Int J Speech Technol (2007) 10: 197–218
DOI 10.1007/s10772-009-9028-6

Wavelet packet approximation of critical bands for speaker verification

Mihalis Siafarikas · Todor Ganchev · Nikos Fakotakis · George Kokkinakis

Received: 29 November 2006 / Accepted: 6 April 2009 / Published online: 15 April 2009

Abstract Exploiting the capabilities offered by the plethora of existing wavelets, together with the powerful set of orthonormal bases provided by wavelet packets, we construct a novel wavelet packet-based set of speech features that is optimized for the task of speaker verification. Our approach differs from previous wavelet-based work, primarily in the wavelet-packet tree design that follows the concept of critical bands, as well as in the particular wavelet basis function that has been used. In comparative experiments, we investigate several alternative speech parameterizations with respect to their usefulness for differentiating among human voices. The experimental results confirm that the proposed speech features outperform Mel-Frequency Cepstral Coefficients (MFCC) and previously used wavelet features on the task of speaker verification. A relative reduction of the equal error rate by 15%, 15% and 8% was observed for the proposed speech features, when compared to the wavelet packet features introduced by Farooq and Datta, the MFCC of Slaney, and the subband based cepstral coefficients of Sarikaya et al., respectively.

Keywords Wavelet packets · Speech features · Speaker verification · Speaker recognition

M. Siafarikas · T. Ganchev (✉) · N. Fakotakis · G. Kokkinakis
Wire Communications Laboratory, Dept. of Electrical and Computer Engineering, University of Patras, Rion-Patras 26500, Greece
e-mail: tganchev@ieee.org

M. Siafarikas
e-mail: msiafarikas@upatras.gr

N. Fakotakis
e-mail: fakotakis@upatras.gr

G. Kokkinakis
e-mail: gkokkin@upatras.gr

1 Introduction

Although many speech processing tasks, like speech and speaker recognition, have reached satisfactory performance levels on specific applications, and even though a variety of commercial products were launched in the last decade, many problems still remain as an open research area and flawless solutions haven't yet been found. For example, such a problem is providing a suitable parameterization of the speech signal for the needs of speaker recognition. In fact, contemporary speaker recognition systems are composed of a feature extraction stage which aims at extracting speaker's characteristics while evading any sources of adverse variability, and a classification stage that identifies the feature vector with certain class. The classification stage that is based on the probability density functions of the acoustic vectors is seriously deranged in the case of inappropriate choice of speech features which are suboptimal for the particular task.

Historically, the following speech features have dominated the speech and speaker recognition areas: Real Cepstral Coefficients (RCC) introduced by Oppenheim (1969), Linear Prediction Coefficients (LPC) proposed by Atal and Hanauer (1971), Linear Predictive Cepstral Coefficients (LPCC) derived by Atal (1974), and MFCC (Davis and Mermelstein 1980). Other speech features such as Perceptual Linear Prediction (PLP) coefficients (Hermansky 1990), Adaptive Component Weighting (ACW) cepstral coefficients (Assaleh and Mammone 1994a, 1994b), and various wavelet-based features (Sarikaya et al. 1998; Sarikaya and Hansen 2000; Farooq and Datta 2001; etc.), although presenting reasonable solutions for the same tasks, did not gain widespread practical use, often due to their relatively more sophisticated computation. Nowadays, many earlier computational limitations have been overcome, due to the significant performance boost of contemporary microprocessors. This opens possibilities for reevaluation of the traditional solutions when speech features are selected for a specific task.

In Davis and Mermelstein (1980), it was demonstrated that the biologically motivated 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. This success of MFCC, combined with their robust and cost-effective computation, turned them into a standard choice in the speech recognition applications. MFCC became widely used on speaker recognition tasks, too, although they might not represent the human voice uniqueness with sufficient accuracy. In fact, when MFCC are used for speech recognition, it is feasible to suppress the individuality of different voices, while the linguistic information remains unaffected by this process. However, in the text-independent speaker recognition task, the linguistic information is not a beneficial source of information, and yet, its presence makes the speaker recognition process even more difficult.

Over the past two decades, wavelet analysis has proven to be an effective signal processing technique for a variety of problems. In particular, in feature extraction schemes designed for the purpose of speech recognition, wavelets have been used twofold. The first approach uses wavelet transform as an effective decorrelator instead of Discrete Cosine Transform in the feature extraction stage (Tufekci and Gowdy 2000). According to the second approach, wavelet transform is applied directly on the speech signal. In this case, either wavelet coefficients with high energy are taken as features (Long and Datta 1996), which nonetheless suffer from shift variance, or subband energies are used instead of the Mel filter-bank subband energies as in Sarikaya and Hansen (2000). In particular, in the speech recognition area, the wavelet packet transform, employed for the computation of the spectrum, was first proposed in Erzin et al. (1995). Later on, wavelet packet bases were used in Sarikaya et al. (1998), Sarikaya and Hansen (2000) and Farooq and Datta (2001, 2002) for the construction of features that were close approximations of the Mel-frequency division using Daubechies' orthogonal filters with 32 and 12 coefficients, respectively. Recently, Nogueira et al. (2006) studied three filter-banks that are based on the Advanced Combinational Encoder (ACE) "NofM" strategy (Nogueira et al. 2005). Specifically, the authors investigated the appropriateness of three different basis functions, namely the Haar wavelet, the Daubechies' wavelet of order 3 and the Symlets family of wavelets, for improving the speech intelligibility in cochlear implants. All experiments were performed on a common decomposition tree that closely follows the frequency bands associated with the electrodes in the ACE strategy.

In the present study by means of wavelet packets, we seek a more general approach that allows easy handling of the spectral content of a speech signal, flexible utilization of the important frequency bands, and a variable frequency resolution in each subband. Specifically, as an alternative to the well-known MFCC features, wavelet packets are exploited to approximate the psychoacoustic effect explained by the critical bands concept, which was introduced in Fletcher (1940). Wavelet packets are especially suitable for this approximation as they provide various orthonormal transforms each one with different time-frequency localization properties especially assigned to particular purposes. Moreover, of equal importance is that the wavelet packets technique employed here allows a beneficial selection of the underlying basis functions, which in cepstral-coefficient-like Discrete Fourier Transform (DFT)-based schemes is fixed to sinusoidal functions.

Our proposal differs from the aforementioned related studies, chiefly in the wavelet packets tree design, but also in the particular wavelet that has been used. In two earlier works (Siafarikas et al. 2004; Ganchev et al. 2004), the authors demonstrated two examples of successful wavelet packet-based speech features, fine-tuned for the task of speaker verification. In the present study, we fully develop the methodology for building such trees. All potential wavelet packet tree candidates are explored in a systematic way and the best one is selected in an objective manner. In addition, both the criteria for selecting an appropriate wavelet packet basis and the particular choice of the most suitable wavelet function are discussed. As the experimental results presented in Sect. 8 demonstrate, the speech features proposed in the present work outperform MFCC, as well as the wavelet-based features introduced in Sarikaya et al. (1998), Sarikaya and Hansen (2000) and Farooq and Datta (2002). The advantage of the proposed speech features is attributed to the following four reasons: (1) the underlying wavelet function was selected in an objective way to maximize the speaker verification performance, (2) the wavelet packet tree design, although inspired by the critical bands concept, was fine-tuned in a systematic way for achieving a better speaker discrimination power, (3) the optimal selection of frequency resolution in the different subbands that accounts for the recent advances in the theory of critical bands, (4) the availability of a larger set of relevant and non-redundant speech features, when compared to the MFCC parameters. In fact, the advantage (4) is a direct consequence of the larger number of frequency subbands due to reasons (2) and (3).

The remainder of this article is organized as follows: In Sect. 2, we introduce briefly the speaker verification problem and terminology, outline the baseline MFCC parameters and the speaker verification system that serves as platform for comparative evaluation of various speech parameterization techniques, and define the performance assessment measures. Section 3 outlines the speech corpora employed in the development and evaluation of the proposed speech parameterization scheme. The critical bands concept is presented briefly in Sect. 4. Section 5 offers a short introduction to the Discrete Wavelet Packet Transform, which is the basic instrument in our research. Section 6 describes the proposed speech parameterization approach. Specifically, in the subsections we firstly select the particular wavelet function for the wavelet-packets tree design, perform systematic evaluation of sixteen wavelet packet trees, and offer a recipe for the computation of the proposed speech features. Section 7 outlines two alternative DWPT-based speech parameterization schemes that were reported to outperform the MFCC. In Sect. 8, we perform a comparative evaluation of the baseline, proposed and alternative speech parameters. Finally, in Sect. 9 conclusion remarks are offered.

2 The speaker verification problem

In general, the speaker verification (SV) problem can be described as making a decision about acceptance or rejection of an identity claim, judging on the base of the questioned claim and a given speech utterance. Thus, during its operation any automatic SV system receives two inputs: an identity claim (PIN, user's name, keyword, etc.) made by the speaking person and a certain amount of speech, representing his/her or someone else's voice. Another term frequently used for description of the SV task is *one-speaker detection*, which emphasizes the fact that the outcome of the SV process is always a binary decision: *yes*—the speaker is accepted with the claimed personality, or *no*—s/he is rejected. The actual decision depends on the degree of similarity between the speech sample and a predefined model for the enrolled user, whose identity the speaker claims. When an enrolled user, i.e. *client* of the system, claims her/his own identity, we designate the input utterance as a *target* trial. When a non-user addresses a SV system, or when an enrolled user claims identity belonging to another user, we denote that utterance as a *non-target* trial. The non-target trials are also referred to as *impostor* trials.

With respect to the linguistic contents of the speech data, the SV process can be *text-dependent* or *text-independent*. The text-dependent scenarios are mostly used in applications which require high degree of security. Most often these are military, monetary, or other restricted-access or highly guarded subdivisions, where the major requirement is low probability of false acceptance of non-authorized persons. In that connection, during the verification process a strong cooperation by the speaker is required since s/he has to follow a strict predefined protocol, and carefully follow the instructions of the system. Furthermore, in these applications the user is identified through detecting something he has, examining something he knows, and verifying something he is, e.g. his voice. As it is obvious, in the text-dependent task the personal comfort of users is a secondary consideration. In opposite, in case of text-independent speaker verification, which is the most challenging among all SV tasks, the speaker is not obligated to follow a specific predefined scenario or instructions from the system, such as pronouncing a password, or prompted by the system sequence of numbers and/or sentences. Therefore, in the text-independent scenario the SV decision is solely based on a given identity claim and voice, and not on something the person has or something s/he knows. Since an explicit cooperation is not required the text-independent scenario is more comfortable for the user and the SV process may remain hidden for her/him.

2.1 Description of the speaker verification system

The SV system briefly described in the following serves as a platform on which we evaluate the practical usefulness of several alternative speech parameterization techniques. This SV system (Ganchev et al. 2002a, 2002b), which successfully participated in the 2002 NIST Speaker Recognition Evaluation (NIST 2002) has a modular structure, where each enrolled user is detected by an individual expert. Each expert considers two hypotheses—either the input speech originates from the same person, whose identity the speaker claims, or it originates from another person, which has different identity. In order to test each of these two hypotheses, we build two models: one for the voice of the enrolled user, and another one representing the rest of the world. The latter model is also designated as a *reference*. Since the reference model has to be sufficiently flat not to interfere with the models of the individual users, it is built by exploiting large amounts of speech from multiple speakers.

In Fig. 1, a simplified block diagram of the Probabilistic Neural Network (PNN)-based SV system is presented. The upper part of the figure summarizes the process of training, where the process of building of the reference model, as well as construction of the individual codebooks for the target speakers is shown. As the figure presents, an individual PNN for each of the target users is trained, by utilizing the reference codebook and the one created for the corresponding user. The lower part of the figure illustrates the operational mode of the system. The processing steps, the SV system performs for each test trial in order to make a final decision, are shown. In the following subsections, the main building blocks of our SV system are described in details.

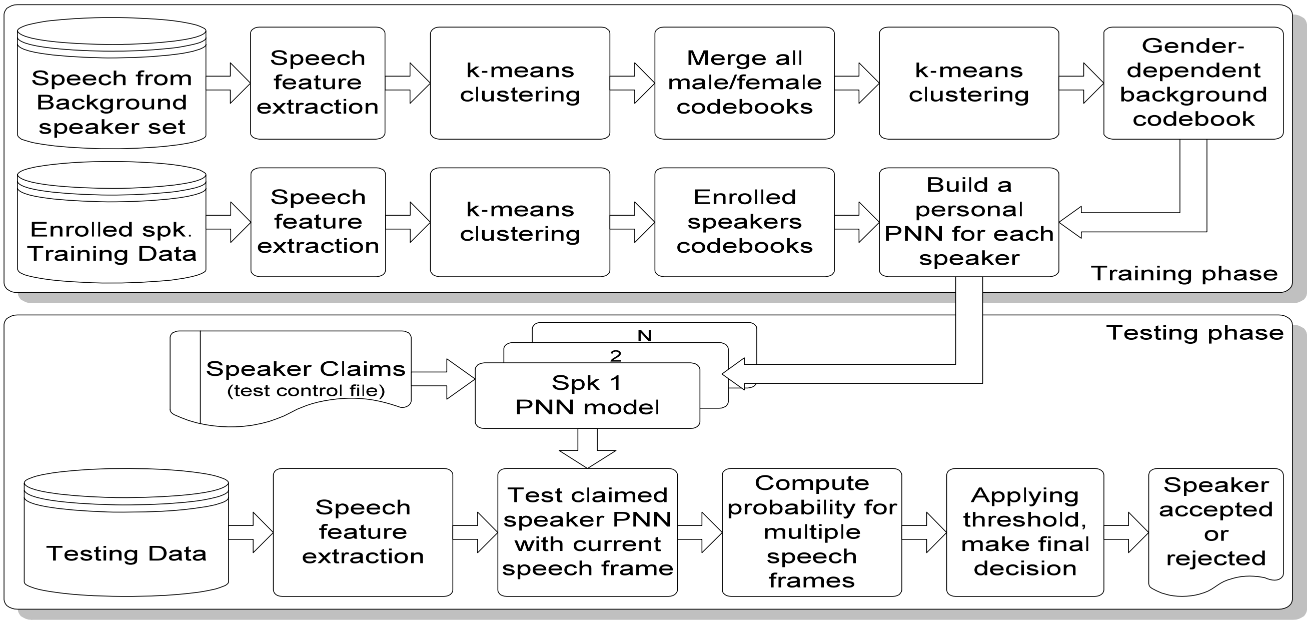

Speech from Background speaker set
Speech feature extraction
k-means clustering
Merge all male/female codebooks
k-means clustering
Gender-dependent background codebook
Enrolled spk. Training Data
Speech feature extraction
k-means clustering
Enrolled speakers codebooks
Build a personal PNN for each speaker
Training phase
Testing phase
Speaker Claims (test control file)
N
2
Spk 1 PNN model
Testing Data
Speech feature extraction
Test claimed speaker PNN with current speech frame
Compute probability for multiple speech frames
Applying threshold, make final decision
Speaker accepted or rejected

Fig. 1 A simplified block diagram of the speaker verification system

2.1.1 *Speech pre-processing and speech parameterization*

The present subsection offers a comprehensive outline of the speech pre-processing and the computation of the MFCCs. In the comparative evaluation of speech parameters (Sect. 8), the MFCCs serve as the baseline speech features.

The speaker recognition corpora utilized in this work (see Sect. 3), consists of telephone quality speech, sampled at 8 kHz. Due to the Lombard effect, or to differences in the temperament and speaking style of talkers, saturation by level is a common phenomenon for telephone speech signals. In order to reduce the spectral distortions it causes, a band-pass filtering of speech is performed as a first step of the feature extraction process. A fifth-order Butterworth filter with pass-band from 80 Hz to 3800 Hz is used for both training and testing. Then the speech signal is pre-emphasized with the filter

$$H(z) = 1 - az^{-1}, \tag{1}$$

where $a = 0.97$, and subsequently windowed into frames of 32 ms duration with skip rate of 16 ms, using a Hamming window. A voiced/unvoiced speech separation is performed by a modification of the autocorrelation method with clipping (Rabiner et al. 1976). Only the voiced speech frames are used due to their relatively better robustness to interference. Next, each voiced speech frame is subjected to $N = 1024$-point short-time DFT, and afterwards it is passed through a set of $M = 32$ equal-area triangular band-pass filter-bank channels:

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

where $i = 1, 2, \ldots, M$ stands for the ith filter, f_{b_i} are $M + 2$ boundary points that specify the M filters, and $k = 1, 2, \ldots, N$ corresponds to the kth coefficient of the N-point DFT. Due to the term $2/(f_{b_{i+1}} - f_{b_{i-1}})$ the filter-bank (2) is normalized in such a way that the sum of coefficients for every filter equals one. We have accepted an approximation of the Mel-scale, with 13 linearly spaced filter-banks, lowest central frequency 200 Hz, highest 1000 Hz and 19 log-spaced with highest central frequency 3690 Hz. This filter-bank is essentially the filter-bank introduced in Slaney (1998) adapted for sampling frequency 8 kHz. Subsequently, thirty-two ($J = 32$) MFCC parameters are computed after applying Discrete Cosine Transform (DCT):

$$C_j = \sum_{i=1}^{M} X_i \cos\left(j\left(i - \frac{1}{2}\right)\frac{\pi}{M}\right), \quad j = 0, 1, \ldots, J - 1 \tag{3}$$

to the log-filter-bank outputs X_i:

$$X_i = \log_{10}\left(\sum_{k=0}^{N-1} |X(k)|H_i(k)\right), \quad i = 1, 2, \ldots, M. \tag{4}$$

A comprehensive description of the MFCC computation steps is offered in Ganchev et al. (2005), where various implementation strategies are evaluated.

2.1.2 *The client's and reference models design*

Because the complexity and computational demands of the PNNs depend strongly on the number and dimensionality of the training vectors, a k-means clustering algorithm (Hartigan and Wong 1979) is used to compact the training data. Codebooks are built for both clients and non-clients of the system. The codebook for a client is built from the speech recordings, collected during her/his enrollment session. The non-clients codebooks, built by utilizing a representative speech database, are further utilized in the gender-specific reference model. The common reference model is employed for counterbalancing the scores produced by the individual user models. In the baseline version of our SV system, we used a codebook of 128-vectors for the clients, and a codebook of 256-vectors for the reference model. The size of the codebooks was chosen as a trade-off between computational demands and performance (details in Ganchev 2005, Chap. 3).

2.1.3 *The PNN-based classifier*

The PNNs, introduced in Specht (1990), combine nonparametric probability density estimation with minimum risk decision making. The density estimation implements the Parzen window estimator (Parzen 1962) by using a mixture of Gaussian basis functions. After the probability density functions for all classes are estimated, the posterior probabilities are computed, and afterwards the Bayes' optimal decision rule is applied to select the winning class.

Since the SV process is a two-class separation problem, a PNN for classification in $K = 2$ classes is considered. The probability density function $\mathrm{f}_i(\mathbf{x}_p)$ of each of the two classes k_i, $i = 1, 2$, is computed by:

$$\mathrm{f}_i(\mathbf{x}_p) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{M_i} \times \sum_{j=1}^{M_i} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}_p - \mathbf{x}_{ij})^T(\mathbf{x}_p - \mathbf{x}_{ij})\right), \quad i = 1, 2, \tag{5}$$

where $\mathbf{x}_{ij}$ is the jth training vector from class k_i; $\mathbf{x}_p$ is the pth input vector; d is the dimension of the speech feature vectors; and M_i is the number of training patterns in class k_i. Each training vector $\mathbf{x}_{ij}$ is assumed a centre of a kernel function, and consequently the number of pattern units in the first hidden layer of the neural network is given as a sum of the pattern units for all classes. The standard deviation σ acts as a smoothing factor, which softens the surface defined by the multiple Gaussian functions. As presented in (5), σ has the same value for all the pattern units, and therefore, a homoscedastic PNN is considered.

After the estimations of the class-conditional probability density functions is obtained through (5), the Bayes' optimal decision rule (6) is applied to distinguish class k_i, to which the pth input vector $\mathbf{x}_p$ belongs:

$$\mathrm{D}(\mathbf{x}_p) = \underset{i}{\operatorname{argmax}}\{h_i c_i \mathrm{f}_i(\mathbf{x}_p)\}, \quad i = 1, 2, \tag{6}$$

where h_i is the a priori probability of occurrence of the patterns of category k_i, and c_i is the cost function in case of misclassification of a vector belonging to class k_i. A comprehensive description of the PNN is available in Specht (1990).

2.1.4 *The output score computation, making the final decision*

The probability $P(k_i|\mathbf{X})$ all test vectors $\mathbf{x}_p$ of a given test trial $\mathbf{X} = \{\mathbf{x}_p\}$, $p = 1, 2, \ldots, P$ to belong to class k_i is computed as:

$$P(k_i|\mathbf{X}) = \frac{N(\mathrm{D}(\mathbf{x}_p) = k_i)}{\sum_{j=1}^{K} N(\mathrm{D}(\mathbf{x}_p) = k_j)}, \quad i = 1, 2, \tag{7}$$

where $N(\mathrm{D}(\mathbf{x}_p) = k_i)$ is the number of vectors $\mathbf{x}_p$ categorized by the Bayes' optimal decision rule (6) as belonging to class k_i. Since the SV task assumes an exhaustive taxonomy, any of the inputs $\mathbf{x}_p$ falls in one of the two classes k_i. Next, for a given test trial, the averaged probability for all output decisions of a particular PNN, obtained by testing with multiple feature vectors, is utilized to compute a score:

$$\chi = \eta(P(k_1|\mathbf{X}) - \beta), \tag{8}$$

where η and β are constants for tuning the scale and the offset of the produced score, respectively.

A speaker-independent threshold θ, pre-computed on a development dataset in a manner that satisfies the decision strategy of the prospective application, is applied to the score (8), and a final decision $O(\theta)$ is made:

$$O(\theta) = \begin{cases} 1, \text{ i.e. accept} & \text{for } \chi \geq \theta, \\ 0, \text{ i.e. reject} & \text{for } \chi < \theta. \end{cases} \tag{9}$$

When the score χ is above or equal to the threshold, the claimant is accepted. Otherwise, the utterance is considered to belong to an impostor speaker.

2.1.5 *The operational mode*

In summary, the SV system decides whether or not the input trial belongs to the claimed speaker, depending on the degree of similarity of the input feature vectors to the speaker's model and to the reference model. Equation (5) estimates the degree of similarity by computing the corresponding Euclidean distances. For every input speech frame, a binary decision is made by applying the Bayes' optimal decision rule (6). Next, through (7) an estimate of the probability a given test trial to belong to the claimed user is obtained. Finally, a speaker-independent threshold is applied to the score computed through (8), and a final decision as specified by (9) is made. A comprehensive description of our SV system is available in Ganchev et al. (2002a, 2002b).

2.2 Cost-based speaker verification performance measure

Two types of errors can occur in the SV process. The first one, called a *false rejection* (FR) error, occurs when the true target speaker is falsely rejected as being an impostor, and as a result, the system *misses* recognizing an attempt belonging to the true authorized user. The second type, called a *false acceptance* (FA) error, occurs when a tryout from an impostor is accepted as if it came from the true authorized user. The latter error is also known as a *false alarm*, because a non-target trial is accepted as a target one. The FR and FA are employed together to characterize the performance of the SV systems under investigation.

A cost-based performance measure C_{Det} (10) was used in the experiments to assess the SV performance. It is defined (NIST 2002) as a weighted sum of the false acceptance and false rejection error probabilities, designated as $P(FalseAlarm|NonTarget)$ and $P(Miss|Target)$, respectively:

$$\begin{aligned} C_{Det} = {} & C_{Miss} P(Miss|Target) P(Target) \\ & + C_{FalseAlarm} P(FalseAlarm|NonTarget) \\ & \times (1 - P(Target)), \end{aligned} \tag{10}$$

where the parameters C_{Miss} and $C_{FalseAlarm}$ are the relative costs of detection errors, and $P(Target)$ is the a priori probability of the specified target speaker. According to the rules of the 2001 NIST Speaker Recognition Evaluation (SRE), the target speaker probability in the experiments with the NIST 2001 SRE database is $P(Target) = 0.01$, and $P(NonTarget) = 1 - P(Target) = 0.99$, and the costs of the FR and FA are $C_{Miss} = 10$ and $C_{FalseAlarm} = 1$, respectively. The cost measure C_{Det} is further normalized as:

$$\begin{aligned} & C_{Norm} = C_{Det}/C_{Default}, \\ & \quad \text{where } C_{Default} = \min\{C_{Miss} P(Target), \\ & \qquad\qquad C_{FalseAlarm} P(NonTarget)\} \end{aligned} \tag{11}$$

for making its values more intuitive. Here, $C_{Default}$ represents the zero value (a system providing no information), which is the cost obtained without processing the data, always making the same decision—either accept or reject. Finally, the range of values received by C_{Norm} is between zero, for a system that makes no mistakes, and a positive constant that depends on the ratio of the products $C_{Miss} P(Target)$ and $C_{FalseAlarm} P(NonTarget)$, for a worthless system. The actual decision cost, *DCFact*, is the decision cost C_{Norm} computed after the final decision is made. Thus, the *DCFact* depends not only on the quality of modeling, but also on the relevance of a priori estimated speaker-independent threshold. The optimal decision cost, *DCFopt*, gives an impression about the prospective performance of a system when "the optimal" speaker-independent threshold is applied.

Since the values of C_{Norm} are not as intuitive as other widely used performance measures, we also provide the Equal Error Rate (EER) decision point, where the false rejection and the false acceptance error probabilities are equal, i.e., when computing the EER, we assume equal weights for the SV cost parameters $C_{Miss} = C_{FalseAlarm} = 1$. The EER decision point is accepted as intuitive and more balanced performance estimation. However, it has the disadvantage that the final decision is made a posteriori, and thus, the reported SV performance is too optimistic. While the EER gives an application-independent assessment of the potential performance of a system, the *DCFact* and *DCFopt* are application-specific due to the cost coefficients C_{Miss} and $C_{FalseAlarm}$. In practice, these costs are application-dependent. Their ratio varies from one application to another within the range of 1:10 to 10:1, depending on whether the emphasis is placed on security or comfort of use.

Since the present study aims at comparing the practical usefulness of various feature extraction techniques rather than optimizing an absolute SV performance, we accepted the *DCFopt* and EER as the main performance measures. Together they offer a unique representation of every experimental result.

3 Speaker recognition corpora

There exists a multitude of speaker recognition corpora collected to capture the variability of real-world conditions. These corpora are representative for a particular application and have been created to address specific aspects of the speaker recognition problem, such as the influence of the microphone/handset type; transmission channel effect; environmental interferences; emotional speech; speech under stress, etc. None of these corpora provides universal environment for research that is exhaustive for a real-world application. In fact, these corpora offer the opportunity some specific aspects of interest to be studied under controlled conditions. A typical member of this group is the Polycost speaker recognition database, which has been collected to assist research in land phone-based teleservices, which incorporate speaker recognition functionality. The good compactness of Polycost makes it a convenient choice for the development experiments in Sect. 6, which aim at the optimization of the proposed speech parameterization scheme.

A second kind of speaker recognition databases, which are purposely created by selecting and post-processing portions of large existing corpora, are specially designed for basic technology evaluation. These databases are used in periodic evaluation campaigns (e.g. the annual NIST SRE technology evaluations) and are explicitly designed to have many (but controlled) degrees of variability. The latter allows assessment of various aspects of the evaluated technology just by analyzing the results from a single experiment. The 2001 NIST SRE—one-speaker detection database, which consists of compilation of post-processed and re-segmented mobile phone speech recordings is representative of this second kind of databases. In Sect. 8, we utilize the 2001 NIST SRE data for validating the significance of our approach and assessing the practical value of the proposed speech parameterization scheme.

3.1 The Polycost speaker recognition database

Polycost contains real-world telephone data (English spoken by foreigners) collected across the international telephone networks of Europe. The speech data are representative for research related to telephone-driven telephone services and automated call-centers. The Polycost database contains 1285 calls (around 10 sessions per speaker) recorded by 134 speakers (74 males and 60 females) from 13 different European countries. Each session comprises 10 prompts with connected digits uttered in English, two prompts with sentences uttered in English, and two prompts in the speaker's mother tongue (17 different languages or dialects). One of the prompts in the speaker's mother tongue consists of free speech. Detailed descriptions of the Polycost database can be found in Hennebert et al. (1996, 2000). In the present work, we use version v1.0 of the Polycost speaker recognition database with bugs 1÷5 fixed (Polycost Bugs 1999).

3.2 2001 NIST SRE—one-speaker detection database

The 2001 NIST SRE—one-speaker detection database is an excerpt from the Switchboard-Cellular corpora, which had been post-processed in order to remove any significant pauses in the speech signal and cancel transmission channel echoes. The training data consist of spontaneous speech from 74 male and 100 female speakers, recorded in different environmental conditions: {'inside', 'outside', 'vehicle'}. All training speech had been acquired over the mobile cellular networks of USA. Each target user is represented by about 2 minutes of spontaneous speech, extracted from a single conversation. The test data consist of speech recorded over {'TDMA', 'CDMA', 'Cellular', 'GSM', and 'Land'} transmission channels. Both same and different phone number calls (the latter imply different handsets) and different transmission channels are available for each user. Depending on the amount of speech the test trials contain, they are separated in the following five categories: {'00-15', '16-25', '26-35', '36-45', and '46-60'} seconds. The complete one-speaker detection task includes all test trials, and therefore covers all aforementioned sources of variability. In the present study, we only consider the complete one-speaker detection task, and no details are given for the sub-tasks. A comprehensive description of the evaluation database and evaluation rules is available in the 2001 NIST SRE Plan (NIST 2001).

4 Critical bands concept

During the past decades, considerable progress has been made in the exploration of the human auditory system (Fletcher 1940; Zwicker 1961; Glasberg and Moore 1990). Fletcher (1940) suggested that the peripheral auditory system behaves as if it consisted of a bank of bandpass filters with overlapping passbands, now referred to as auditory filters. The frequency selectivity of the auditory system and the characteristics of its corresponding auditory filters can be investigated by conducting perceptual experiments based on the technique of masking. The masking effect prevents the human auditory system from being sensitive to the detailed spectral structure of a sound within the bandwidth of a single auditory filter. To describe the effective bandwidth of the auditory filter over which the main masking effect takes place, Fletcher introduced the term of *critical bandwidth* (CB). He also used the phrase *critical bands* to refer to the concept of the auditory filters.

Since Fletcher's first description of the critical bands, many experimenters attempted to estimate it. Zwicker (1961) estimated that the critical bandwidth is constant and equal to 100 Hz for frequencies below 500 Hz, while for higher frequencies, it increases approximately in proportion with the centre frequency. Exploiting this first approximation, Davis and Mermelstein (1980) proposed the renowned MFCC features based also on another similar perceptual (subjective) measure, namely the Mel-scale, which represents the perceived frequency or pitch of a tone as a function of its acoustic frequency. However, MFCC features, as computed in Davis and Mermelstein (1980), include approximations of the critical bands that do not completely conform to our today's understanding of that matter, as explained in the paragraphs that follow.

The critical bandwidth relationship, derived by Zwicker, was estimated when there were only few estimates available for low centre frequencies. However, in more recent attempts to determine the shape of the auditory filters and estimate the Equivalent Rectangular Bandwidth (ERB), Moore (2003) demonstrated that there are discrepancies in comparison with Zwicker's findings (Zwicker 1961), particularly at frequencies below 500 Hz where the critical bandwidth continues to decrease with frequency.

The ERB might be regarded as a measure of the CB and, according to Moore (2003), it is equal to the bandwidth of a perfect rectangular filter, whose pass-band is equal to the maximum transmission of the specified filter and transmits the same power of white noise as the specified filter. Equation (12) presents the ERB as a function of centre frequency f using moderate sound levels for young people with normal hearing (Glasberg and Moore 1990):

$$\mathrm{ERB} = 24.7(4.37f + 1), \tag{12}$$

where the values of ERB and f are specified in Hz and kHz, respectively. As presented in Moore (2003), (12) fits roughly the values of ERB estimated in many different laboratories. Therefore, ERB approximates the CB, which in turn is a subjective measure of the bandwidth of the auditory filters. However, the CB function derived by Zwicker fits markedly worse to the above equation at low frequencies.

The incorporation of that knowledge in automatic speech and speaker recognition led to an increased performance due to the emergence of new feature extraction techniques. Many innovative speech features were designed, for example: PLP (Hermansky 1990), ACW (Assaleh and Mammone 1994a, 1994b), but the MFCC paradigm preserved its predominance. In Sect. 6, exploiting the most recent advances in the understanding of the human auditory system and based on the insights of Moore (2003), the authors provide an approximation of the perceptual behavior of the auditory system which in turn ensures the framework for the construction of a successful wavelet packets tree. Before that, however, for comprehensiveness of our exposition, in Sect. 5 we briefly introduce the wavelet packet analysis and the discrete wavelet packets transform.

5 Wavelet packet analysis

Wavelet analysis is relatively new and is usually considered complementary to existing analysis techniques such as Fourier analysis. However, in addition to its profound theoretical background which stems from group theory of representations (Daubechies 1992), in many cases, wavelets successfully contributed to the solution of problems for which limited progress had been made prior to their introduction, as for instance: noisy signal estimation and compression. The present section introduces wavelet analysis and underlines its major advantages over Fourier analysis, paying particular attention to wavelet packet analysis.

5.1 Discrete Wavelet Transform

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. A comprehensive analytic presentation of wavelet analysis and its basic transform, CWT, can be found in Mallat (1998).

The basic tool for a 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 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 $t_j = 2^j k, k \in \mathbb{Z}$ 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 a signal.

Although DWT can be regarded as an attempt to preserve the key features of the CWT in a succinct manner, it can be formulated entirely in its own right. Analytically, the DWT of level $J \in \mathbb{N}$ of a discrete-time signal $x[n], n = 0, 1, \ldots, N-1$ with $N = 2^J$, is an N-dimensional vector, $\mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2\ \cdots\ \mathbf{W}_J\ \mathbf{V}_J]^T$. Each $\mathbf{W}_j$ is an $N/2^j$-dimensional vector of wavelet coefficients each one of which is associated with adjacent time intervals of width 2^j and frequency interval $[1/2^{j+1}, 1/2^j]$, while $\mathbf{V}_J$ is a one dimensional vector containing the scaling coefficient associated with the whole time interval of width 2^J and frequency interval $[0, 1/2^{J+1}]$. Therefore, in contrast to the frequency analysis provided by DFT, the DWT provides a time-frequency decomposition of a signal in the manner illustrated in Fig. 2. The tiling of the time-frequency plane with rectangles of different size means that energy components of the signal within different rectangles of specific time and frequency coordinates can be discerned. Therefore, the rectangles are an indication of the optimal resolution achieved by the time-frequency capability of DWT. As presented in Fig. 2, DWT provides an octave-based decomposition of the frequency domain and gives good frequency resolution in the lower frequencies that gets worse as we move to higher frequencies. On the contrary, DWT provides a good time resolution in the higher frequencies that gets worse as we move towards lower frequencies.

In practice, the DWT is computed by the successive application of two discrete filters, called *wavelet filter* and *scaling filter*, initially on the signal, and subsequently on the lower frequency part of the resulting signal, followed

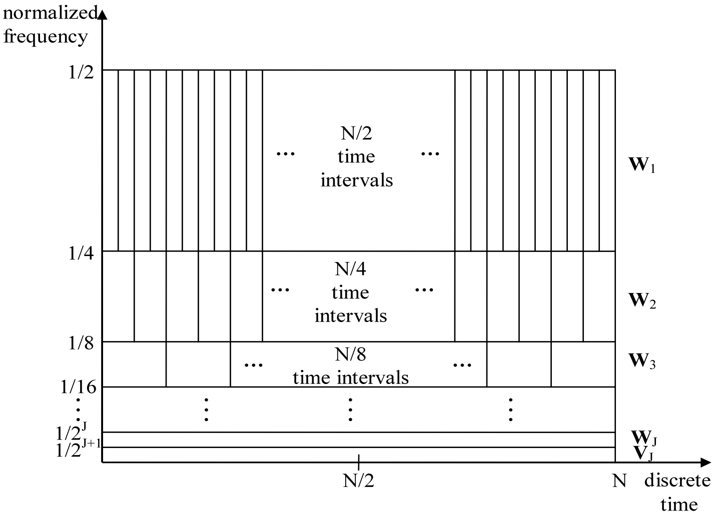


Fig. 2 Discrete Wavelet Transform time-frequency analysis

by subsampling by a factor of two. A filter $\{h_l : l = 0, 1, \ldots, L-1, L = 2k, k \in \mathbb{N}\}$ is called a wavelet filter if

$$\sum_{l=0}^{L-1} h_l = 0 \quad \text{and}$$
$$\sum_{l=0}^{L-1} h_l h_{l+2n} = \begin{cases} 1, & \text{if } n = 0, \\ 0, & \text{if } n \in \mathbb{Z},\ n \neq 0, \end{cases} \tag{13}$$

where the last summation expresses the orthonormality property of a wavelet filter ($\mathbb{Z}$ is the set of integer numbers). As its name suggests, an explicit connection exists between this filter and the wavelet function used in CWT (Percival and Walden 2000). The scaling filter is defined in terms of the wavelet filter via the quadrature mirror relationship $g_l = (-1)^{l+1} h_{L-1-l}$.

Application of the wavelet filter $\{h_l\}$ is equivalent to selecting the higher frequency part of a signal, while application of the scaling filter $\{g_l\}$ is equivalent to selecting the lower frequency part. Therefore, in the decomposition of a signal with the DWT, only the lower frequency band is decomposed giving a right recursive binary tree structure, where its right child represents the lower frequency band and its left child represents the higher frequency band.

Ideally, it would be desirable to have wavelet and scaling filters the frequency response of which were limited to the frequency bands [1/4, 1/2] and [0, 1/4], respectively. In this case, the actual time-frequency tiling of the DWT would coincide with the optimal tiling depicted in Fig. 2. The degree that these nominal frequency intervals are approximated (by the actual frequency decomposition achieved by DWT) depends on the frequency characteristics of the wavelet filter and therefore the scaling filter.

DWT exploits the flexibility provided by the possibility of choosing the wavelet filter according to specific needs. Thus, different wavelet filters can provide fine-tuned DWT analyses, although the optimal time-frequency analysis of DWT is depicted in Fig. 2. Furthermore, wavelet and scaling filters play a similar role in the formulation of Discrete Wavelet Packet Transform (DWPT) presented in the following subsection.

5.2 Discrete Wavelet Packet Transform

Discrete Wavelet Packet Transform is a generalization of the DWT that allows an effective representation of the time-frequency properties of a discrete signal so that useful features for a particular purpose can be appropriately extracted. In this subsection, we present the analysis of a time series with DWPT showing the advantages over both DFT and DWT as far as their capability for time-frequency transform is concerned.

Let $x[n], n = 0, 1, \ldots, N-1$, where N is an integer multiple of 2^J for some positive integer J, denote a real valued discrete time signal. For $0 \leq j \leq J$, the level j DWPT of $x[n]$ is an orthonormal transform yielding an N dimensional vector of coefficients that can be partitioned as $[\mathbf{W}_j^{2^j-1}\ \mathbf{W}_j^{2^j-2}\ \ldots\ \mathbf{W}_j^1\ \mathbf{W}_j^0]^T$, where each $\mathbf{W}_j^n$ 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 $I_j^n = [\frac{n}{2^{j+1}}, \frac{n+1}{2^{j+1}}]$. 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 $[2^j k, 2^j(k+1)]$, $k = 0, 1, \ldots, N/2^j - 1$. Thus, the DWPT provides localized time-frequency description of a signal. In addition, each level of DWPT provides uniform frequency and time analysis, as shown in Fig. 3, in contrast to DWT that provides an octave based decomposition. However, as it is well known, the time-frequency analysis of any time-frequency transform is restricted by Heisenberg's principle

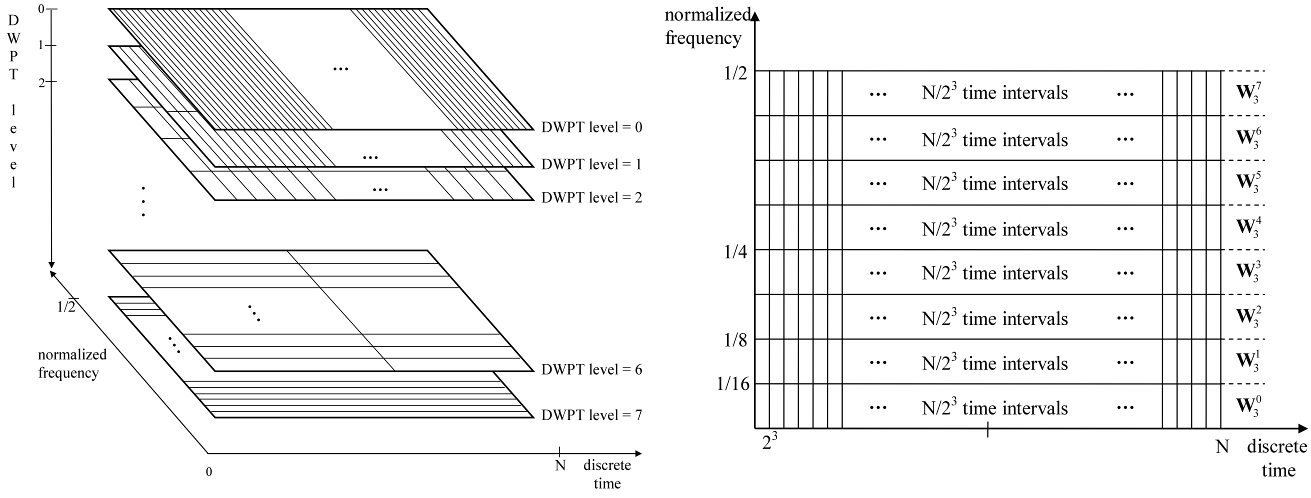


Fig. 3 Three-dimensional illustration of time frequency analysis achieved with levels $j = 0$ through $j = 7$ of Discrete Wavelet Packet Transform. For clarity, DWPT of level $j = 3$ is shown on the right hand side of the figure. At level $j = 0$, time width is 1 and bandwidth is $1/2$; At level j, time width is 2^j and bandwidth is $1/(2 \cdot 2^j) = 1/2^{j+1}$. Therefore, at each level j, the time axis $[0, N]$ is divided into $N/2^j$ intervals while the frequency axis $[0, 1/2]$ is divided into 2^j intervals

stating that good frequency analysis leads to bad time localization and vice-versa. In practice, DWPT of level j is formed by filtering DWPT of level $j - 1$ using the wavelet and scaling filters $\{h_l\}$ and $\{g_l\}$ (defined in Sect. 5.1) and setting $\mathbf{W}_0^0[n] \equiv x[n]$, as shown in the following two relationships:

$$\mathbf{W}_j^{2n}[k] = \sum_{i=0}^{L-1} a_{n,i} \mathbf{W}_{j-1}^{n}[(2k + 1 - i) \mathrm{mod}(N/2^{j-1})], \tag{14}$$
$$k = 0, 1, \ldots, N/2^j - 1, \quad a_{n,i} = \begin{cases} g_i, & \text{if } n \text{ is even,} \\ h_i, & \text{if } n \text{ is odd,} \end{cases}$$

$$\mathbf{W}_j^{2n+1}[k] = \sum_{i=0}^{L-1} b_{n,i} \mathbf{W}_{j-1}^{n}[(2k + 1 - i) \mathrm{mod}(N/2^{j-1})], \tag{15}$$
$$k = 0, 1, \ldots, N/2^j - 1, b_{n,i} = \begin{cases} g_i, & \text{if } n \text{ is odd,} \\ h_i, & \text{if } n \text{ is even.} \end{cases}$$

The most appealing aspect of DWPT is that carefully selected basis vectors belonging to different level DWPTs can be grouped together in order to create an even larger collection of orthonormal transforms. This is achieved by organizing all the DWPTs for levels $j = 0, 1, \ldots, J$ into a tree structure, called wavelet packet (WP) tree. This is possible since, in going from DWPT of level $j - 1$ to the next level j, each *parent node* $\mathbf{W}_{j-1}^{n'}$ is circularly filtered and down-sampled twice: once with the wavelet filter $\{h_l\}$ and once with the scaling filter $\{g_l\}$, yielding two *children nodes* $\mathbf{W}_j^n$ indexed by $n = 2n'$ and $n = 2n' + 1$. Having constructed the WP tree, the coefficient vectors $\mathbf{W}_j^n$ can be collected together to form a set $S = \{\mathbf{W}_j^n : j = 0, 1, \ldots, J, n = 0, 1, \ldots, 2^j - 1\}$, 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 provides a flexible tiling of the time-frequency plane with various frequency resolutions in the corresponding frequency intervals. That property will turn out to be extremely useful in the approximation of the critical bands that the authors present in Sect. 6. Before that, however, in Sect. 6.1 we study several wavelet functions, and subsequently select the one with the highest benefit for the SV task.

6 The proposed DWPT-based speech parameterization scheme

The present section discusses the way the DWPT approximates the critical bands along with the constraints it encounters. Specifically, in Sect. 6.1 the suitability of five wavelet functions for the SV task is studied. Next, utilizing the most successful wavelet function, a number of DWPT trees corresponding to different approximations of the critical bands are constructed and their discrimination power is evaluated on the SV task. The most advantageous DWPT tree is further utilized in the computation of the proposed speech features. Finally, in Sect. 6.3 a recipe for computing the proposed speech features is offered.

6.1 Selection of wavelet function for WP tree design

As it was discussed in Sect. 5.2, the capability of the DWPT to provide various time-frequency representations depends on the wavelet function used and thus the corresponding

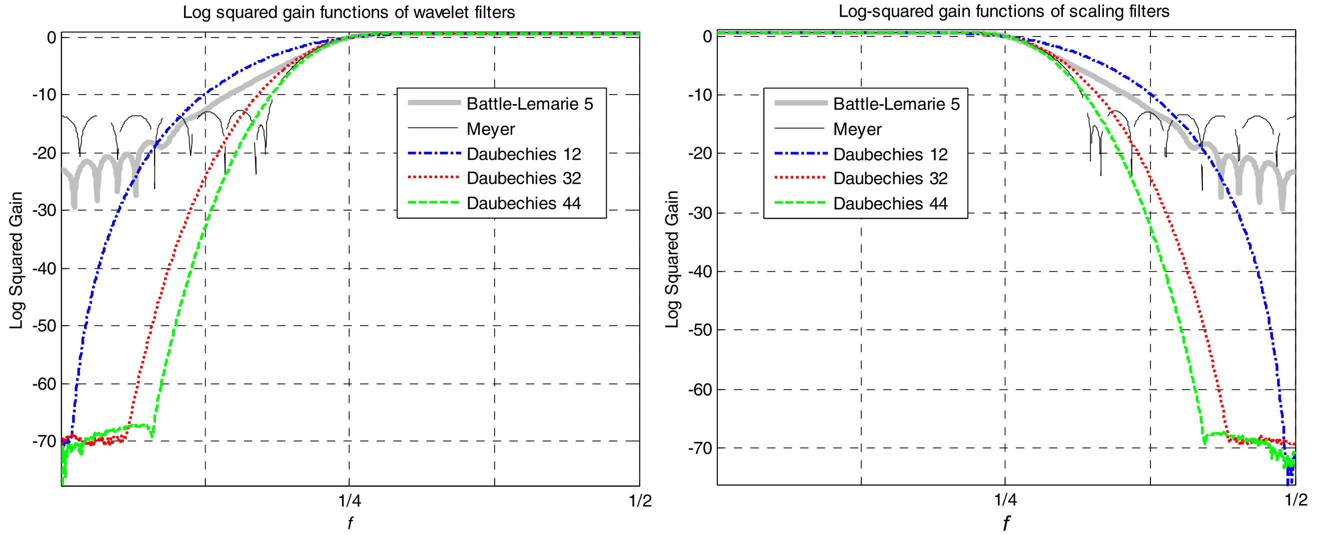


Fig. 4 The log squared gain functions of the wavelet and scaling filters for the: Battle-Lemarié, Daubechies, and discretized Meyer wavelets. The number after the name of the wavelet indicates the order. The ordinate shows the log squared gain in dB, and abscissa f is the normalized frequency

wavelet and scaling filters. Therefore, DWPT analysis can be further enhanced and fine-tuned by carefully selecting a wavelet function that is appropriate for the specific application. In the present work, the variety provided by the many existing wavelet families (Mallat 1998) is explored in order to augment the frequency localization abilities of the selected DWPT.

6.1.1 Initial considerations

Ideally, as in the case with DWT, it is desirable to have a time-frequency tiling that would coincide with the nominal tiling provided by the DWPT tree. This would require the frequency response of the wavelet filter and the scaling filter to be confined to $[1/4, 1/2]$ and $[0, 1/4]$, respectively. Although that is not practically achievable, the criterion for selecting a particular wavelet function and therefore, the corresponding wavelet and scaling filters, is based on the degree of proximity of their frequency responses to the ones of the ideal high-pass and low-pass filters, respectively. Therefore, the requirement to keep the leakage of energy to neighboring frequency bands as low as possible, and at the same time maximizing the amount of energy in the specified frequency band was set.

Comparing the frequency responses of the respective wavelet and scaling filters, three wavelet functions were initially studied: Daubechies' wavelet of order 44, Battle-Lemarié polynomial spline wavelet of order 5 and discretized Meyer wavelet. For reasons of comparison, Daubechies' wavelets of order 12 and 32 utilized in earlier related work (Farooq and Datta 2002; Sarikaya et al. 1998) were also included. A thorough description of all the aforementioned wavelets along with their corresponding wavelet and scaling filters is available in Mallat (1998). For comprehensiveness of our exposition, the frequency responses of the wavelet and scaling filters are illustrated in Fig. 4. As the figure presents, the frequency responses of the filters corresponding to Daubechies' wavelet of order 44, Battle-Lemarié polynomial spline wavelet of order 5, discretized Meyer wavelet and Daubechies' wavelet of order 32 have approximately the same characteristics in the interval $[0, -3]$ dB. On the contrary, the frequency responses of the filters corresponding to Daubechies' wavelet of order 12 are much different. As a consequence, we expect that all wavelets (except Daubechies' wavelet of order 12) would lead to a similar tiling of the time-frequency plane. A further purely theoretical analysis based only on the properties of these five wavelets was regarded inappropriate due to various factors, among which are the approximate character of critical bands and the non-deterministic nature of the speech signal. In addition, the considerable difference among the frequency responses outside the interval $[0, -3]$ dB might play, or might not play, a significant role for the wavelet performance in the specific setup. In view of this, an objective evaluation of the SV performance was performed in order to determine the most suitable amongst the five candidates. Details about the evaluation procedure and the ensuing experimental results are presented in the following subsection.

6.1.2 Objective evaluation of the candidate wavelet functions

Exploiting a common DWPT tree (which is presented in Sect. 6.2 under the provisional designation WP-0) as an ap-

Table 1 Experimental evaluation of five wavelet functions on the speaker verification task

Wavelet function	EER [%]
Battle-Lemarié, order 5	1.53
Discretized Meyer	1.87
Daubechies, order 12	2.49
Daubechies, order 32	2.10
Daubechies, order 44	2.25

proximation of the critical bands, five consecutive experiments were performed. Each time, a different DWPT that corresponds to one of the five wavelets under consideration (Battle-Lemarié polynomial spline wavelet of order 5, discretized Meyer wavelet, and Daubechies' wavelets of order 12, 32 and 44) was implemented. In these experiments the Polycost speaker recognition database described in Sect. 3.1 was employed.

In brief, the client models were trained by employing the English utterances from the first two sessions of each speaker, since a single session did not provide sufficient amount of training speech. In average, about 35 seconds of voiced speech were available for training each speaker model. In our setup, fifty male speakers were enrolled as authorized users and the remaining twenty-four speakers were considered as unknown to the system impostors. The reference model was built by exploiting the same speech material used for training the fifty user models. In total, 500 target and 36500 impostor trials were performed, with 10 target and 730 impostor trials per user model. Both unknown impostors and pseudo-impostors performed the fraud trials—in ten separate attempts per impostor. In average, about 1.3 seconds of voiced speech per test utterance were available. The actual amount of voiced speech in the particular trials ranged between 0.4 and 2.1 seconds.

Despite the observation that all candidate wavelet functions (with the exception of Daubechies' wavelet of order 12) possess similar characteristics in the frequency domain (refer to Fig. 4), the experimental results presented in Table 1 reveal that these wavelets have noticeably different functioning concerning the SV task. In fact, the Battle-Lemarié wavelet provides the lowest error rate exhibiting a considerable advantage over the Discretized Meyer wavelet and Daubechies' wavelets of order 32, 44, and 12. The authors deem the explanation behind this success of the Battle-Lemarié wavelet consists in its frequency-domain properties. Nevertheless, the contribution of its time-domain properties is not excluded either. Therefore, owing to its advantageous performance, the Battle-Lemarié wavelet was adopted as the basis wavelet function that provided the corresponding wavelet and scaling filters utilized in the formulation of DWPT, which is presented in the following subsection.

6.2 Wavelet packet tree design

In the present subsection, the DWPT based on the Battle-Lemarié of order 5 wavelet and scaling filters is employed for the approximation of the critical bands. The major reason for choosing DWPT instead of other candidate transforms (such as DFT) is the undeniable virtues of the DWPT, such as the flexible tiling of the time-frequency plane and its ability for fine-tuning of the time-frequency tiling. Exploiting these advantages, DWPT provides an efficient approximation of the critical bands in a natural manner by adapting its frequency resolution along the frequency axis.

Although from theoretical point of view we were aware about the advantages of the DWPT, we proceeded with assessing the practical worth of using either DFT- or DWPT-based analysis. Thus, the wavelet packet trees evaluated here were implemented in terms of both DFT and DWPT. The experimental evaluation is discussed later on in this subsection. In the following, we place emphasis upon the DWPT-based analysis.

While there is a lot of flexibility in choosing resolutions to cover the whole frequency range $[0, 4000]$ Hz, there are practical limitations concerning the quantity and numerical values of the available resolutions because of the limited length of the analyzed speech segment. When telephone speech sampled at 8 kHz is assumed, our desire to keep short-term stationarity of the speech signal imposes a segment length of about $20 \div 26$ milliseconds, which is in conflict with the requirement of the DWPT decomposition that the number of speech samples has to be equal to a power of two. Having in mind that we are going to exploit only the voiced speech, we consider a segment length of 32 milliseconds, which is equivalent to $256\ (= 2^8)$ samples. This size is a good trade-off allowing us to achieve a good resolution of the DWPT analysis while still preserving a reasonable stationarity of the signal in any particular voiced speech frame.

The maximum resolution achieved by the DWPT depends on the time length of the signal under analysis. Thus, for time signals of N samples, the maximum resolution of the DWPT is dictated by the maximum decomposition level, $j = \log_2(N)$. Therefore, the best resolution is $(1/2)^j F_N = 1/2^{j+1}$ where $F_N = 1/2$ is the Nyquist frequency. In our case $N = 256$, the maximum decomposition level is $j = 8$, and best resolution is $1/2^9 = 1/512$, while other potential resolutions belong to the set $\{1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4\}$. Considering a sampling frequency of 8 kHz, meaningful resolutions are $\{15.625$ Hz, 31.25 Hz, 62.5 Hz, 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz$\}$.

According to relationship (12), the ERB is an increasing function of frequency f. We consider the values of ERB that are equal to the resolutions provided by the DWPT. Thus, the exact frequencies at which ERB values change from

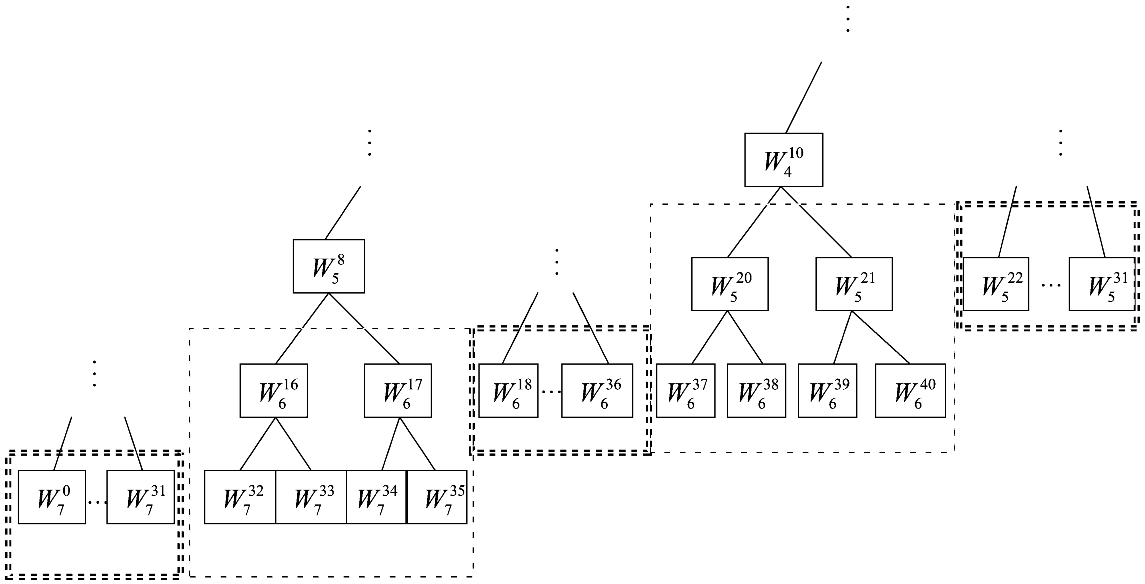


Fig. 5 A portion of the DWPT tree decomposition that illustrates the process of building all 16 candidate trees. *Dashed squares* present the parts of the tree that vary in frequency resolution. *Double dashed squares* present the parts that retain the same resolution for all the candidate trees

Table 2 DWPT resolutions suggested by ERBs and implemented with DWPTs frequency bands

Frequency Bands [Hz]		DWPT resolutions [Hz]	
Suggested by ERBs	Implemented with DWPTs	Tested candidates	Selected for the initial tree WP-0
60–350	62.5–375	15.625, 31.25	31.25
350–930	375–1000	31.25, 62.5	31.25
930–2090	1000–2250	62.5, 125	62.5
2090–4000	2250–4000	125, 250	125

one DWPT resolution to the next can be obtained. By solving (12) for f, we locate the precise frequencies that correspond to these specific values of ERB. Therefore, according to relation (12) the ERB values of $\{31.25$ Hz, 62.5 Hz, 125 Hz, 250 Hz, 500 Hz$\}$ correspond to frequencies $\{60$ Hz, 350 Hz, 930 Hz, 2090 Hz, 4400 Hz$\}$, respectively. Considering the approximate character of the ERB, the values computed for f constitute rough boundaries obtained during the selection of appropriate resolutions that correspond to various values of ERB.

In Fig. 5, the DWPT resolutions that were tested in the specific frequency bands of interest are presented. In the experiments, we systematically fluctuated around ERB values by varying the resolution a step higher or lower according to the DWPT structure. As Table 2 reveals, the frequency bands suggested by the ERBs were accommodated in a way that their boundaries become multiple of the values of both candidate resolutions. Thus, the DWPT trees are constructed for the frequency bands defined in the second column of Table 2, by utilizing the resolutions specified in the fourth column of the same table. The resolutions of the DWPT are 31.25 Hz, 62.5 Hz and 125 Hz in the frequency bands $[0, 1000]$ Hz, $[1000, 2250]$ Hz and $[2250, 4000]$ Hz, respectively. The tree constructed in this way is described by the set of vectors: $S_{init} = S_0 = \{\mathbf{W}_7^0 - \mathbf{W}_7^{31}, \mathbf{W}_6^{16} - \mathbf{W}_6^{35}, \mathbf{W}_5^{18} - \mathbf{W}_5^{31}\}$.

In each frequency band ($[62.5, 375]$, $[375, 1000]$, $[1000, 2250]$, $[2250, 4000]$ Hz) two resolutions were tested $(15.625, 31.25)$, $(31.25, 62.5)$, $(62.5, 125)$, and $(125, 250)$ Hz, respectively—one just above the CB value for the specific center frequency and the other below. Through extensive experimentation, we derived the following general conclusions: (a) starting from 0 Hz, and going up to 350 Hz, the most appropriate DWPT resolution is 31.25 Hz; (b) in the range from 350 Hz to 4000 Hz, the appropriate DWPT resolution is half the CB; or, in mathematical terms:

$$\textit{DWPT resolution} = \begin{cases} 31.25\ \text{Hz}, & \text{for } f \in [0, 350)\ \text{Hz}, \\ CB/2\ \text{Hz}, & \text{for } f \in [350, 4000]\ \text{Hz}. \end{cases} \tag{16}$$

Principally, the DWPT tree was designed according to the results summarized in Table 2. The approximate character of the critical bands as explained in Sect. 4, and in depth in Moore (2003), motivated us to test a variety of combinations at the points on the frequency axis where the resolution changes. For example, at frequency $f_1 = 1000$ Hz the DWPT resolution changes from 31.25 Hz to 62.5 Hz, and at frequency $f_2 = 2250$ Hz it changes from 62.5 Hz to 125 Hz. A set of 16 different WP trees, provisionally denoted as WP-0 to WP-15, was constructed each having a combination of different frequency resolutions in the frequency bands $[1000, 1250]$ Hz and $[2250, 2750]$ Hz. For example, the wavelet packet tree WP-9 is constructed with the frequency resolutions 31.25, 62.5, 125, and 62.5 Hz in frequency bands $[1000, 1125]$ Hz, $[1125, 1250]$ Hz, $[2250, 2500]$ Hz, and $[2500, 2750]$ Hz, respectively. Two of the sixteen trees, namely the WP-5 and WP-13, were empirically derived earlier and employed for the task of speaker verification in Siafarikas et al. (2004), Ganchev et al. (2004). In the present study, all potential solutions are explored in a systematic manner, and the best one among all WP trees is selected.

Table 3 describes the sixteen DWPT tree candidates that were compared in the experiment. Columns 2 to 8 contain the DWPT vectors used in each tree along with the different resolutions provided. The EER results for the DWPT are presented in column 9. Column 10 presents the results for the DFT-based filter-banks that have the corresponding frequency resolutions. As a whole, the DWPT-based features outperformed the corresponding DFT-based features. This is a consequence of the specific benefits of wavelet packet transform in comparison with the DFT, as those have been presented in Sect. 5.2. In summary, the superior performance of DWPT can be principally attributed to the following three reasons:

(a) The time-frequency localization of the DWPT, in contrast to the DFT which provides only frequency localization of a signal.
(b) The possibility of selecting among a wide range of wavelet functions which consist of the basis functions of the DWPTs providing the opportunity to select a wavelet that best suits the specific task under consideration. In contrast, the DFT relies only on a basis of sinusoidal functions.
(c) The possibility of selecting among a variety of different DWPTs especially designed for some particular task.

As the experimental results illustrate, the exact location of the switch-points, where the resolution changes from a finer to a coarser level e.g. the frequencies 1000 Hz and 2500 Hz, influences the SV performance. The objective evaluation performed here led us to the conclusion that the DWPT tree, designated as WP-2, provides the lowest overall EER among all DWPTs and DFTs. According to Table 3, WP-2 has the following resolutions: 31.25 Hz in the frequency band $[0, 1000]$ Hz, 62.5 Hz in the band $[1000, 2500]$ Hz and 125 Hz in the band $[2500, 4000]$ Hz, which turned out to provide advantage over the other candidates. Finally, according to Table 3, the WP-2 tree is defined as the subset of vectors $S_2 = \{\mathbf{W}_7^0 - \mathbf{W}_7^{31}, \mathbf{W}_6^{16} - \mathbf{W}_6^{39}, \mathbf{W}_5^{20} - \mathbf{W}_5^{31}\}$.

In view of the experimental results, the DWPT-based WP-2 is the most favorable time-frequency representation of the speech signal for the purpose of SV, and thereof is utilized in the speech parameterization proposed in the following subsection.

6.3 Computation of the proposed speech parameters

According to the experimental results presented in Sects. 6.1 and 6.2, the Battle-Lemarié wavelet of order 5 was found to be the most appropriate basis function, and the wavelet packet tree WP-2 was observed to provide the most suitable tiling of the time-frequency space as concerns the speaker verification task. Based on the DWPT decomposition that is being implemented using the wavelet packet tree WP-2 and the Battle-Lemarié wavelet function, we introduce a novel set of speech features, which is purposely designed for the tasks of speaker recognition. For simplicity and clarity, in the rest of our exposition, the proposed feature set is denoted as WP1-proposed, or most often just as WP1. The notation WP1, as well as the other notations of wavelet-based speech features (for example: WP2 and WP3) used in Sects. 7 and 8 are conventional, and thus, should not be perceived as derivative of the wavelet trees described in the previous subsection.

The signal pre-processing and feature computation steps for the new speech features were purposely kept coherent with the ones of the baseline MFCC (refer to Sect. 2.1.1). This facilitates a common experimental setup and a fair comparison of the results obtained for various speech parameterization schemes (Sect. 8). In summary, the computation of the proposed speech features, WP1, is performed as follows:

- The sampled at 8 kHz speech signal is filtered by a fifth-order Butterworth filter with pass-band $[80, 3800]$ Hz in order to remove any possible drift of the signal, and to reduce the effect of saturation by level, which might be present in speech.
- The pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ is employed.
- The discrete time speech signal is partitioned into overlapping segments of length 32 milliseconds ($N = 256$ speech samples). As discussed in Sect. 6.2, the segment's length is imposed by the restrictive assumption of the power of two length of the analysis window for DWPT, along with the desire to keep short-term stationarity of the speech segments. A 16 milliseconds skip rate (128 speech samples) provides a reasonable trade-off between continuity and computational efficiency. Due to the compact support of wavelets, no Hamming or other complex

Table 3 EER values for the 16 candidate trees WP-0 to WP-15. Each tree is characterized by a set of DWPT vectors, obtained in going horizontally from the name of the tree to the corresponding EER value. The numbers below the DWPT vectors indicate the resolution in Hz obtained with the specific vectors

WP tree	Frequency Bands [Hz]							EER [%]	
	[0, 1000]	[1000, 1125]	[1125, 1250]	[1250, 2250]	[2250, 2500]	[2500, 2750]	[2750, 4000]	DWPT	DFT
WP-0	$W_7^0 - W_7^{31}$ 31.25	$W_6^{16} - W_6^{17}$ 62.5	$W_6^{18} - W_6^{19}$ 62.5	$W_6^{20} - W_6^{35}$ 62.5	$W_5^{18} - W_5^{19}$ 125	$W_5^{20} - W_5^{21}$ 125	$W_5^{22} - W_5^{31}$ 125	1.53	1.87
WP-1						$W_6^{40} - W_6^{43}$ 62.5		1.45	1.78
WP-2					$W_6^{36} - W_6^{39}$ 62.5	$W_5^{20} - W_5^{21}$ 125		<u>**1.24**</u>	1.42
WP-3						$W_6^{40} - W_6^{43}$ 62.5		1.54	1.65
WP-4			$W_7^{36} - W_7^{39}$ 31.25		$W_5^{18} - W_5^{19}$ 125	$W_5^{20} - W_5^{21}$ 125		1.37	1.77
WP-5						$W_6^{40} - W_6^{43}$ 62.5		1.67	1.67
WP-6					$W_6^{36} - W_6^{39}$ 62.5	$W_5^{20} - W_5^{21}$ 125		1.33	1.55
WP-7						$W_6^{40} - W_6^{43}$ 62.5		1.44	1.55
WP-8		$W_7^{32} - W_7^{35}$ 31.25	$W_6^{18} - W_6^{19}$ 62.5		$W_5^{18} - W_5^{19}$ 125	$W_5^{20} - W_5^{21}$ 125		1.58	1.58
WP-9						$W_6^{40} - W_6^{43}$ 62.5		1.33	1.55
WP-10					$W_6^{36} - W_6^{39}$ 62.5	$W_5^{20} - W_5^{21}$ 125		1.35	1.55
WP-11						$W_6^{40} - W_6^{43}$ 62.5		1.31	1.69
WP-12			$W_7^{36} - W_7^{39}$ 31.25		$W_5^{18} - W_5^{19}$ 125	$W_5^{20} - W_5^{21}$ 125		1.55	1.76
WP-13						$W_6^{40} - W_6^{43}$ 62.5		1.33	1.52
WP-14					$W_6^{36} - W_6^{39}$ 62.5	$W_5^{20} - W_5^{21}$ 125		1.33	1.67
WP-15						$W_6^{40} - W_6^{43}$ 62.5		1.47	1.33

window is required, and therefore a rectangular one is implied.

- A voiced/unvoiced decision is obtained. In our experiments, the modified autocorrelation method with clipping (Rabiner et al. 1976) was used, but any reliable pitch estimation algorithm is applicable. Only voiced speech frames are used to represent the speakers' identity.
- DWPT with WP-2 tree and Battle-Lemarié wavelet function of order 5 is applied to the voiced speech frames. This DWPT provides a total of $B = 68$ frequency subbands. It was noticed that the first four subbands do not carry useful information (partially due to the band-limitation inherent to the telephone quality speech signal) and therefore, they were discarded.
- 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,

$$E_p = \frac{\sum_{i=1}^{N/2^j} (\mathbf{W}_j^k(i))^2}{N/2^j},$$

$$\mathbf{W}_j^k \in S_1, \;\; p = 1, 2, \ldots, B, \tag{17}$$

where $W_j^k(i)$ is the ith coefficient of the DWPT vector W_j^k.

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

$$F(i) = \sum_{p=1}^{B} \log_{10}(E_p) \cos\left(\frac{i(p-1/2)}{B}\right),$$
$$i = 0, 1, \ldots, r-1, \qquad (18)$$

where r is the number of feature parameters.

The full dimensionality of the feature vector $F(i)$ is sixty-four, i.e. $r = 64$. However, since most of the energy of the feature vector is carried by the first few coefficients, in many applications the first r coefficients alone might be sufficient as signal descriptors. To investigate the issue of sufficient dimensionality for the needs of speaker verification, we proceeded with computing the entire set of sixty-four coefficients and subsequently experimented with subsets of different size. The experimental setup and comparative results are discussed in Sect. 8.

7 Alternative DWPT-based speech parameterization schemes

Two alternative DWPT-based speech parameterization schemes, introduced by Farooq and Datta (2002) and Sarikaya et al. (1998) are briefly outlined here. In these studies, the Farooq-Datta's and Sarikaya's speech features were reported to outperform the baseline MFCC parameters on the speech and speaker recognition tasks, respectively. Here we are interested mostly in the frequency division utilized in these speech parameterization schemes.

7.1 The Farooq-Datta's wavelet packet-based speech features

Considering the phoneme recognition task, Farooq and Datta (2001, 2002) performed a wavelet packet decomposition of the frequency range [0, 8] kHz such that the obtained twenty-four frequency subbands closely follow the Mel scale. Following this decomposition, the phonemes were passed through the twenty-four wavelet packet filter-bank and the total energy E_p in the subband p was calculated as follows:

$$E_p = \sum_{j=1}^{N_p} (C_{j,p})^2, \quad p = 1, 2, \ldots, L, \qquad (19)$$

$$F_p = E_p / N_p, \quad p = 1, 2, \ldots, L, \qquad (20)$$

where $C_{j,p}$ is the jth coefficient in the pth subband, N is the number of wavelet coefficients in the pth subband and L is the number of subbands. The calculated energy is then divided by the number of wavelet coefficients in the corresponding subband thereby giving average energy per wavelet coefficients per subband F_p. In order to obtain features with emphasis on the lower frequency subbands, the authors selected Daubechies' wavelet filter of order 12. The logarithmically compressed subband energies obtained at the output of the filter-bank were decorrelated by applying the DCT and the first thirteen coefficients were kept as the feature set.

For the purpose of fair comparison with the other speech parameterization schemes evaluated in Sect. 8, certain modifications to the original settings proposed by Farooq and Datta were effected as follows:

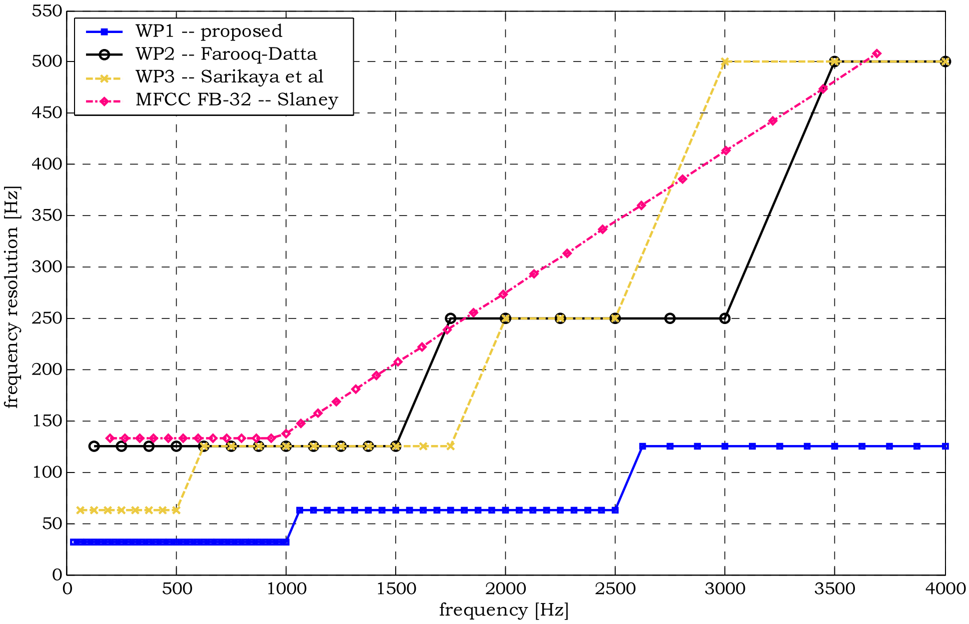
- Firstly, Daubechies' wavelet filter of order 12 employed in the original scheme was substituted with the Battle-Lemarié wavelet of order 5. As it was experimentally proved in Sect. 6.1, the Battle-Lemarié wavelet provides superior SV performance.
- Secondly, instead of the frequency range [0, 8] kHz utilized by the authors, in this work the frequency range [0, 4] kHz is considered. For the elimination of this discrepancy, the original wavelet decomposition, as it was described in Farooq and Datta (2002), is kept but we did confine it to [0, 4] kHz by discarding the upper four wavelet packet subbands, which cover the frequency range [4, 8] kHz. Thus, only the lowest twenty filters were retained from the original filter-bank. In Fig. 6, the frequency resolution of the twenty frequency subbands corresponding to Farooq and Datta's wavelet packet tree is presented with black solid line and 'o'-marks. For the purpose of comparison similar plots are presented for the other speech parameterization techniques under consideration here.
- Thirdly, after the decorrelation with the DCT, all twenty non-redundant coefficients were utilized as opposed to the first thirteen coefficients originally retained by Farooq and Datta. This allowed the evaluation of the different kinds of speech parameters for a common size of the feature vector.

The parameters obtained after these changes are referred to as Farooq-Datta's speech features, or simply WP2.

7.2 The Sarikaya's et al. wavelet packet-based speech features

In Sarikaya et al. (1998), considering the speaker identification problem, the authors performed a wavelet packet decomposition of the frequency range [0, 4] kHz such that the twenty-four frequency subbands obtained follow the Mel

Fig. 6 Frequency warping for the speech parameterization schemes under consideration. Each plot presents the frequency resolution in relation to the central frequency of the filters



scale. In brief, after some experimentation Sarikaya et al. found that the specific wavelet packet decomposition described in their work (Sarikaya et al. 1998) provided the best overall result among a reasonable number of wavelet packet trees. The proposed tree assigns more subbands between low to mid frequencies while keeping roughly a log-like distribution of the subbands across frequency. In Fig. 6, the frequency resolutions versus the corresponding frequency band for their tree is graphically illustrated with dashed light (yellow) line with 'x'-marks. Following the decomposition, the energy of each subband is computed and then scaled by the number of transform coefficients in that subband. The subband signal energies are computed for each frame as follows:

$$S_i = \frac{\sum_{m\in i}[(W_\psi x)(i), m]^2}{N_i}, \quad i = 1, 2, \dots, 24, \tag{21}$$

where $W_\psi x$ is the wavelet packet transform of signal x, i is the subband frequency index and N_i is the number of coefficients in the ith subband. Wavelet packet transform was implemented by using Daubechies' wavelet filter of order 32. The speech features, which Sarikaya et al. named Subband Based Cepstral coefficients (SBC), are derived from subband energies by applying the DCT transformation:

$$\mathrm{SBC}(n) = \sum_{i=1}^{L} \log(S_i) \cos\left(\frac{n(i-1/2)}{L}\pi\right)^2, \quad n = 0, 1, \dots, r-1, \tag{22}$$

where r is the number of SBC parameters and L is the total number of frequency bands.

For the purpose of fair comparison with the other speech parameterization schemes evaluated in Sect. 8, the following two modifications to the original settings proposed by Sarikaya et al. were effected:

- Firstly, Daubechies' wavelet of order 32 employed in the original approach of Sarikaya et al. was substituted with Battle-Lemarié wavelet of order 5.
- Secondly, only the first twenty coefficients were retained out of the total number of twenty-four, while Sarikaya et al. utilized the whole set of twenty-four coefficients.

The speech parameters obtained after these modifications are referred to as Sarikaya's features, or simply WP3.

7.3 Comparison among the various parameterization schemes

Despite their explicit dependence on the DWPT, the proposed speech parameters WP1 are considerably different from the related work (Farooq and Datta 2002) and (Sarikaya et al. 1998) both in their design philosophy and implementation. In brief, the main differences result from the design strategy that we utilized: Firstly, the underlying wavelet function was selected among others in an objective way aiming at the maximization of the SV performance. Secondly, the optimal frequency resolution in the various subbands has been selected in order to account for the recent advances in the theory of critical bands. Thirdly, the wavelet packet tree design was fine-tuned in a systematic way to provide frequency division that better suits for discrimination among human voices.

Aiming at a clearer illustration of the dissimilarities among the frequency warping schemes under consideration, in Fig. 6 we present a direct comparison of the frequency resolution in relation to the corresponding frequency band. As the figure presents, the general trend of all frequency warping schemes is to attribute the finest resolution to low frequency bands, whereas the coarsest resolutions are assigned to the higher frequency bands. The frequency resolution of the proposed frequency division is at all times finer (with a factor of two to four) than those of the other schemes. In fact, the frequency resolution of the proposed speech features, WP1, is in accordance with the relationship (16) which accounts for the recent advances in the theory of critical bands.

8 Experiments and results

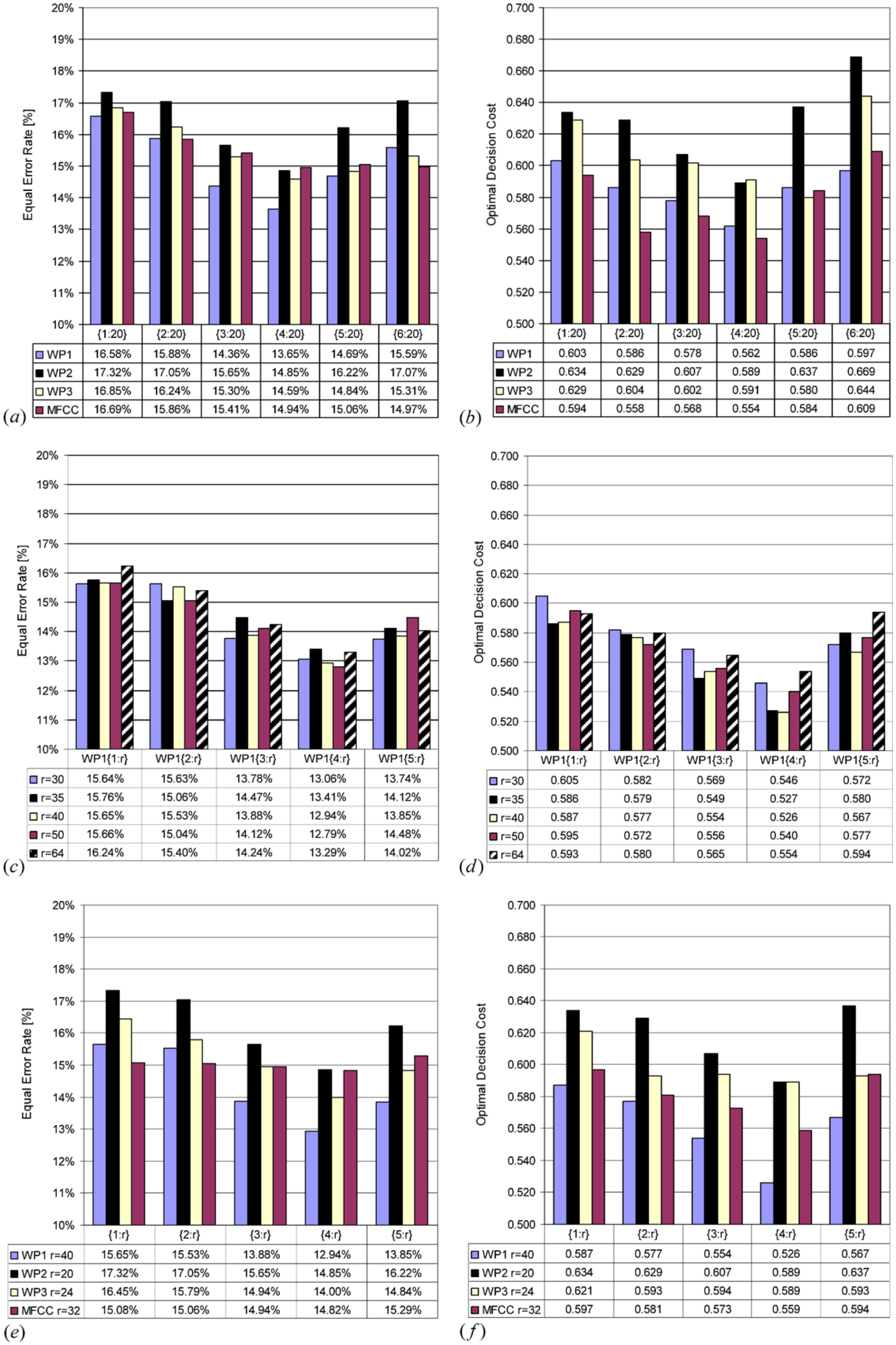
In a thorough evaluation, the proposed feature set WP1 was compared with the baseline MFCC and with two alternative DWPT-based speech features: WP2 and WP3. The proposed WP1 features were obtained as described in Sect. 6.3. The Sarikaya's and Farooq-Datta's features were computed following the methodology proposed by the corresponding authors but adapted to the needs of impartial comparison as presented in Sect. 7. In brief, since in the present work we focus on the approximation of the critical bands with wavelet packets, a direct comparison with other speech parameterization schemes is required. To this end, for all feature sets, each voiced speech frame is being decomposed with the DWPT in the frequency range $[0, 4]$ kHz using a common wavelet function (Battle-Lemarié of order 5) and the first twenty coefficients are retained. Therefore, features sets WP1, WP2 and WP3 differ solely in the wavelet packet tree and subsequently in the frequency warping along the frequency range $[0, 4]$ kHz. The computation of the baseline MFCC parameters is outlined in Sect. 2.1.1. This implementation of the MFCC parameters was found more successful for SV (details in Ganchev et al. 2005) than other MFCC implementations (Davis and Mermelstein 1980; Young 1993, etc.).

All validation experiments were performed on the NIST 2001 speaker recognition database which was described in Sect. 3. A common protocol was followed in all experiments according to the rules described in the 2001 NIST SRE Plan (NIST 2001). In brief, approximately 40 seconds of voiced speech were detected in the training data—a single two-minute recording per target speaker. The speech parameters computed for the voiced speech frames were utilized for building the clients' models. The common reference model was created by exploiting the male training speech available in the 2002 NIST SRE database (NIST 2002). Approximately one hour and forty minutes of voiced speech was available for that purpose. After the training, the user models were tested carrying out all male speech trials as defined in the complete one-speaker detection task (index file 'detect1.ndx'). Each SV experiment included 850 target and 8500 impostor trials with a duration from 0 to 60 seconds of speech, and employed all transmission channel types.

To this end, we performed two different experiments—the first one with uniform number of coefficients (twenty) for all speech parameterization techniques under consideration here, and the second one with the genuine set of speech features as they were proposed by the corresponding authors. These two experiments, as well as a supplementary one, which studies the performance of WP1 for various sizes of the feature vector, aim at providing a better understanding of the advantages of the evaluated parameterization techniques and the practical worth they offer.

For the purpose of fair comparison, in the first experiment we did alignment of the feature vector size to 20 parameters. Thus, only the first twenty parameters were kept for the feature vectors that have more coefficients. Figure 7 presents the results obtained for the evaluated speech parameterization techniques and various subsets of coefficients, in terms of EER and *DCFopt*, Fig. 7(a) and 7(b), respectively. As expected, reduction of the error rate was observed when the first coefficient (speaking about "the first" we refer to the coefficient with index "0"—equations (3), (18), (22)) was excluded from the feature vector. It is widely acknowledged that the value of the first coefficient in the cepstral-like features is sensitive to handset/communication channel mismatches between training and testing. This sensitivity is due to the fact that the first coefficient is related to the logarithm of the energy of the corresponding speech frame. Afterwards, we proceeded with an examination of the SV performance, when the second, third, fourth, and fifth coefficients are also excluded from the feature set. Surprisingly, an even more significant drop of the error rate was observed, for the cases when the second and third coefficients were discarded, along with the first one. This observation indirectly suggests that every mismatch between training and operational conditions (due to different handsets/transmission channels/environmental conditions, linguistic contents, etc.) notably affects not only the first coefficient but the first three ones. However, when coefficients beyond the third one were excluded from the feature set, higher error rates were observed.

As the experimental results presented in Fig. 7(a) and 7(b) suggest, the best subset for all features is $\{4{:}20\}$. The proposed speech features, WP1$\{4{:}20\}$ lead to the lowest EER among all subsets, and the MFCC$\{4{:}20\}$ to the lowest *DCFopt*. The WP2 and WP3 features outperform the MFCC in terms of EER, but are significantly inferior in terms of *DCFopt*. A more illustrative representation of the SV performance is presented in Fig. 8(a) where all results are mapped in the two-dimensional *DCFopt*—EER space. The lower left-hand part of the figure corresponds to the lowest



(*a*)

	{1:20}	{2:20}	{3:20}	{4:20}	{5:20}	{6:20}
🟪 WP1	16.58%	15.88%	14.36%	13.65%	14.69%	15.59%
■ WP2	17.32%	17.05%	15.65%	14.85%	16.22%	17.07%
□ WP3	16.85%	16.24%	15.30%	14.59%	14.84%	15.31%
🟥 MFCC	16.69%	15.86%	15.41%	14.94%	15.06%	14.97%

(*b*)

	{1:20}	{2:20}	{3:20}	{4:20}	{5:20}	{6:20}
🟪 WP1	0.603	0.586	0.578	0.562	0.586	0.597
■ WP2	0.634	0.629	0.607	0.589	0.637	0.669
□ WP3	0.629	0.604	0.602	0.591	0.580	0.644
🟥 MFCC	0.594	0.558	0.568	0.554	0.584	0.609

(*c*)

	WP1{1:r}	WP1{2:r}	WP1{3:r}	WP1{4:r}	WP1{5:r}
🟪 r=30	15.64%	15.63%	13.78%	13.06%	13.74%
■ r=35	15.76%	15.06%	14.47%	13.41%	14.12%
□ r=40	15.65%	15.53%	13.88%	12.94%	13.85%
🟥 r=50	15.66%	15.04%	14.12%	12.79%	14.48%
▨ r=64	16.24%	15.40%	14.24%	13.29%	14.02%

(*d*)

	WP1{1:r}	WP1{2:r}	WP1{3:r}	WP1{4:r}	WP1{5:r}
🟪 r=30	0.605	0.582	0.569	0.546	0.572
■ r=35	0.586	0.579	0.549	0.527	0.580
□ r=40	0.587	0.577	0.554	0.526	0.567
🟥 r=50	0.595	0.572	0.556	0.540	0.577
▨ r=64	0.593	0.580	0.565	0.554	0.594

(*e*)

	{1:r}	{2:r}	{3:r}	{4:r}	{5:r}
🟪 WP1 r=40	15.65%	15.53%	13.88%	12.94%	13.85%
■ WP2 r=20	17.32%	17.05%	15.65%	14.85%	16.22%
□ WP3 r=24	16.45%	15.79%	14.94%	14.00%	14.84%
🟥 MFCC r=32	15.08%	15.06%	14.94%	14.82%	15.29%

(*f*)

	{1:r}	{2:r}	{3:r}	{4:r}	{5:r}
🟪 WP1 r=40	0.587	0.577	0.554	0.526	0.567
■ WP2 r=20	0.634	0.629	0.607	0.589	0.637
□ WP3 r=24	0.621	0.593	0.594	0.589	0.593
🟥 MFCC r=32	0.597	0.581	0.573	0.559	0.594

Fig. 7 The speaker verification performance for various subsets of speech features in terms of Equal Error Rate (EER) in percentage and Optimal Decision Cost (*DCFopt*)

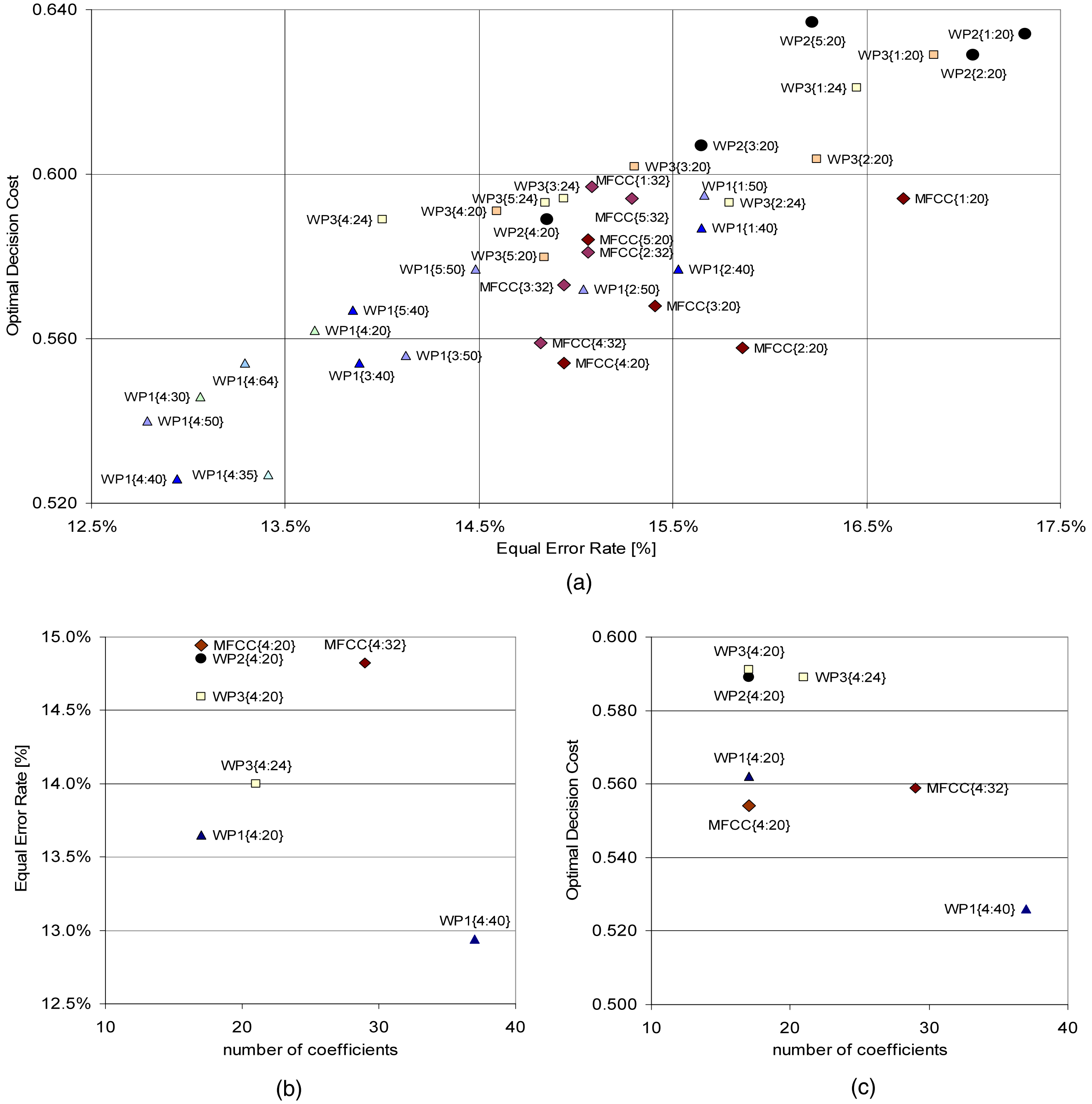


Fig. 8 Performance comparison: (**a**) mapping in the EER-*DCFopt* space, (**b**) mapping EER vs. number coefficients, and (**c**) *DCFopt* vs. number of coefficients

EER and decision cost *DCFopt,* and therefore to the best performance. The proposed features, WP1, which are plotted with triangles, demonstrate the best performance (set WP1{4:20}). Next, the Sarikaya's features, WP3, plotted with squares, follow with performance nearly equal to the one of the MFCC, plotted with rhombs. Finally, the Farooq-Datta's features, WP2, plotted with circles, exhibit the highest EER and decision cost.

In order to study the potential benefits of the larger number of coefficients that the speech features WP1 offer, we performed supplementary experiments with increased size of the feature vector. These experiments did not aim at identifying the best performance that can be achieved (on the NIST 2001 database) with a subset of WP1 features, but instead to study the general trend. Figure 7(c) and 7(d) present the EER and *DCFopt* for subsets of WP1 that include the first 30, 35, 40 and 50 coefficients, as well as the entire set of 64 coefficients. In the same manner as before we kept out the first coefficient and then the first two, three and four, coefficients from the feature vector. As the experimental results suggest, adding more coefficients to the feature vector is beneficial for the SV performance. However, for the very large subsets, which have more than 40 coefficients, the performance drops due to the curse of dimensionality. We deem that on a larger database that provides more training data even larger feature vectors will be beneficial.

In the second experiment, for which results are presented in Fig. 7(e) and 7(f), we evaluated the speech parameterization techniques under consideration, for the full set of coefficients. The only exception are the speech features WP1 for which we used only the first 40 parameters. (As we do not have evidence that the coefficients beyond the first 40 contribute to gaining a better SV performance, we do not include them in this experiment.) As the figures and the corresponding tables present, the significant advantage of the WP1 (subset WP1{4:40}), in terms of both EER and *DCFopt* is convincing. Next, the WP3 (subset WP3{4:24}) have the second best performance in terms of EER, however it is outperformed by the MFCC (subset MFCC{4:32}) in terms of decision cost *DCFopt*. Finally, the WP2 (subset WP2{4:20}) express the highest EER and decision cost *DCFopt*, and therefore offers the lowest SV performance.

In Fig. 8(b) and 8(c), the EER and *DCFopt* vs. the size of the feature vector for the evaluated speech features are presented, respectively. As the experimental results suggest, the lowest EER and the lowest decision cost *DCFopt*, i.e. the best performance, was obtained for the subset WP1{4:40}. The other feature sets express either higher EER or *DCFopt*. The changeable ranking in terms of EER and *DCFopt* for some subsets is due to the change in the slope of the DET performance plots (refer to Chap. 4, Fig. 4.9 in Ganchev 2005). The DET plots are not shown here due to space limitations, however, in Fig. 7 and 8 the effect from the change in the DET plots' slope is obvious. For instance, in Fig. 8(b) and 8(c), MFCC{4:20} and MFCC{4:32} interchange their ranks in terms of EER and *DCFopt*. The same holds for MFCC{4:20} and WP1{4:20}. Due to this phenomenon, by selecting specific subsets of features, one is able to trade EER vs. *DCFopt*, depending on the requirements of the specific SV application.

In brief, comparing the best subsets for all speech parameterization schemes (refer to the EER and *DCFopt* presented in the tables in Fig. 7(a) and 7(b), and in 7(e) and 7(f), as well as to the mapping in the *DCFopt*-EER space in Fig. 8(a)), we can conclude that the Farooq-Datta's features, WP2, exhibit the worst SV performance among the tested speech features, while the proposed features, WP1, present the best one. In terms of EER, the Sarikaya's features, WP3, perform slightly better than the MFCCs, but are entirely outperformed by the proposed features, WP1. For an equal size of the feature vector, a relative reduction of the EER (*DCFopt*) by 7%, 9%, and 9%, (5%, 5%, −1%) was observed for the proposed speech features, WP1, when compared to the WP3, WP2, and MFCC FB-32, respectively. Finally, for the best feature vectors of each type, the relative reduction of the EER (*DCFopt*) is 8%, 15%, 15%, (and 12%, 12%, 6%), respectively.

9 Conclusion

A novel set of wavelet packet-based speech features, appropriate for the task of speaker verification, was proposed. Our contribution is mainly in the wavelet-packet tree design, which roughly follows the critical band concept but is fine-tuned to provide frequency division that better suits the speaker recognition tasks. A comparative experimental evaluation of the proposed features performed on a well-known speaker recognition database, proved the practical significance of our approach. In particular, the proposed features demonstrated a superior performance when compared to other wavelet packet based features and to the Mel-scale cepstral coefficients. The superior performance of the proposed speech features is deemed to the reason that the (1) wavelet function, (2) design of the wavelet packet tree, (3) selection of frequency resolution were optimized in a systematic way to emphasize the dissimilarity between different voices. Finally, the proposed speech parameterization scheme offers the advantage of computing a larger number of relevant non-redundant parameters from a speech frame that further contributes for obtaining a better speaker verification performance.

References

Assaleh, K. T., & Mammone, R. J. (1994a). Robust cepstral features for speaker identification. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing*, (*ICASSP'94*) (Vol. 1, pp. 129–132). Adelaide, Australia.

Assaleh, K. T., & Mammone, R. J. (1994b). New LP-derived features for speaker identification. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 630–638.

Atal, B. S. (1974). 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.

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

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: SIAM.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, *28*(4), 357–366.

Erzin, E., Cetin, A. E., & Yardimci, Y. (1995). Subband analysis for speech recognition in the presence of car noise. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (*ICASSP-95*) (Vol. 1, pp. 417–420). Detroit, MI, USA.

Farooq, O., & Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letters*, *8*(7), 196–198.

Farooq, O., & Datta, S. (2002). Mel-scaled wavelet filter based features for noisy unvoiced phoneme recognition. In *Proceedings of the 7th international conference on spoken language processing* (*ICSLP 2002*) (pp. 1017–1020). Denver, Colorado, USA.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, *12*, 47–65.

Ganchev, T. (2005). Speaker recognition. Ph.D. dissertation, Dept. of Electrical and Computer Engineering, University of Patras, Greece, Nov. 2005.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2002a). Text-independent speaker verification based on Probabilistic Neural Networks. In *Proceedings of the acoustics 2002* (pp. 159–166). Patras, Greece.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2002b). A speaker verification system based on Probabilistic Neural Networks. In *2002 NIST speaker recognition evaluation, results CD workshop presentations & final release of results*, Vienna, Virginia, USA.

Ganchev, T., Siafarikas, M., & Fakotakis, N. (2004). *Speaker verification based on wavelet packets*. *Lecture notes in computer science*. Heidelberg: Springer. ISSN: 0302-9743, LNAI 3206/2004:299–306.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceedings of the 10th international conference on speech and computer, SPECOM 2005* (Vol. 1, pp. 191–194). October 17–19, 2005, Patras, Greece.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138.

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100–108.

Hennebert, J., Melin, H., Genoud, D., & Petrovska-Delacretaz, D. (1996). *The POLYCOST 250 Database (v1.0)*, COST250 report.

Hennebert, J., Melin, H., Petrovska, D., & Genoud, D. (2000). Polycost: a telephone-speech database for speaker recognition. *Speech Communication*, *31*(2–3), 265–270.

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

Long, C. J., & Datta, S. (1996). Wavelet based feature extraction for phoneme recognition. In *Proceedings of the 4th international conference on spoken language processing* (*ICSLP-96*) (Vol. 1, pp. 264–267). Philadelphia, USA.

Mallat, S. (1998). *A wavelet tour of signal processing*. San Diego: Academic Press.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing* (5th edn.). San Diego: Academic Press.

NIST (2001). The NIST year 2001 speaker recognition evaluation plan. National Institute of Standards and Technology of USA. Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf.

NIST (2002). The NIST year 2002 speaker recognition evaluation plan. National Institute of Standards and Technology of USA. Available: http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf.

Nogueira, W., Büchner, A., Lenarz, T., & Edler, B. (2005). A Psychoacoustic "NofM"-type speech coding strategy for cochlear implants. *EURASIP Journal on Applied Signal Processing—Special Issue on DSP in Hearing Aids and Cochlear Implants*, *18*, 3044–3059.

Nogueira, W., Giese, A., Edler, B., & Büchner, A. (2006). Wavelet packet filter-bank for speech processing strategies in cochlear implants. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (*ICASSP 2006*) (Vol. 5, pp. 121–124). Toulouse, France.

Oppenheim, A. V. (1969). A speech analysis-synthesis system based on homomorphic filtering. *Journal of the Acoustical Society of America*, *45*, 458–465.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals in Mathematical Statistics*, *33*, 1065–1076.

Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press.

Polycost Bugs (1999). A list of known bugs in version 1.0 of POLYCOST database. The Polycost Web-page. Available: http://circhp.epfl.ch/polycost/polybugs.htm.

Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., & McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *24*(5), 399–418.

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

Sarikaya, R., Pellom, B. L., & Hansen, J. H. L. (1998). Wavelet packet transform features with application to speaker identification. In *Proceedings of the IEEE nordic signal processing symposium: (NORSIG'98)* (pp. 81–84). Visgo, Denmark.

Siafarikas, M., Ganchev, T., & Fakotakis, N. (2004). Wavelet packets based speaker verification. In *Proceedings of the ISCA speaker and language recognition workshop—Odyssey 2004* (pp. 257–264). Toledo, Spain.

Slaney, M. (1998). *Auditory toolbox. Version 2* (Technical Report #1998-010). Interval Research Corporation.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, *3*(1), 109–118.

Tufekci, Z., & Gowdy, J. N. (2000). Feature extraction using discrete wavelet transform for speech recognition. In *Proceedings of the IEEE SoutheastCon 2000* (pp. 116–123). Nashville, Tennessee, USA.

Young, S. J. (1993). *The HTK hidden Markov model toolkit: design and philosophy* (Technical Report TR. 153). Department of Engineering, Cambridge University, UK.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, *33*, 248–249.