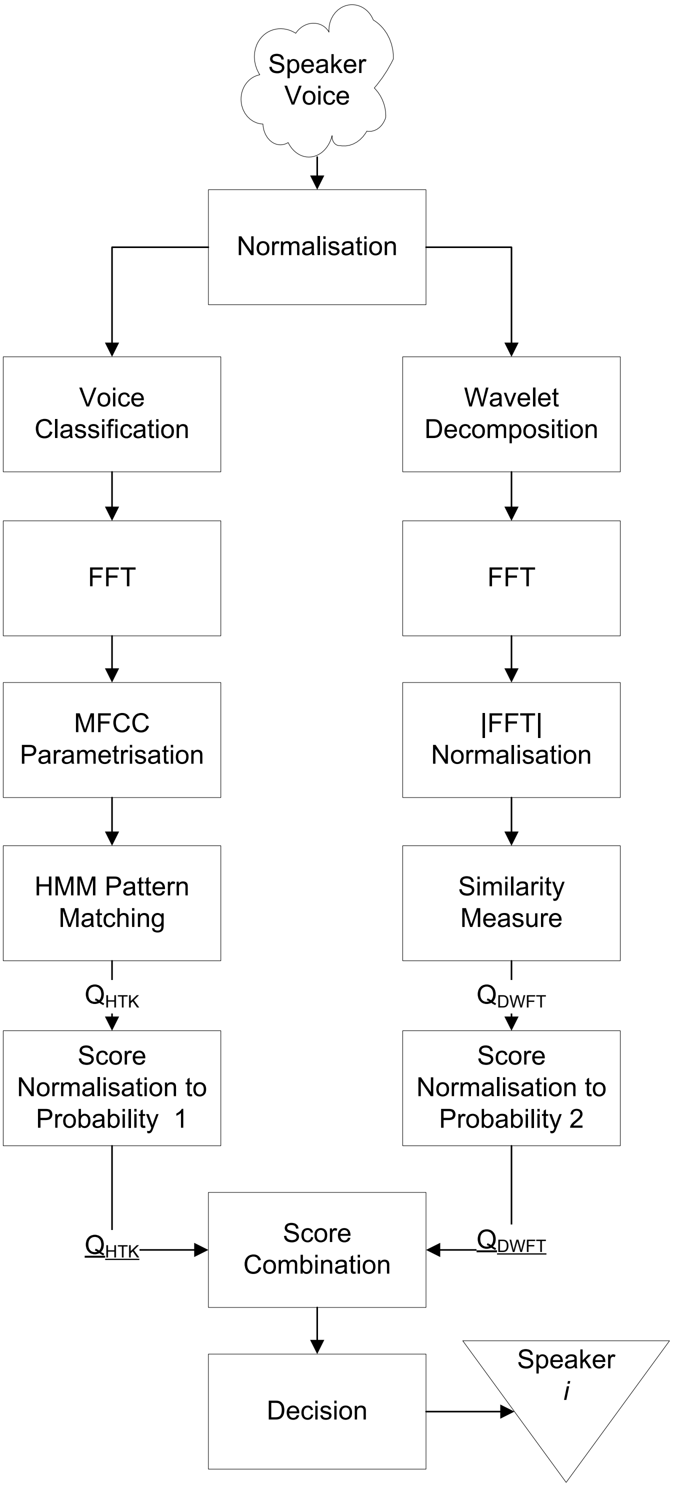

Figure 4: *Architecture of the described hybrid speaker recognition system applying DWFT and HTK*

$$u_{m,n} = N^{-1} \sum_{i=1}^{N} u_{m,n(i)} \tag{16}$$

$$v_{m,n} = N^{-1} \sum_{i=1}^{N} v_{m,n(i)} \tag{17}$$

for $m = 2, \ldots, 6$ frequency bands, create an individual characteristic for each speaker.

The second, more traditional branch of the system is based on HTK with 23 MFCCs, 26 filter banks and no cepstral mean normalisation. This set was found the most effective in our previous experiments. HTK voiced frames models have 8 states with average 10 mixture components per state and unvoiced frames have 3 states with 3 mixture components per state. Statistics for assigning number of states for voiced frames were used.

The speaker recognition procedure relies on a comparison of the spectra of persons to be identified with characteristics of all speakers collected in a database. In the first branch of the system, a voice recorded for an unknown person is processed in a way described by formulae (6)-(17) to obtain its spectra a_x, $v_{m,x}$ and $u_{m,x}$ where x is an index of a speaker being verified. In the input of the system we have two utterances which are going to be compared. The length of both statements are the same and in our experiments it was set to 10 [s], what is equal to $K = 160\,000$ samples. The similarity measure from the first branch between verified and n-th speaker is defined as l^1 metric

$$Q_{DWFT}(n, x) = |a_x - a_n| + \sum_{m=1}^{11} \left(|u_{m,x} - u_{m,n}| + |v_{m,x} - v_{m,n}| \right). \tag{18}$$

The obtained result gives information about similarity of two voices. The smaller value means a higher similarity.

The second branch provides Q_{HTK}, which is a measure from HMM and Viterbi algorithm. Then Q_{DWFT} and Q_{HTK} are normalised to probability-like values represented by $\underline{Q}$. The final score is provided by weighting

$$Q(n, x) = w\underline{Q}_{DWFT}(n) + (1 - w)\underline{Q}_{HTK} , \tag{19}$$

where $w \in [0\ 1]$ can be found experimentally.

Our solution is so far only a prototype to test the described method. This is why the results presented in the next section are given as a similarity matrix rather then some scenario simulating practical usage. It helps to evaluate the efficiency of the hybdrid algorithm, as much as possible on limited data. The algorithm will be evaluated on a larger set of data and turned to a working, practical application later on, for crime investigations and corporation call-centre services.

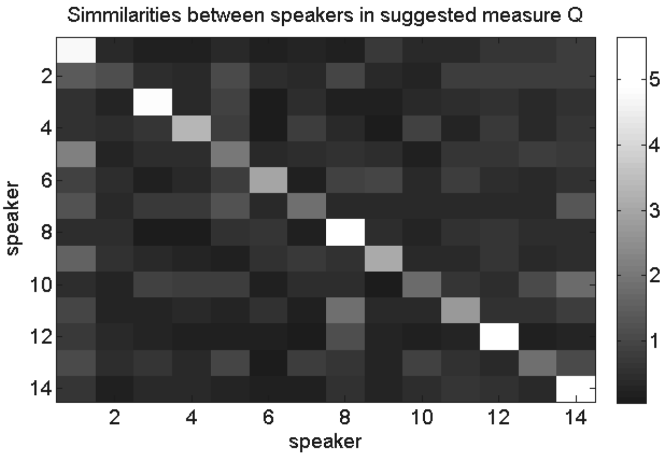

Figure 5: *Matrix of speaker-to-speaker similarities using measure (19)*

5 Results

The speech signals applied to check properties of the recognition are part of TIMIT database. TIMIT is an acoustic-phonetic continous speech corpus developed by many institutions, including Texas Instruments (TI) and Massachusetts Institute of Technology (MIT), hence the corpus' name. The corpus includes 16-bit, 16 kHz speech waveforms with ortographic, phonetic and word transcriptions. The speech was recorded at TI and can be considered as studio recordings with very high signal-to-noise ratio.

In every conducted test 14 waveforms length of 10 seconds recorded by different speakers were chosen to calculate speaker-to-speaker similarities according to (19). Only male voices were considered to make the comparision of all pairs of speakers as difficult as possible. In Fig. 5 matrix of speaker-to-speaker similarities cointaining $14 \cdot 14 = 196$ comparisions from one of the experiments is presented. The matrix is not symmetric because columns represent the average spectra and rows represent spectra for one utterance, only. The higher the similarity coefficient $Q(n, x)$ is, the more similar sample x is to an average spectrum of nth speaker. The elements on the diagonal have clearly the highest values in each row.

In the test presented in Fig. 5, all speakers were recognised properly in case of properly chosen weight of both branches of the hybrid system. The recognition for different weights is presented in Fig. 6. It clearly shows that DWFT branch alone gives recognition worse then 80%. However, also HTK-based system alone is not perfect having recognition around 90%. The equally weighted hybrid system using both branches gives 100% recognition rate.

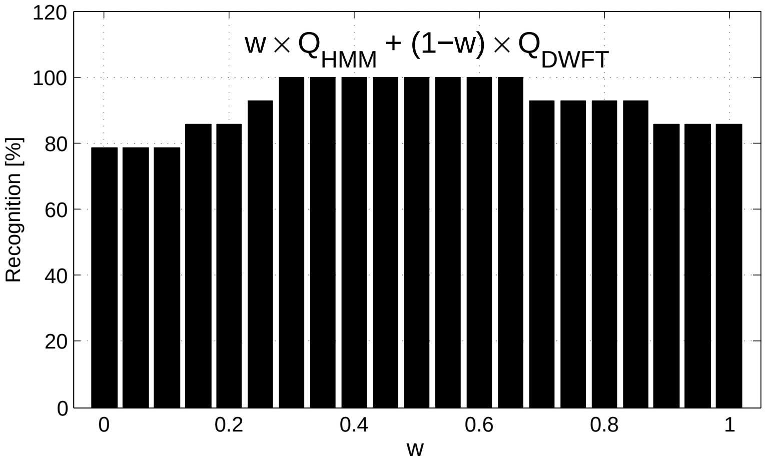

Figure 6: *Recognition for various weights w of impact of DWFT and HTK branches*

6 Conclusions

The transform applied to speech, must not only extract frequency information from a signal, but should keep the individual properties of each speaker as well. The combination of wavelet and Fourier transforms that we used, captures all the same frequencies in the same region, which makes it easier to localise them. Moreover, a composition of these transforms makes it possible to detect the specific voice signal properties. These properties have important features for a speaker recognition system. As shown in Fig.3, information important for further processing has singular distribution. The combination of two methods makes it possible to detect additional speech characteristic properties. They arise from the simultaneous exploitation of the advantages of both, the wavelet and the Fourier methods. It is possible to observe some characteristic irregularities which are not directly visible in either the wavelet or Fourier spectrum.

Information derived from WFT analysis is useful in the case of the analysis of quasi-harmonic signals, like speech signal. It allows us to estimate the significance of some spectrum irregularities. These irregularities together with particular behaviour in the time domain determine the specific tone and colour of the sound. They are features which are different then traditional ones, which allowed to improve HTK based speaker recognition system significantly, up to 100% recognition for a test corpus, however, it is not too large.

All procedures used to compute speech spectra, i.e.: DWT, FFT, are simple and quick. So the method described in this article enables to build a fast speaker recognition system. It is possible to obtain the speaker recognition in real-time systems.

7 Acknowledgements

This work was supported by MNISW grant number OR00001905.

References

[1] G. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data multivariate kernel density (mvkd) versus gaussian mixture model-universal background model (gmm-ubm)," *Speech Communication*, vol. 53, pp. 242–256, 2011.

[2] Y. X. Shan and J. Liu, "Robust speaker recognition in cross-channel condition based on gaussian mixture model," *Multimedia Tools and Applications*, vol. 52, pp. 159–173, 2011.

[3] J. Wu and Y. J. Tsai, "Speaker identification system using empirical mode decomposition and an artificial neural network," *Expert Systems with Applications*, vol. 38, pp. 6112–6117, 2011.

[4] S. Zhang and M. Mak, "Optimized discriminative kernel for svm scoring and its application to speaker verification," *IEEE Transactions on Neural Networks*, vol. 22, 2011.

[5] T. Ganchev, M. Siafarikas, and N. Fakotakis, "Speaker verification based on wavelet packets," *Lecture Notes in Computer Science - Text, Speech and Dialogue, Springer*, 2004.

[6] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Processing Letters*, vol. 8, no. 7, pp. 196–198, 2001.

[7] ——, "Wavelet based robust subband features for phoneme recognition," *IEEE Proceedings: Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 187–193, 2004.

[8] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," *Proceedings of ICASSP*, vol. Istanbul, 2000.

[9] H.-W. Chen and T. Olson, "New aggressive way to search for the best base in wavelet packets," *IEEE Proceedings of Vision and Image Signal Process*, vol. 152, no. 6, 2005.

[10] T. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practice.* Prentice Hall PTR, 2001.

[11] I. Daubechies, *Ten lectures on Wavelets.* Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1992.

[12] Y. Meyer, *Wavelets and applications.* Masson, 1991.

[13] M. Ziółko, J. Gałka, B. Ziółko, and T. Drwięga, “Perceptual wavelet decomposition for speech segmentation,” *Proceedings of the INTERSPEECH, Makuhari*, pp. 2234–2237, 2010.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book.* UK: Cambridge University Engineering Department, 2005.