Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Zakopane–Kościelisko, 1st–6th September 2011

WAVELET-FOURIER ANALYSIS FOR SPEAKER RECOGNITION

Mariusz Ziółko, Rafał Samborski, Jakub Gałka, Bartosz Ziółko

Department of Electronics,
AGH University of Science and Technology, Kraków, Poland
al. Mickiewicza 30, 30-059 Kraków
www.dsp.agh.edu.pl
{ziolko, sambo, jgalka, bziolko}@agh.edu.pl

ABSTRACT

An innovative method of speech analysis is described. The method shown in this paper is based on the composition of two transforms. The wavelet transform was calculated first and next the Fourier transform was applied. The wavelet-Fourier transform to representation and analysis of the speech signal is presented in this paper. This representation is used to find the differences between speakers. Speech signals were used to verify the efficiency of presented methods. It gives the possibility to analyse the efficiency of a speaker recognition system in the wavelet-Fourier domain.

INTRODUCTION

There are a number of dynamically altered parameters in a speech signal which make recognising a speaker possible. A speech signal should be analysed in a specific manner and the appropriate representation of speech is an important problem. The original representation in the time domain usually gives little information about the speech signal properties. To make information more noticeable it is necessary to use some transform. The choice of transform depends on the purpose of analysis. The frequency properties of a speech signal permanently change by continuous reconfiguration of human voice tract and resonant chambers.

The aim of a speaker recognition system is to provide an efficient and accurate mechanism to distinguish the individual properties of each speaker. Similarities in individual speech elements are always the base of speaker recognition system. Especially it is important if a new speaker representation is introduced. Speaker recognition requires precise analysis of the speech signal. It must be cleared out whether the representation carries suitable signal features or not. The goal of the work is to analyse, if the wavelet-Fourier transform (WFT) meets the demand of speaker recognition systems. The system should be independent of spoken text and based on the language characteristics such as accents, speaking styles and dysfluencies. Wavelet packets were already tested for speaker verification [1]. Psychoacoustucally motivated wavelet based methods were tested for speech recognition [2–4]. A method of reduction of a wavelet decomposition tree was presented [5], however, described conclusions about their solutions being optimal is questionable.

As has been noticed at the beginning, the speech signals were used to compute the wavelet-Fourier spectra. The speech signals which were used to check properties of the wavelet-Fourier transform are part of TIMIT database, however, only male voices were chosen. The method is based on a similar approach already applied to automatic speech recognition [6].

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.

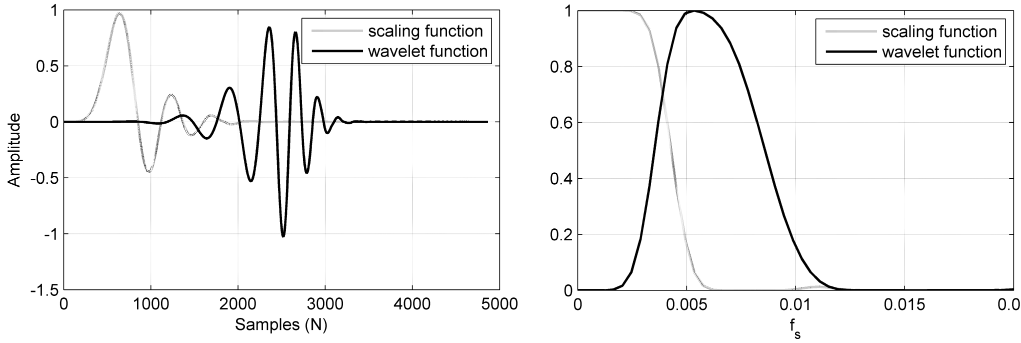

Figure 1. Daubechies 10-th order wavelet, scale function and their amplitude spectra

The discrete wavelet transformation (DWT) belongs to the group of frequency transformations and is used to obtain a time-frequency spectrum [7,9] 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 the 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 a wavelet spectrum of a 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

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

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 [7]) the scaling function $\varphi(t)$ is defined. These two functions have an 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 Discrete Wavelet Transform (DWT) for five resolution levels. The coefficients of the lower levels are calculated by applying the well-known [7] 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 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 [8]. 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 6 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

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

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)

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

is computed by applying the Fast Fourier Transform (FFT) to each level separately.

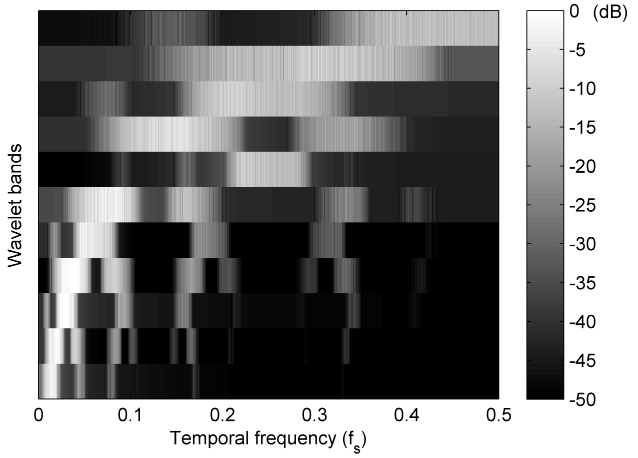

Figure 2. Example of DWFT spectra for 11 resolution levels in a [dB] scale. DWFT allows easy and detailed analysis, similarly to traditional spectrograms but are more efficient

The discrete wavelet-Fourier spectrum gives the specific and individual frequency characteristics for voices of each speaker. The sampling frequency has been set to 16 [kHz]. The discrete wavelet-Fourier spectra were calculated for eleven resolution levels. The example of DWFT spectrum is presented in Fig.2. Every row of DWFT spectrum is a Fourier spectrum of one resolution level. The wavelet decomposition is applied on the real signals, therefores the decomposed signals are also real. Hence their Fourier spectra are symmetrical to the center of spectrum and only a half of a spectrum is shown. It is clear that for resolution levels containing high frequencies (upper part of the figure), their Fourier spectra have more spectral power for the higher frequencies (right part of the figure).

SYSTEM ARCHITECTURE

Voices of each of fourteen speakers were recorded. Let us assign the number of utterances by N. WFT spectra obtaned in the way described above are easy to compare. Fig. 3 presents the architecture of the comparison system.

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 $m = 2, \ldots, 6$, where $n = 1, \ldots, M$ is the number of the speaker and $i = 1, \ldots, N$ is the number of his utterance.

The average value of normalised amplitude spectra

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

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

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

for $m = 1,\ldots,6$ frequency bands, create an individual characteristic for each speaker.

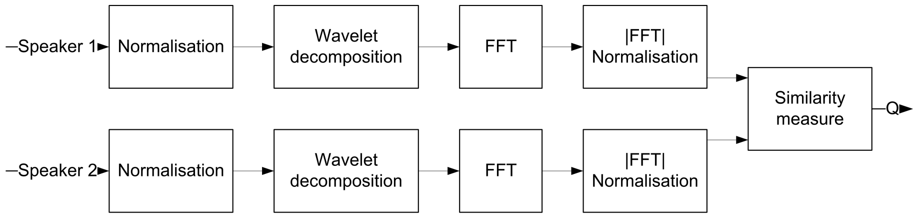

Figure 3. Architecture of the described speaker comparison system applying DWFT

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. A voice recorded for an unknown person is processed in a way described by formulae (6)-(13) to obtain its spectra $u_{m,v}$ for $m = 1,\ldots,M$, where v 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 was set to be 10 s what is equal to $N = 160\,000$ samples. The similarity measure between verified and n-th speaker is defined as l^1 metric

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

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

RESULTS

To examine the presented method two data bases were employed. First one consisted the average DWFT spectras and the second on consisted DWFT samples. These data bases were based on different sentences. The square matrix of speaker-to-speaker distances presented in Fig. 4 represents the similarity coefficients calculated as in (18). The size of the matrix depends on a number of compared speakers. In our experiment fourteen different speakers were used what results in the matrix size of $14 \cdot 14 = 196$ elements. Matrix is not symmetric because columns represent the average spectra and rows represent spectra for one utterance, only. The smaller the similarity coefficient is, the more similar is DWFT sample to average spectra. It is clear that in most cases the element on the diagonal has the smallest value in the row. Recognition quality is well-characterised by a average position of the proper speaker $\overline{p}$. In our experiments $\overline{p} = 1.3$ for 14 speakers in the data base.

CONCLUSIONS

The frequency method of a speech signal analysis is a useful tool, both in speaker, and speech recognition.

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 possible the

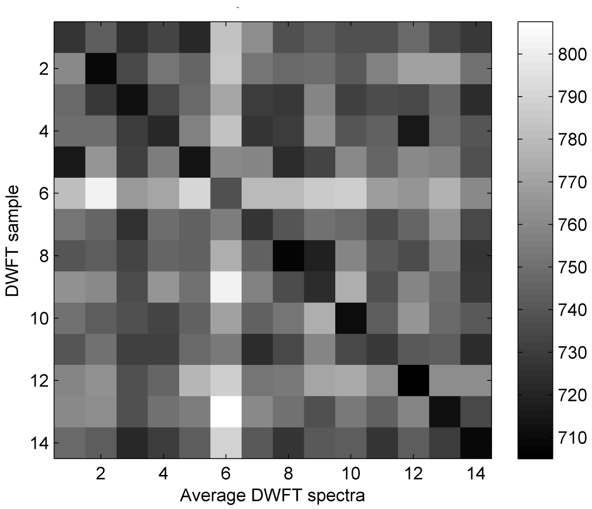

Figure 4. Matrix of speaker-to-speaker distances using measure (18). Lower is better.

detection of the specific voice signal properties. These properties have important features for a speaker recognition system.

The combination of two methods, the WFT, makes possible the detection of 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. The wavelet-Fourier analysis seems to be an effective tool to distinguish the voices.

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

ACKNOWLEDGEMENTS

This work was supported by MNISW grant number OR00001905.

REFERENCES

[1] T. Ganchev and M. Siafarikas and N. Fakotakis: *Speaker Verification Based on Wavelet Packets*, Lecture Notes in Computer Science - Text, Speech and Dialogue, Springer (2004).

[2] O. Farooq and S. Datta: *Mel Filter-like admissible wavelet packet structure for speech recognition*, IEEE Signal Processing Letters **8** (2001), 196-198.

[3] ———: *Wavelet based robust subband features for phoneme recognition*, IEEE Proceedings: Vision, Image and Signal Processing **151** (2004), 187-193.

[4] J. N. Gowdy and Z. Tufekci: *Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition*, Proceedings of ICASSP (2000).

[5] 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 **152** (2005).

[6] M. Ziółko and J. Gałka and B. Ziółko and T. Jadczyk and D. Skurzok and J. Wicijowski: *Automatic speech recognition system based on wavelet analysis*, In 2010 IEEE Fourth International Conference on Semantic Computing, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA (2010), 450-451.

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

[8] M. Ziółko and J. Gałka and B. Ziółko and T. Drwięga: *Perceptual Wavelet Decomposition for Speech Segmentation*, Proceedings of the INTERSPEECH, Makuhari (2010), 2234-2237.

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