Speech Recognition using Wavelet Packet Features

Mihalis Siafarikas, Iosif Mporas, Todor Ganchev, Nikos Fakotakis

Wire Communications Laboratory,

Department of Electrical and Computer Engineering, University of Patras

26500 Rion-Patras, Greece

{msiaf,imporas,tganchev,fakotaki}@wcl.ee.upatras.gr

Abstract

In view of the growing use of automatic speech recognition in the modern society, we study various alternative representations of the speech signal that have the potential to contribute to the improvement of the recognition performance. Specifically, the main targets of the present article are to overview and evaluate the practical importance of some recently proposed, and thus less studied, wavelet packet-based speech parameterization methods on the speech recognition task, illustrating their merits compared to other well known approaches. To this end, working on the widely acknowledged TIMIT speech database and relying on the Sphinx-III speech recognizer, we contrast the performance of four wavelet packet-based speech parameterizations against traditional Fourier-based techniques that have been considered for the task of speech recognition for over two decades, including MFCC and PLP cepstral coefficients that presently dominate the speech recognition field. The experimental results demonstrate that the wavelet packet-based speech features of interest provide a superior performance over the baseline parameters. This validates the wavelet packet-based speech parameterization schemes as a promising research direction that could bring further speech recognition error rate reduction.

Index Terms: speech parameterization, speech recognition, wavelet packet transform

1. Introduction

Automatic speech recognition (ASR) aims at converting spoken language to text. At present, speech recognizers are based predominantly on statistical analysis of speech, which involves training of multiple sub-word models by utilizing large speech corpora. The state-of-the-art technology in the field of ASR is based on the combination of efficient methods for pattern recognition, such as the Hidden

Markov Models (HMM) [1, 2], with techniques for dynamic programming, such as the Viterbi algorithm [3, 4]. This combination allows for the efficient processing of time-series, where each event is represented by a set of speech feature vectors which, having the task to present the information carried by the speech signal in a compact form so that it can be efficiently utilized by the HMM classifier. Although presently the ASR technology is sufficiently mature and a large number of commercial applications have been launched, one problem that has not been solved yet with adequate elegance is the speech parameterization process despite the fact that a comprehensive knowledge of the speech features that should be retained or suppressed has already been developed. Furthermore, thorough research has been carried out in the processing of the speech features for the purpose of eliminating the signal variability due to several factors, such as adverse environmental conditions [5], differences in the vocal tract among speakers [6-8], etc. However, accurate speech recognition remains a difficult task to achieve often due to a number of other practical issues that have to solved [9]. Furthermore, an optimal and generally accepted solution for the construction of speech features, especially designed for speech recognition in real world conditions, has yet to be found.

Historically, the following speech features have dominated the speech recognition area: Real Cepstral Coefficients (RCC) introduced by Oppenheim [10], Linear Prediction Coefficients (LPC) proposed by Atal and Hanauer [11], Linear Predictive Cepstral Coefficients (LPCC) derived by Atal [12], Mel Frequency Cepstral Coefficients (MFCC) of Davis and Mermelstein [13] and Perceptual Linear Predictive (PLP) parameters of Hermansky [14]. In [13], it was demonstrated that the psychophysically inspired MFCC outperform LPC, LPCC, and other features, on the task of speech recognition. From a perceptual point of view, MFCC roughly resemble the human auditory system, since they account for the nonlinear nature of pitch perception, as well as for the nonlinear relation between intensity and loudness. That makes MFCC more adequate features for speech recognition than other formerly used speech parameters like RCC, LPC, and LPCC, further reinforced by their robust and cost-effective computation. As a result, speech recognition tasks are heavily dependent on these beneficial attributes of MFCC features whenever the selection of speech features is brought into question. MFCC features have been reported to be slightly outperformed by other features such as PLP only in specific conditions [15]. As concerns various other speech features, such as the perceptual linear prediction Adaptive Component Weighting (ACW) cepstral coefficients [16] and several wavelet-based features such as the Subband Based Cepstral (SBC) [17], Wavelet Packet Features

(WPF) [18], Wavelet Packet parameters for Speaker Recognition (WPSR) [19], Overlapping Wavelet Packet Features (OWPF) [20], etc, although presenting reasonable solution for the same tasks, did not gain widespread practical use, often due to their relatively more complicated computation.

However, present processing performance of computers has increased dramatically so that any restrictions due to their computational limitations can be easily circumvented. In view of the growing utilization of ASR technology and the emerge of multiple commercial services, this progress might be utilized for re-examination of traditional techniques in the extraction of speech features for a specific application. For instance, one typical set-up is the server-based speech interaction, where the speech recognizer is deployed on a remote voice server located at the service provider premises. In such applications, the user satisfaction is the foremost criterion for estimating the worth of the voice interaction system, and this raises the need of squeezing the maximum performance out of the ASR component, leaving aside issues as computational demands, memory requirements, power efficiency, etc. Various approaches for improving the performance of the speech recognition component were proposed in the literature but here we consider only issues related to the choice of speech parameterization technique.

In the present work, we employ the Sphinx-III speech recognizer [21] and the TIMIT (Texas Instruments and Massachusetts Institute of Technology) speech database [22] to evaluate a number of recent wavelet packet-based speech parameterization techniques [17-20] against traditional Fourier-based approaches [14, 23] in a common experimental setup. Eventually, we identify the relative ranking of the evaluated speech features, we perform a statistical test to estimate the significance of the differences in their performance as compared to MFCC and PLP features [23] and we attempt a theoretical investigation for the superior performance of wavelet packet-based features.

The remaining of this article is organized as follows: In Section 2, we provide description of the wavelet packet-based speech parameterization techniques of interest. Section 3 outlines the traditional MFCC and PLP speech parameterization techniques. Section 4 describes the experimental setup along with the training and testing procedures used in the present study. Section 5 presents comparative results for the speech features extraction methods used in the present work. This article concludes with Section 6, which offers brief summary and conclusions.

2. Wavelet packet-based speech parameterization

Before discussing in depth the wavelet packet-based speech parameterization techniques, we offer, in subsection 2.1, a brief overview of wavelet packet (WP) analysis focusing on the Discrete Wavelet Packet Transform (DWPT) that deals with discrete-time signals. A detailed presentation of signal analysis with wavelets can be found in [24]. Moreover, in the same subsection, an outline of a specific generalization of DWPT to Overlapping discrete Wavelet Packet Transform (OWPT), initially described in [20], is shortly presented.

2.1 Wavelet packet analysis

Historically, wavelet analysis begins with continuous wavelet transform (CWT). It provides a time-scale representation of a continuous function where scale plays a role analogous to frequency in the analysis with the well-known Fourier Transform (FT). More precisely, wavelet analysis uses dilations of a single function, called wavelet, to analyze a signal with different scales or resolutions.

The basic tool for the practical analysis of discrete-time signals via wavelets is the discrete wavelet transform (DWT). DWT bears a relation to the CWT analogous to the relation that the Discrete Fourier Transform (DFT) bears to the FT. DWT is orthonormal, and thus can be regarded as a sub-sampling of the two dimensional CWT on dyadic scales $s_j = 2^j, j \in \mathbb{N}$ and on selected times $t_j = k \cdot 2^j, k \in \mathbb{Z}$ in a given dyadic scale s_j. In this way, a one dimensional time-scale representation of a signal is obtained, in contrast to the DFT that provides solely a frequency representation of the signal.

Discrete Wavelet Packet Transform (DWPT) is a generalization of the DWT that allows an effective representation of the time-frequency properties of a discrete-time signal so that useful features for a particular purpose can be appropriately extracted. If $x[n], n = 0,\ldots,N-1$, where N is an integer multiple of 2^J for some positive integer J, denotes a real valued discrete-time signal, then for $0 \leq j \leq J$, the jth level DWPT of $x[n]$ is an orthonormal transform yielding an N dimensional vector of coefficients that can be partitioned as $\left[\mathbf{W}_j^{2^j-1} \quad \mathbf{W}_j^{2^j-2} \quad \cdots \quad \mathbf{W}_j^1 \quad \mathbf{W}_j^0\right]^T$, where $\mathbf{W}_j^n$ is a $N/2^j$ dimensional vector, each element of which is nominally associated with adjacent time intervals of width 2^j and frequency interval $I_j^n = \left[\frac{n}{2^{j+1}}, \frac{n+1}{2^{j+1}}\right]$. These 2^j vectors divide the Nyquist frequency interval $[0,1/2]$ into 2^j intervals of equal width (so the bandwidth associated with each jth level DWPT coefficient is $1/2^{j+1}$) and each one of its $N/2^j$ elements provides information associated with the time interval $\left[k \cdot 2^j, (k+1) \cdot 2^j\right], k = 0,\ldots N/2^j - 1$. Thus, the DWPT provides localized time-

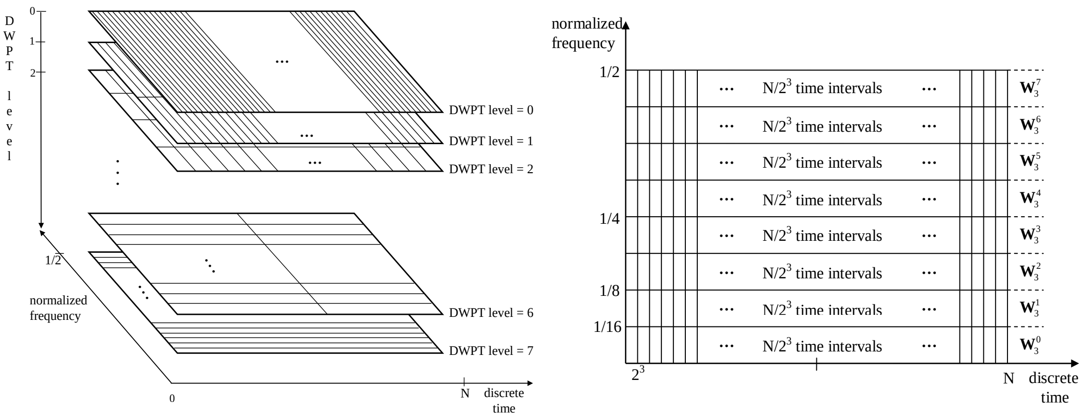

Figure 1. Three-dimensional illustration of time-frequency analysis achieved with DWPT (left hand side) and two dimensional illustration of DWPT at a specific example level, *j*=3 (right hand side).

frequency description of a signal while each level of DWPT provides homogeneous frequency and time analysis, as it is visualized in Figure 1, in contrast to DWT that provides an octave based decomposition. The most appealing fact, though, of DWPT is that coefficient vectors $\mathbf{W}_j^n$ from different levels can be grouped together to form a set $S = \left\{\mathbf{W}_j^n : j = 0,\dots J, n = 0,\dots 2^j - 1\right\}$, where each $\mathbf{W}_j^n \in S$ is nominally associated with the frequency band I_j^n. Any subset $S_1 \subset S$ that provides a non-overlapping complete coverage of $[0,1/2]$ with coefficient vectors $\mathbf{W}_j^n$ yields an orthonormal DWPT. In this way, DWPT can provide a flexible tiling of the time-frequency plane with various frequency resolutions in the corresponding frequency intervals.

In [20], Siafarikas et al. introduced a generalization of the DWPT, referred to as Overlapping discrete Wavelet Packet Transform (OWPT). They grouped together carefully selected basis vectors belonging to different levels of the DWPT in such a way that multiple wavelet packet resolutions are utilized for specific frequency bands of interest. This led to the construction of an even larger collection of wavelet packet transforms which are not necessarily orthonormal. The main difference with DWPT is that the coefficient vectors W_j^n that constitute the OWPT form a subset $S_2 \subset S$, $S = \left\{W_j^n : j = 0,\dots J, n = 0,\dots 2^j - 1\right\}$, which provides a complete coverage of the interval $[0,1/2]$ with overlapping frequency bands I_j^n. In this way, OWPT provides a flexible tiling of the time-frequency plane with various frequency resolutions in the corresponding frequency intervals along with emphasis in specific frequency subbands.

2.2 Wavelet packet-based speech parameterization schemes

In the present subsection, we outline four wavelet packet-based speech parameterization schemes that differ mainly in the WP time-frequency analysis and the wavelet function employed in the DWPT analysis. The first two schemes ([17], [18]) are well-known references in the area of speech parameterization with wavelet packets, reporting successful results for the speech recognition task. The last two schemes ([19], [20]) have further elaborated on the same subject introducing new concepts such as approximation of the critical bands with wavelet packets and overlapping wavelet packets, reporting better results than both previous wavelet packet based approaches as well as traditional DFT based techniques. To the best of our knowledge, the aforementioned four wavelet packet based speech parameterization methods are amongst the most successful state-of-the-art ones and that urged us to this specific selection of wavelet packet based speech parameterization schemes.

For the purpose of fair comparison we adapted these schemes to a common experimental setup, by unifying the frequency range and the signal pre-processing steps. Specifically, we accepted the frequency range [133, 6855] Hz of the MFCC [23] as binding for all other speech parameterization methods under consideration in this study. Details about the signal pre-processing are provided in Section 4, and thus in the following subsections we focus only on the speech features extraction steps.

2.2.1 Sarikaya & Hansen's SBC features

Working on the stressed speech monophone recognition problem, Sarikaya & Hansen [17] considered analysis of the speech signal in the frequency interval [0, 4] kHz with wavelet packets such that the 24 frequency subbands closely approximate the Mel scale. Specifically, in [17] the authors worked with speech signal sampled at 8 kHz and relied on a frame size of 24 milliseconds with a 10 milliseconds skip rate. The choice of frame size is in correspondence with the requirement that the total number of samples in the frame should be divisible by 64, while keeping comparable frame sizes for all parameters under consideration. After framing, the speech signal was windowed with Hamming window and next pre-emphasized. After some experimentation, the authors found a specific wavelet packet time-frequency analysis (Figure 2(a)) that led to an optimal performance amongst a large number of similar wavelet packet analyses.

By incorporating the highest resolution of 62.5 Hz in the lower part of the frequency scale, their approach puts special emphasis on the lower frequency part [0, 500] Hz, which normally contains large portion of the signal energy. As it can be seen in Figure 2(a), the partition of the next frequency range

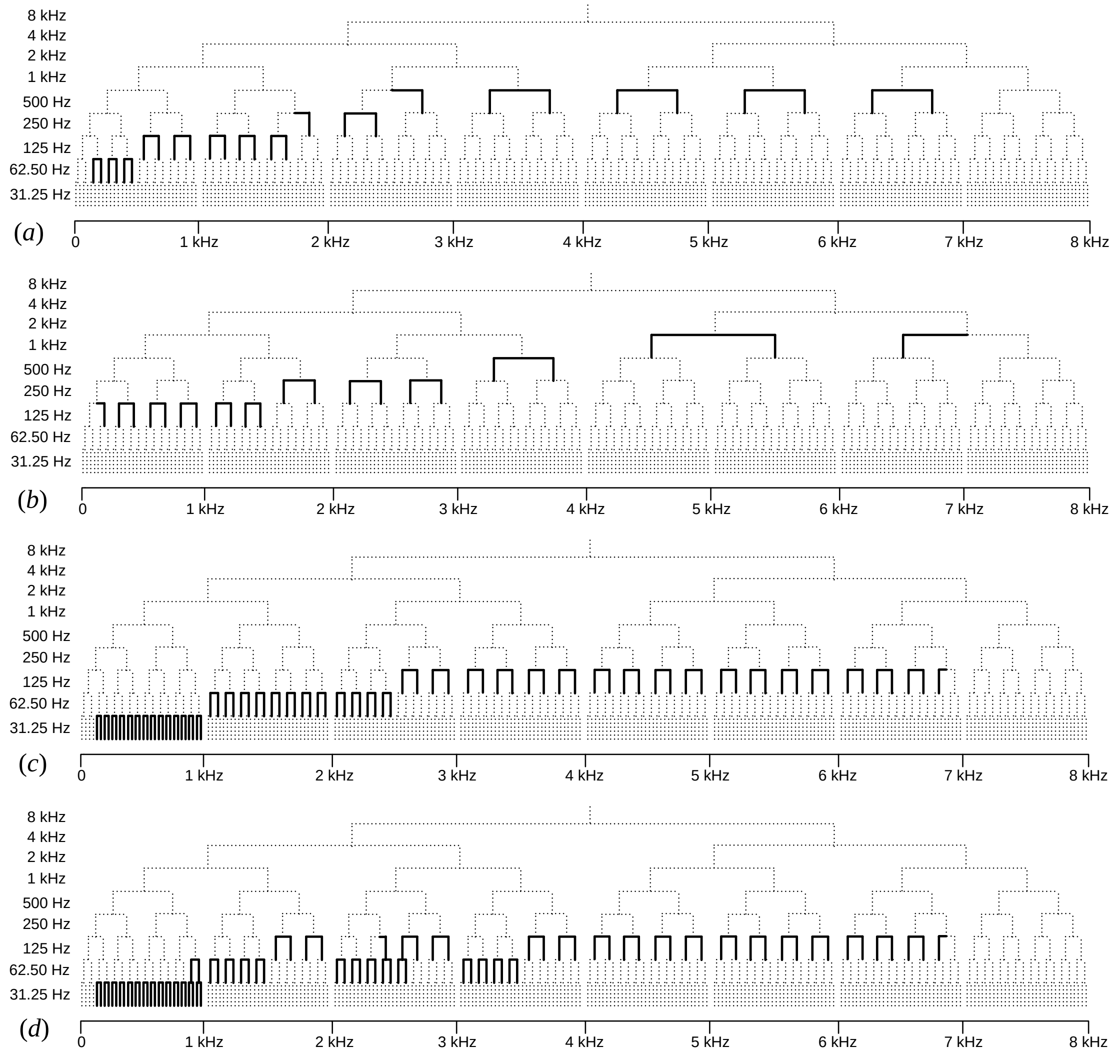

Figure 2. Time-frequency analyses for the following wavelet packet-based features: (*a*) SBC (*b*) WPF (*c*) WPSR and (*d*) OWPT.

[500, 1750] Hz is in subbands of 125 Hz. The upper frequency range approximates the division defined by the Mel-scale. In brief, the wavelet packet analysis designed in this manner assigns more subbands in the lower part of the spectrum and progressively fewer subbands as we ascend the frequency axis.

The DWPT for the proposed analysis results in a sequence of subband signals or equivalently a sequence of wavelet packet transform coefficients. For each frame, the wavelet packet decomposition is followed by the computation of the energy in each subband i and the scaling by the number of transform coefficients N_i in that subband:

$$E_i = \frac{\sum_{m\in i}\left[\left(W_\psi x\right)(i,m)\right]^2}{N_i}, \tag{2}$$

where $W_\psi x(i,m)$ is the wavelet packet transformed signal x evaluated at frequency subband i and time m. For the wavelet packet transform, Daubechies' wavelet of order 32 was used (Figure 3(*b*)).

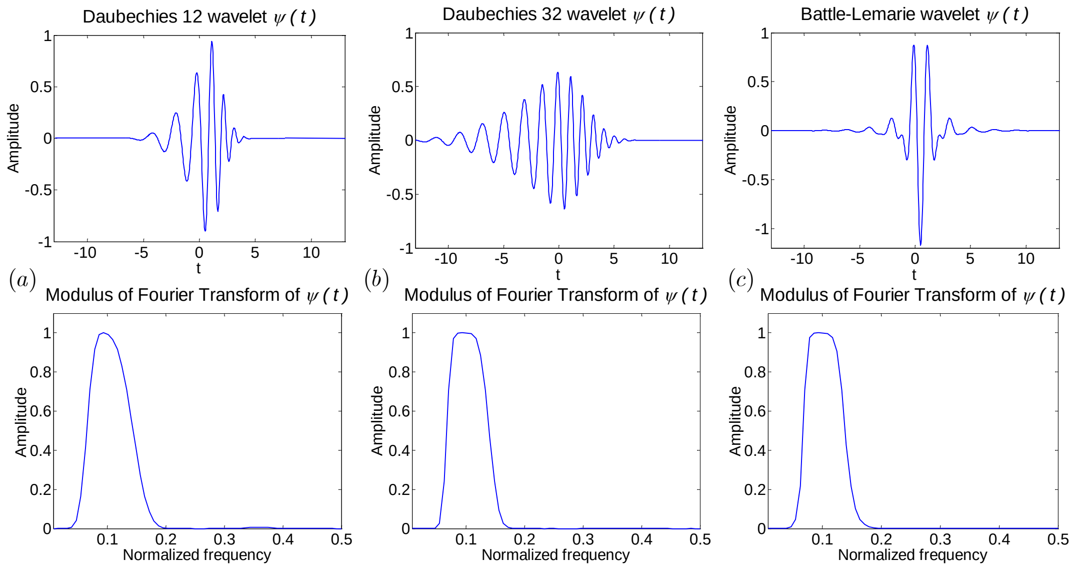

Figure 3. Wavelet functions and modulus of the Fourier transform for: (a) Daubechies 12, (b) Daubechies 32 and (c) Battle-Lemarié of order 5.

Then, the Discrete Cosine Transform (DCT) was applied on the logarithmically compressed subband energies, resulting in the so called Subband Based Cepstral (SBC) coefficients:

$$SBC(j) = \sum_{i=1}^{M} \log E_i \cos\left(\frac{j(i-0.5)}{M}\pi\right)^2, \text{ for } \quad j = 1,\ldots,J, \tag{3}$$

where J is the number of subband speech parameters and M is the total number of frequency bands. In detail, the DCT serves as a decorrelation stage, i.e. significantly reduces the correlation between the cepstral coefficients, and thus permits the use of a diagonal covariance matrix for the density estimation in the HMM states. This facilitates the training of the HMMs since fewer free variables in their covariance matrices need to be estimated from the training data.

In the present work, for the purpose of adjusting the filter-bank of SBC parameters to the desired frequency range, the following two modifications were performed: (i) The first two subbands of bandwidth 62.5 Hz each were omitted and (ii) six new subbands of bandwidth 500 Hz each were appended at the end of the original frequency range [0, 4] KHz. The above modifications resulted in 28 frequency subbands covering the desired frequency range of [125, 7000] Hz while preserving the Mel-scale like frequency warping. This construction is the closest feasible implementation of the desired bandwidth.

2.2.2 Farooq & Datta's WPF features

Farooq and Datta, in [18], analyzed the speech signal with wavelet packets focusing on phoneme recognition. The speech in the TIMIT database, sampled at 16 kHz, was considered. Firstly, the authors carried out a full three level WP decomposition, which partitioned the frequency axis into eight bands of 1 kHz each. The lowest band of [0, 1] kHz was further decomposed by applying again full three level WP decomposition likewise dividing the [0, 1] kHz band into eight subbands each of bandwidth 125 Hz, which is close to 100 Hz bandwidth for the corresponding Mel filters. The frequency band [1, 2] kHz was further decomposed by applying two level WP decomposition, thereby, giving four subbands of 250 Hz. The frequency bands of [1, 1.25] and [1.25, 1.5] kHz were further decomposed once, thus increasing the number of bands to six in the [1, 2] kHz range. Next, the [2, 3] kHz band was further decomposed using full two level WP decomposition providing four bands each of 250 Hz bandwidth. The band [3, 4] kHz was decomposed once by WP, giving two bands of [3, 3.5] and [3.5, 4] kHz, while the frequency band of [4, 8] kHz was not further decomposed. The above analysis gave a total of 24 frequency bands (Figure 2(b)). In the present work, for the purpose of adjusting the filter-bank of WPF parameters to the desired frequency range, the first subband of bandwidth 125 Hz and the last subband of bandwidth 1 KHz were omitted. The above modification resulted in 22 frequency subbands covering the desired frequency range of [125, 7000] Hz. This construction is the closest feasible implementation of the desired bandwidth.

Farooq and Datta applied wavelet packet transform for the analysis of a 32 milliseconds long phoneme utilizing Daubechies' wavelet function of order 12 (Figure 3(a)) in order to obtain features with emphasis on the lower frequency subbands. Next, the energy E_p in each subband p was calculated as follows:

$$E_p = \sum_{j=1}^{N_p} \left(C_{j,p}\right)^2, \quad p = 1,\ldots,M , \tag{4}$$

where $C_{j,p}$ is the j th coefficient of the wavelet packet transformed signal in the p th subband, N_p is the number of wavelet packet coefficients in the same subband and M is the number of subbands. Then, the normalized subband energies:

$$F_p = E_p / N_p, \quad p = 1,\ldots,M , \tag{5}$$

were further decorrelated with the application of DCT after they had first been logarithmically compressed. As speech features, denoted as WPF F&D, the authors selected the first 13 coefficients of the resultant decorrelated energy vector.

2.2.3 Siafarikas' et al. WPSR features

As their name suggest, Wavelet Packet features for Speaker Recognition (WPSR) of Siafarikas et al.

[19] were developed for the task of speaker recognition but, in this work, they are adapted to the speech recognition task. The authors performed wavelet packet decomposition of the speech signal with Battle-Lemarié polynomial spline wavelet of order 5 (Figure 3(c)), which they found to be the best choice as a wavelet basis function among a reasonable number of evaluated wavelet functions.

Their WPSR features relied on the psychoacoustic effect of the critical bands, introduced by Fletcher in [25], in contrast to the SBC and WPF F&D speech features which were constructed following approximately the Mel scale. The resulting WP analysis vector set, referred to as S_1, is shown in Figure 2(c). The original WPSR features used 68 filters to efficiently cover the frequency range [0, 4000] Hz, as follows: resolution 31.25 Hz for the range [0, 1000] Hz, corresponding to 32 subbands; resolution 62.5 Hz for [1000, 2500] Hz, 24 subbands; resolution 125 Hz, for [2500, 4000] Hz, 12 subbands. Here, the desired frequency range is implemented by discarding the first four subbands and adding 23 new subbands with resolution 125 Hz. This led to a total of 87 subbands in the range [125, 6875] Hz.

Next, the energy in each frequency band is computed, and then divided by the total number of coefficients present in that particular band. In detail, the subband signal energies are computed for each frame as:

$$E_p = \frac{\sum_{i=1}^{N/2^j} \left(\mathbf{W}_j^k(i)\right)^2}{N/2^j}, \quad \mathbf{W}_j^k \in S_1, \quad p = 1,\ldots,B, \tag{6}$$

where $W_j^k(i)$ is the i th coefficient of the DWPT vector W_j^k belonging to WP analysis scheme S_1.

Finally, a logarithmic compression is performed and a DCT is applied on the logarithmic subband energies in order to obtain decorrelated coefficients:

$$F(i) = \sum\nolimits_{p=1}^{B} \log_{10} E_p \cos\left(\frac{i(p - 1/2)}{B}\right), i = 1,\ldots,r \tag{7}$$

where r is the number of feature parameters. The full dimensionality of the feature vector $F(i)$ is $r = 87$. However, since most of the energy of the feature vector is carried by the first few coefficients, in many applications the first coefficients alone might be sufficient as signal descriptors.

2.2.4 Siafarikas' et al. OWPF features

In [20], elaborating on the OWPT, the authors introduced the OWPF. As presented in [20], there exists a huge number of overlapping transforms that can be selected among. During the design of the OWPF, the authors focused their search in specific areas of the frequency axis that concentrate the positions of the formants. These areas were also found relatively more important for distinguishing of different voices, and for this reason, they ought to have a stronger contribution in the speech feature vector.

In [20], it was found that the OWPT that provided the best performance covered the frequency bands (in kHz) [0, 1], [0.875, 1.5], [1.5, 2], [2, 2.625], [2.375, 3], [3, 3.5] and [3.5, 4] with resolutions (in Hz) 31.25, 62.5, 125, 62.5, 125, 62.5 and 125, respectively. Evidently, the areas of overlapping that provided the best speaker verification results were the frequency range [0.875, 1] kHz covered with resolutions 31.25 Hz and 62.5 Hz and the frequency range [2.375, 2.625] kHz covered with resolutions 62.5 Hz and 125 Hz. The resultant WP analysis scheme is shown in Figure 2(d). Here, to adjust the frequency range to the desired one, we discarded the first four subbands, which do not contribute much to the speech content, and then added 23 bands with resolution 125 Hz in the interval [4, 6.785]. Thus, a total of eighty-eight frequency bands, which cover the frequency range [125, 6785] Hz, were obtained. Utilizing the results from an earlier research, the Battle-Lemarié of order 5 (Figure 3(c)) wavelet function was employed for the computation of the OWPFs.

The normalized energy in each frequency band is computed as:

$$E_j = \frac{\sum_{i=1}^{N/2^j + \left|M_j^n\right|} \left(W_j^p(i)\right)^2}{N/2^j + \left|M_j^n\right|}, \quad \mathbf{W}_j^k \in S_2, \quad j = 1, \ldots, B \tag{8}$$

where $W_j^p(i)$ is the i th coefficient of the wavelet packet vector W_j^p belonging to vector set S_2 of the OWPT, and B is the total number of frequency bands. M_j^n are integers that satisfy the following equation:

$$\sum_{j=0}^{J} \sum_{n=0}^{2^j-1} \left|M_j^n\right| = M \tag{9}$$

where M is a non-negative integer representing the redundancy factor of the OWPT.

Finally, logarithmic compression is performed and DCT is applied on the logarithmic subband energies to decorrelate the parameters:

$$F(i) = \sum_{n=1}^{B} \log_{10} E_n \cos\left(\frac{i(n - 1/2)}{B}\right), i = 1, \ldots, r\,, \tag{10}$$

where r is the number of parameters in the feature vector.

3. DFT-based Speech Parameterization Techniques

Among the numerous DFT-based speech parameterization methods proposed in the literature [10-16, 23, 26-28], we will consider only two well known schemes, namely the MFCC-FB40 [23] and the PLP-FB19 [14], which at present are the typical choice for speech recognition.

3.1 MFCC speech features (MFCC-FB40)

Over the last few decades, MFCC speech features have proved a most effective parameterization scheme for the task of speech recognition resulting in various implementations by a number of authors [13, 23, 27, 28, etc]. In this work, for purposes of comparative evaluation of the speech parameterization methods reported in Section 5, we depend on the implementation proposed in [23] due both to its reliability as well as the fact that Sphinx-III speech recognizer utilizes that as its default speech parameterization scheme.

In brief, assuming sampling frequency 16 kHz, the author in [23] implemented a filter bank of 40 equal area filters, which cover the frequency range [133, 6855] Hz. The centre frequencies of the first 13 filters are linearly spaced in the range [200, 1000] Hz with a step of 66.67 Hz and the ones of the next 27 are logarithmically spaced in the range [1071, 6400] Hz with a step $s = 1.0711703$, computed as:

$$s = \exp\left(\frac{\ln\left(\frac{f_{c_{40}}}{1000}\right)}{N_{\mathrm{logfilt}}}\right). \tag{11}$$

Here $f_{c_{40}} = 6400\ \mathrm{Hz}$ is the centre frequency of the last of the logarithmically spaced filters, and $N_{\mathrm{logfilt}} = 27$ is the number of logarithmically spaced filters. Each one of these equal-area triangular filters is defined as:

$$H_i(k) = \begin{cases} 0 & \text{for } k < f_{b_{i-1}} \\ \dfrac{2\left(k - f_{b_{i-1}}\right)}{\left(f_{b_i} - f_{b_{i-1}}\right)\left(f_{b_{i+1}} - f_{b_{i-1}}\right)} & \text{for } f_{b_{i-1}} \le k \le f_{b_i} \\ \dfrac{2\left(f_{b_{i+1}} - k\right)}{\left(f_{b_{i+1}} - f_{b_i}\right)\left(f_{b_{i+1}} - f_{b_{i-1}}\right)} & \text{for } f_{b_i} \le k \le f_{b_{i+1}} \\ 0 & \text{for } k > f_{b_{i+1}} \end{cases}, \tag{12}$$

where $i = 1,\dots,M$ stands for the i th filter, f_{b_i} are $M+2$ boundary points that specify the M filters, and $k = 1,\dots,N$ corresponds to the k th coefficient of the N- point DFT:

$$X(k)=\sum_{n=0}^{N-1}x(n)\cdot\exp\left(\frac{-j2\pi nk}{N}\right),\quad k=0,\ldots,N-1, \tag{13}$$

computed for the discrete input signal $x(n)$. The boundary points f_{b_i} of the filters in (12) are expressed in terms of position, as specified by:

$$f_{b_i}=\left(\frac{N}{F_s}\right)\cdot\hat{f}_{mel}^{-1}\left(\hat{f}_{mel}\left(f_{low}\right)+i\cdot\frac{\hat{f}_{mel}\left(f_{high}\right)-\hat{f}_{mel}\left(f_{low}\right)}{M+1}\right), \tag{14}$$

where f_{low} and f_{high} are respectively the low and high boundary frequencies for the entire filter bank, M is the number of filters. Here, the function $\hat{f}_{mel}(\cdot)$ states the transformation:

$$\hat{f}_{mel}=1127\cdot\ln\left(1+\frac{f_{lin}}{700}\right), \tag{15}$$

and $\hat{f}_{mel}^{-1}$ is the inverse of transformation (15), formulated as:

$$\hat{f}_{mel}^{-1}=f_{lin}=700\cdot\left[\exp\left(\frac{\hat{f}_{mel}}{1127}\right)-1\right]. \tag{16}$$

The key to equalization of the area below the filters in the filter-bank (12) lies in the term:

$$\frac{2}{\left(f_{b_{i+1}}-f_{b_{i-1}}\right)}. \tag{17}$$

Due to the term (17), the filter bank (12) is normalized in such a way that the sum of coefficients for every filter equals one. Thus, the i th filter satisfies:

$$\sum\nolimits_{k=1}^{N}H_i(k)=1,\ \text{for}\ \ i=1,\ldots,M\ . \tag{18}$$

Next, the equal area filter bank (12) is employed in the computation of the log-energy output:

$$X_i=\log_{10}\left(\sum_{k=0}^{N-1}\left|X(k)\right|\cdot H_i(k)\right),\qquad i=1,\ldots,M\ . \tag{19}$$

Finally, the DCT (20) provides the MFCC-FB40 parameters:

$$C_j=\sum_{i=1}^{M}X_i\cdot\cos\left(j\cdot(i-1/2)\cdot\frac{\pi}{M}\right),\ \text{with}\ \ j=0,\ldots,J-1, \tag{20}$$

where j is the serial index of the cepstral coefficient, and J is the number of MFCC that are needed. Obviously $J\leq M$, and for speech recognition a widely accepted value is $J=13$. In many real world applications the 0 th cepstral coefficients is excluded from the feature vector for reducing the dependence on the environmental variability, but for generality we utilize all J parameters.

3.2 PLP speech features (PLP-FB19)

For the computation of the PLP parameters [14], the discrete time domain input signal $x(n)$ is initially transformed by the N- point DFT (13). Subsequently, for the computation of the critical-band power spectrum, the power spectrum is convolved with the piece-wise approximation of the critical-band curve [14, 29], as follows:

$$\theta(B_i) = \sum_{B=-1.3}^{2.5} \left|X(B - B_i)\right|^2 \Psi(B) \qquad (21)$$

where

$$\Psi(B) = \begin{cases} 0 & \text{for} \quad B < -1.3 \\ 10^{2.5(B+0.5)} & \text{for} \quad -1.3 \le B \le -0.5 \\ 1 & \text{for} \quad -0.5 \le B \le +0.5 \\ 10^{-1(B-0.5)} & \text{for} \quad +0.5 \le B \le +2.5 \\ 0 & \text{for} \quad +2.5 < B \end{cases} \qquad (22)$$

In the above relation, B is the Bark warped frequency obtained through the Hertz-to-Bark conversion:

$$B(f) = 6\ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right), \quad f = 0,...,Fs/2\,. \qquad (23)$$

Equal loudness pre-emphasis is applied on the down-sampled $\theta(B)$ and then intensity-loudness compression is performed. To the result obtained so far, an inverse DFT is performed to obtain the equivalent autocorrelation function. Finally, the PLP coefficients are computed after autoregressive modelling and conversion of the autoregressive coefficients to cepstral coefficients.

For the needs of the present work, we modified the filter bank by omitting the first (lowest frequency) filter and all filters whose centre frequencies reside beyond 6855 Hz. This adaptation provided us with a coverage of the frequency range [100, 6400] Hz with a filter-bank of 19 filters. The current implementation is the closest feasible coverage of the desired frequency range.

4. Experimental Setup

The above described speech parameterization techniques were evaluated on the TIMIT database [22]. TIMIT is a widely-used database in the area of speech processing. The evaluation on it gives the ability for further comparison with other parameterization techniques that are not included in the present work. For the speech recognition task, the Sphinx-III system [21, 30] was utilized. Sphinx III is an open-source speech recognizer, based on the HMMs. It functions in training (learning) and decoding (recognition) mode. Further details about the speech corpus and the setup of the speech recognizer are given below.

4.1 Speech corpora and experimental protocol

TIMIT database contains recordings of the eight major dialect regions of the United States of America. The total number of speakers is 630. The audio material consists of single-channel 16-bit microphone-quality speech recordings, with sampling rate 16 kHz. The database has been subdivided into portions for training and testing. The test subset contains about 27 % of the total speech material, while the rest is provided for training. This standard division of training and test data was followed at the present work.

The pronunciation of the words included in the transcriptions of the database was carried out with the utilization of the lexicon provided with the database. The lexicon offers phonetic writing of 6229 American-English words. We used a phoneme set of 38 symbols {*aa, ae, ah, ahr, aw, ay, b, ch, d, dh, eh, er, ey, f, g, hh, ih, iy, jh, k, l, m, n, ng, ow, oy, p, r, s, sh, t, th, uh, uw, v, w, y, z*}.

4.2 Training of the acoustic models

For each of the above described wavelet packet-based and Fourier-based parameterization techniques we trained one acoustic model, following a two-stage procedure. Initially, an acoustic model was trained using the speech waveforms and their corresponding transcriptions. This acoustic model was force-aligned against the transcription of the training data, and the pronunciations of the words with multiple phonetic representations were extracted. After that, the force-aligned transcriptions were further utilized to train new acoustic models following the same procedure. In particular, 13-dimensional feature vectors were used. We used pre-emphasis filter with factor equal to 0.97 for the speech parameterization methods that perform the pre-emphasis in time domain. Subsequently, the speech waveforms were segmented to frames of length of 16 or 25.625 milliseconds with step of length of 10 milliseconds. In addition, for the MFCC and PLP features, each frame was passed through a Hamming window, while for the DWPT-based schemes this is not needed due to their compact representation. The first and second derivative coefficients were tagged on the base 13, which resulted to feature vectors of dimension equal to 39.

The 3-state Bakis-topology HMMs [31] with a non-emitting terminating state was employed here. For each of the 38 monophones one HMM model was constructed plus one model for the silence. C*ontext-independent* (CI) phone models were trained through the Baum-Welch algorithm. The training process was terminated when the convergence ratio between two successive iterations became less than a predefined threshold. Next, *context-dependent* (CD) untied triphone models were trained for every triphone that had occurred at least 8 times in the training data. The CI model parameters were used to initialize the values of the parameters of the CD models. After initialization the CD models were further re-trained through the Baum-Welch algorithm. The training procedure was identical to the one used in the CI step. Next, decision trees were built. The decision trees determine which HMM states of all untied models are similar in order to be merged to common states or senones. In total 1000 senones were trained as empirical rules indicated [30]. At the next step, the decision trees were pruned, so as the number of leaves to reduce to the predefined number of tied states, without including the CI states. At last, the CD tied models were constructed. Every state of all the HMMs was modelled by a mixture of 8 continuous Gaussian distributions.

4.3 Decoding

The recognition was carried out by the Sphinx-III decoder. During recognition, every speech waveform of the test subset was parameterized identically to the training ones. Throughout the recognition process the most probable sequence of words is considered as the recognized one. This result comes out of the product of two factors, namely the acoustic score that the HMM models provide and the probability of the existence of the sequence of words called language weight. The acoustic score is estimated using the trained acoustic models. The language score is computed by a language model. Here, a 3-gram language model was used, constructed from the CMU Language Model Toolkit [32]. The training corpus of the language model included all the transcriptions of the database. Details about the setup of the decoder's parameters can be found in Table 1. Parameters that are not specified keep their default values as specified in the Sphinx-III documentation [30].

Table 1. Parameter setup of the Sphinx-III decoder

Decoder's Parameter	**Value**
Beam selecting HMMs at each frame	1.0e-55
Beam selecting word-final HMMs at each frame	1.0e-55
Beam selecting HMMs transitioning to successors at each frame	1.0e-55
Max # of histories to maintain at each frame	120
Max # of active HMMs to maintain at each frame	30000
Max # of distinct word exists to maintain at each frame	25
Silence word probability	1.0
Two alternative language weights	9.5 or 12.0

5. Experimental Results

For the parameterization techniques under consideration, the speech recognition performance was examined following the experimental setup described in Section 4. The Word Error Rate (WER) in percentages for frame length 25.625 milliseconds is presented in Figure 4, where the left bars correspond to language weight 9.5, and the right bars to language weight 12. In the same figure, the number 16 in brackets next to the denotations for the DWPT-based speech features serve to remind the fact that these speech parameterization techniques utilize only the first 16 milliseconds of each speech frame (i.e. only the first 256 speech samples). This is to respect the requirement of the DWPT analysis, which imposes on having the number of speech samples an exact power of two.

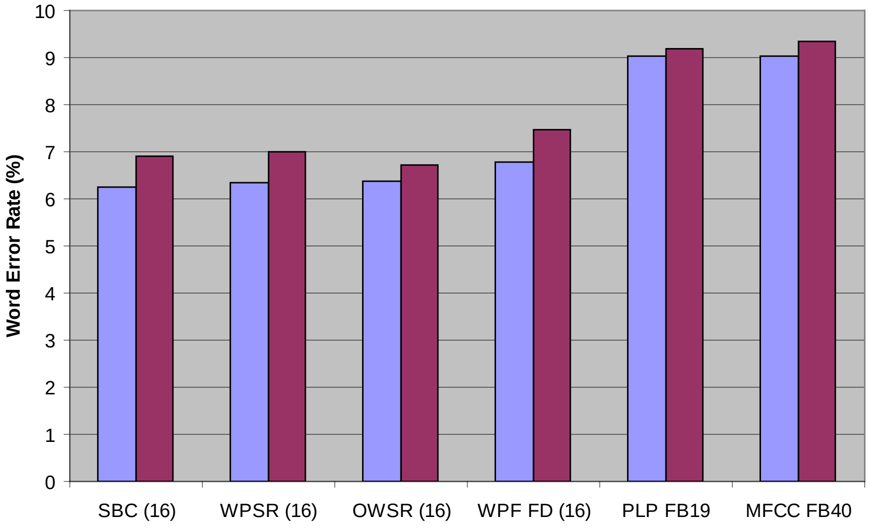

Figure 4. Results for speech frame length of 25.625 milliseconds: The WER for language weight = 9.5 (left bar) and language weight = 12.0 (right bar)

As comes out from Figure 4 all wavelet-based speech features achieved lower error rates, and thus significantly outperformed the baseline MFCCs and PLPs. It is interesting to examine the ordering and the improvement the DWPT-based speech features provide when compared to the baseline. As the experimental results illustrate, the lowest error rates for language weight 9.5 were achieved for the SBC, followed by the WPSR and OWPF, the WPF, and next by the baseline DFT-based speech features. For the case of language weight 12.0, the OWPF outperformed the SBC.

Comparing the WER for the left and right bar for each speech feature set (please refer to Figure 4), we can observe that there is only slight increase of the error rates when the language weight is increased form 9.5 to 12, while change of the language weight form 9.5 to 12 corresponds to a speed-up of the decoding operation by about 1.5 times.

In order to compare parameterization techniques fairly, the experiments involving MFCC and PLP speech features were repeated for frame length of 16 milliseconds, identical to the frame length that the

DWPT-derived speech features use. The WERs for language weights 9.5 (left) and 12.0 (right) are shown in Figure 5. For convenience, the results for DWPT-based features are duplicated from Figure 4.

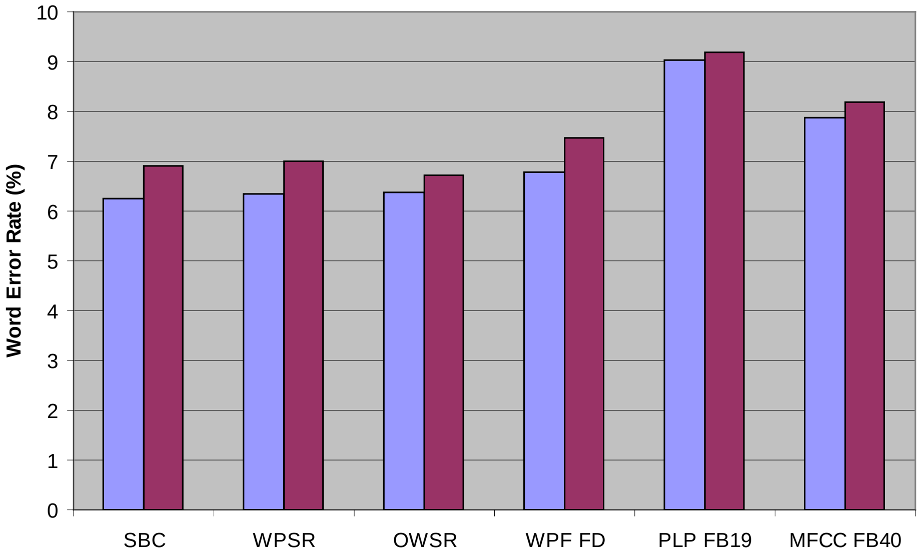

Figure 5. Results for speech frame length of 16 milliseconds: The WER for language weight = 9.5 (left) and language weight = 12.0 (right)

As the experimental results in Figure 5 reveal, the DWPT-derived speech features retained superiority over the baseline MFCC and PLP. With small exceptions, the ordering of the speech features remained the same as in Figure 4.

In order to investigate the statistical significance between the speech features, pair-wise t-test [33] was carried out for every pair of results. The results of the t-test showed that the SBC and the WPSR and OWPF are statistically identical in terms of word error rate performance and statistically different from the DFT-based features. Furthermore, the result for WPF is statistically different to the other DWPT-based features, as well as to the MFCCs and PLPs. The outcome of the t-test explains well the slightly different ranking of the results for the language weights 9.5 and 12 (Figure 5).

Summarizing the results presented in Figure 5, we can see that the SBC of Sarikaya & Hansen demonstrated a relative reduction of the WER by more than 20% and 30%, when compared to the baseline MFCC and the PLP, respectively. The other DWPT-based speech features also demonstrated significant reduction of the error rates.

The superior performance for the DWPT-based speech features for both frame lengths (25.625 and 16 milliseconds) may be mainly attributed to a number of reasons. Firstly, the above described wavelet-packet parameterization algorithms provide a balanced time-frequency resolution, in contrast to the

DFT-based techniques, which utilize uniform frequency analysis. In addition, the basis functions, used in the DWPT-based techniques, seem to be a more efficient choice for analysis of non-stationary signals, like speech waveforms, when compared to the cosine functions of the DFT-based techniques. Finally, the flexibility of selecting a particular basis for the optimal representation of a specific speech signal among a plethora of representation bases.

6. Summary and conclusion

The present study is an explicit demonstration of the advantages of wavelet packet-based speech parameterization techniques over widely acknowledged Fourier-based similar schemes. This is achieved through the comparative evaluation of four wavelet packet-based methods to the baseline parameterizations of MFCC and PLP. Since wavelet packet-based speech parameterization is a relatively new area, this work reinforces the trend towards wavelet packets as a promising research direction bringing potentially further reduction of the error rates on the automatic speech recognition task. This, in effect, will hopefully motivate researchers for further elaboration on wavelet packets for speech parameterization, not only for the needs of speech recognition, but also for other tasks related to speech processing, such as speaker recognition, language identification, emotion recognition from speech, etc. Furthermore, the current evaluation can be a precious guide to those who build speech recognition systems, since it renders unnecessary the implementation and comparison of various speech parameterization schemes, especially those based on wavelet packets.

7. References

[1] Baum, L.E., Petrie, T. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[2] Rabiner, L.R. “A tutorial on hidden Markov models and selected applications in speech recognition”. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[3] Viterbi, A. J. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[4] Forney, G.D. “The Viterbi algorithm”. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[5] Gong, Y. “Speech recognition in noisy environments: A survey”. *Speech Communication*, 16(3):261–291, 1995.

[6] Welling, L., Kanthak, S., Ney, H. “Improved methods for vocal tract normalization”. *Proc. of the ICASSP 1999*, vol. 2, pp.761–764.

[7] Zhan, P., Westphal, M. “Speaker normalization based on frequency warping”. *Proc. of the ICASSP 1997*, pp. 1039–1042.

[8] Garau, G., Renals, S., Hain, T. “Applying vocal tract length normalization to meeting recordings”. *Proc. of the INTERSPEECH 2005*, Lisbon, Portugal, pp. 265–268.

[9] Deng, L., Huang, X. “Challenges in adopting speech recognition”. *Communications of the ACM*, 47(1):69–75, 2004.

[10] Oppenheim, A.V. “A speech analysis-synthesis system based on homomorphic filtering”. *Journal of the Acoustical Society of America*, 45:458–465, 1969.

[11] Atal, B.S., Hanauer, S.L. “Speech analysis and synthesis by linear prediction of the speech wave”. *Journal of the Acoustical Society of America*, 50(2):637–655, 1971.

[12] Atal, B.S. “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”. *Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[13] Davis, S.B., Mermelstein, P. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[14] Hermansky, H. “Perceptual linear predictive (PLP) analysis for speech”. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[15] Kim, D.S., Lee, S.Y., Kil, R.M. “Auditory processing of speech signals for robust speech recognition in real-world noisy environments”. *IEEE Transaction on Speech and Audio Processing*, 7(1):55–69, 1999.

[16] Assaleh, K.T., Mammone, R.J. “Robust cepstral features for speaker identification”. Proc. of the ICASSP’94, Adelaide, Australia. Vol.1, 1994, pp. 129–132.

[17] Sarikaya, R., Hansen, H.L. “High resolution speech feature parameterization for monophone-based stressed speech recognition”. *IEEE Signal Processing Letters*, 7(7):182–185, 2000.

[18] Farooq, O., Datta, S. “Mel filter-like admissible wavelet packet structure for speech recognition”. *IEEE Signal Processing Letters*, 8(7):196-198, 2001.

[19] Siafarikas, M., Ganchev, T., Fakotakis N., Kokkinakis G. “Wavelet Packet Approximation of Critical Bands for Speaker Verification”. Submitted to the *International Journal of Speech Technology, Kluwer Academic Publishers.*

[20] Siafarikas, M., Ganchev, T., Fakotakis, N., Kokkinakis, G. “Overlapping Wavelet Packet Features for Speaker Verification”. *Proc. of the INTERSPEECH 2005*, Lisbon, Portugal, pp. 3121–3124, 2005.

[21] Lee, K.-F., Hon, H.-W., Reddy, R. “An overview of the SPHINX speech recognition system”. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(1):35–45, 1990.

[22] Garofolo, J. “Getting started with the DARPA-TIMIT CD-ROM: An acoustic phonetic continuous speech database”. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA, 1988.

[23] Slaney, M. “Auditory toolbox. Version 2”. Technical Report #1998-010, Interval Research Corporation, 1998.

[24] Percival, D.B., Walden, A.T. (2000). Wavelet methods for time series analysis. *Cambridge University Press*, USA.

[25] Fletcher, H. “Auditory patterns”. *Reviews of Modern Physics*, 12:47–65, 1940.

[26] Skowronski, M.D., Harris, J.G. “Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition”. *Journal of the Acoustical Society of America*, 116(3):1774–1780, 2004.

[27] Zheng, F., Zhang, G., Song, Z. “Comparison of Different Implementations of MFCC”, *Journal of Computer Science & Technology*, 16(6):582-589, 2001.

[28] Young, S.J., Odell, J., Ollason, D., Valtchev, V., Woodland, P. “The HTK Book. Version 2.1”, *Department of Engineering, Cambridge University*, UK, 1995.

[29] Moore, B.C.J., Glasberg, B.R. “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns”. *Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[30] “RobustGroup’s Open Source Tutorial – Learning to use the CMU SPHINX Automatic Speech Recognition system”. Available: http://www.speech.cs.cmu.edu/sphinx/tutorial.html.

[31] Bakis, R., “Continuous speech word recognition via centi-second acoustic states”, *Proc. of ASA Meeting (Washington, DC)*, April 1976.

[32] “The CMU-Cambridge Statistical Language Modelling Toolkit, v2”. Available: http://www.speech.cs.cmu.edu/SLM/toolkit_ documentation.html.

[33] Freedman, D.A, Pisan, R., Purves, R.A. Statistics. Third Edition, *W. W. Norton & Company, Inc.* New York (1998).