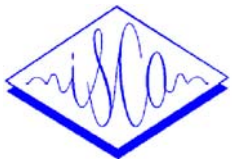
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Objective Wavelet Packet Features for Speaker Verification

Mihalis Siafarikas, Todor Ganchev, Nikos Fakotakis

Wire Communications Laboratory,
University of Patras, Rio-Patras 26500, Greece
{msiaf,tganchev,fakotaki}@wcl.ee.upatras.gr

Abstract

Studying ways for achieving a better demarcation of human voices for the task of speaker verification and taking advantage of the flexibility provided by wavelet packet analysis, we investigate in an objective way the relative importance of constituent disjoint frequency subbands of speech signals. Based on experimental results measuring the actual contribution of these subbands in relation to the corresponding frequency resolution, we propose a novel wavelet packet-based speech feature set that is effectively designed for speaker verification. The practical significance of our approach has been evaluated in comparative experiments performed on 2001 NIST Speaker Recognition Evaluation database. The proposed wavelet packet feature set has proven to outperform the widely used Mel-frequency scaled cepstral coefficients (MFCCs), as well as other wavelet packet based features that have been successfully used for speaker recognition.

10.21437/Interspeech.2004-532

1. Introduction

Although cepstrum-based speech features, such as MFCCs, are currently the most popular choice for speech recognition, it can be argued that they are not the best choice for speaker recognition given the dissimilar requirements for these two tasks. Clearly, speech recognition aims at capturing typical properties of phonemes, irrespective of any speaker-originated variations in the speaking style, while speaker recognition aims at exploiting these individualities. These distinctive traits of speaker recognition tasks motivated us to search for a more general approach (than the Mel-scale cepstral coefficients), which explicitly emphasizes the individuality of human voice.

As an alternative to the traditional Fourier Transform based techniques for analyzing time series, Wavelet Packet Transform (WPT) has been proven an effective signal-processing tool in a variety of speech processing applications. Specifically, WPT has been used in [1] and [2] as an alternative to Discrete Fourier Transform (DFT) for the purpose of speech recognition, closely approximating Mel-frequency scale. Furthermore, Sarikaya et al. [3] used the same features for speaker identification referring that they outperformed MFCCs. Other works do use the DFT analysis technique but instead of strictly adhering to the Mel-scale, they explore the discrimination ability of different frequency subbands [4]. Results from subjective speaker verification tests for a variety of frequency subbands were used for deriving of perceptually motivated features that were reported to outperform the Mel-frequency scale cepstral coefficients.

In the present work, the relative importance of 8 disjoint frequency subbands of speech signal is evaluated, in an objective manner, as concerns the speaker verification task, by using wavelet packet analysis. The results of this study are used to construct a wavelet packet tree that effectively represents the human voice individuality. Based on the proposed wavelet packet tree, a novel, fine-tuned for speaker verification speech features set, named Objective Wavelet Packet Features (OWPFs), was created.

Even though our work was inspired by [1], [2], [3], and [4], there are major differences between these earlier studies and our approach. Apart from using different wavelet and corresponding conjugate mirror filters, the authors of [1], [2], and [3] have used wavelet packets for approximating the Mel-frequency scale, while we utilize the wavelet packets for achieving a more effective division of the frequency scale that provides a better speaker verification performance. According to the experimental results measuring the relative importance of disjoint frequency subbands, a novel wavelet packets tree was proposed, and a novel speech features set corresponding to that tree was derived.

The principal difference between [4] and our work is that we do not depend on subjective speaker verification tests but on the contrary, we perform an objective study of the relative importance of the disjoint frequency subbands. Besides the aforementioned difference, we do not confine to frequency analysis using DFT that provides a fixed tiling of the time-frequency plane and thus requires further weighing of the frequency subbands to facilitate a perceptually based frequency scale. Instead, we exploited the flexibility provided by wavelet packets as concerns the tiling of the time-frequency plane in order to follow the discrimination performance of different subbands according to results of our objective study. These results guided us to attribute different resolutions to separate frequency subbands building accordingly a wavelet packet tree that provided the proposed features set, OWPFs.

As the experimental results presented in Section 5 demonstrate, the OWPFs proposed here outperform MFCCs, as well as the wavelet packet features presented in [1], [2], and [3].

2. Wavelet Packet Bases

Wavelet packet functions generalize the filter bank tree that relates wavelets and conjugate mirror filters [9] leading to a flexible division of the frequency axis into intervals of various bandwidths. In the decomposition of a signal with the WPT, either the low or the high frequency band is decomposed resulting in a balanced tree structure, as shown in Figure 1. To each node of the tree, a wavelet packet space W_j^p is associated, where j is the depth, and p is the number of nodes to the left of this particular node at the given depth. Supposing that this space admits an orthonormal basis $B_j^p = \left\{\psi_j^p\left(t-2^j n\right)\right\}_{n\in\mathbb{Z}}$, the following splitting relations [9] define the wavelet packet orthogonal bases at the children nodes:

$$\psi_{j+1}^{2p}(t) = \sum_{n=-\infty}^{+\infty} h[n]\,\psi_j^p\left(t-2^j n\right) \qquad (1)$$

and

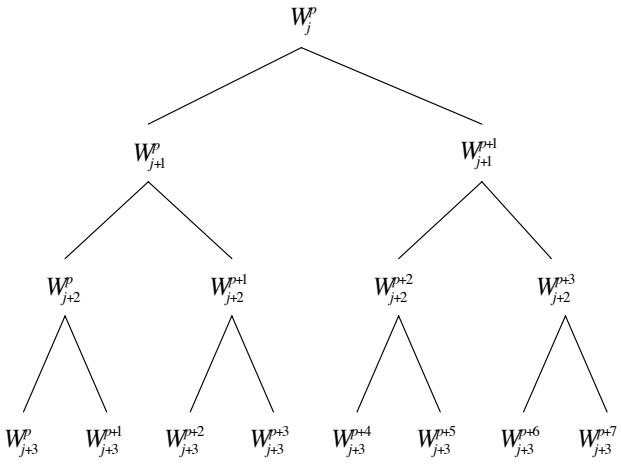


Figure 1: Binary tree of wavelet packet spaces

$$\psi_{j+1}^{2p+1}(t) = \sum_{n=-\infty}^{+\infty} g[n]\psi_j^p(t - 2^j n) \quad (2)$$

where h and g are a pair of conjugate mirror filters. As concerns the frequency localization of wavelet packets, the Fourier transforms of wavelet packet children are related to their parent by the following relations [9]:

$$\hat{\psi}_{j+1}^{2p}(\omega) = \hat{h}(2^j\omega)\hat{\psi}_j^p(\omega) \quad (3)$$

and

$$\hat{\psi}_{j+1}^{2p+1}(\omega) = \hat{g}(2^j\omega)\hat{\psi}_j^p(\omega) \quad (4)$$

where $\hat{h}$ and $\hat{g}$ are the DFT of the discrete filters h and g, respectively.

The major part of the energy of $\hat{\psi}_j^p$ is concentrated over a particular frequency band. The two filters $\hat{h}(2^j\omega)$ and $\hat{g}(2^j\omega)$ select the lower and higher frequency components within this band, leading to the corresponding energy concentration of $\hat{\psi}_{j+1}^{2p}$ and $\hat{\psi}_{j+1}^{2p+1}$.

3. Relative importance of frequency subbands

In order to determine the maximum frequency resolution necessary to capture the speaker identity, we took into consideration the concept of critical bandwidth introduced by Fletcher in [5]. Zwicker in [6] estimated that the critical bandwidth is constant at 100 Hz for centre frequencies up to 500 Hz, while for higher frequencies the bandwidth increases approximately in proportion with centre frequency. However, more recent experiments [7] have provided evidence that the critical bandwidth can be as narrow as 30 Hz for frequencies below 500 Hz. The last led us to construct a wavelet packet tree with a maximum frequency resolution of 31.25 Hz.

With the intention to study the significance of each frequency subband for speakers' voices differentiation, we performed objective speaker verification experiments utilizing the Polycost speaker recognition corpus [10]. In total, speech from all the 74 male speakers were employed, each represented with his first three sessions, and with ten sentences per session. The full frequency range of 0 ÷ 4 kHz of the speech signal is separated into eight disjoint subbands of 500 Hz each. These subbands and their corresponding center frequencies are shown in the first two columns of Table 1. The bandwidth of 500 Hz was chosen with the aim of obtaining a smoothed estimation of the relative importance of the corresponding subband avoiding the threshold effect when dealing with smaller sections.

Instead of performing subjective speaker verification tests, as accomplished in [4], we oppositely evaluated the speaker verification performance by employing the automatic speaker verification system described in [12]. However, in contrast to [12], where MFCCs speech features are considered, here, speaker-specific information available in each tested frequency subband was analyzed by using wavelet packets. This was achieved by taking the WPT at the depth j=7 corresponding to a frequency resolution of 31.25 Hz, and selecting the coefficients representing this particular frequency band, as shown in Figure 1. Speech signal was preprocessed as described in Section 4. The relative importance of spectral information in each of the subbands, shown in column 1 of Table 1, was evaluated in the sense of Equal Error Rate (EER) for the above mentioned frequency resolution of 31.25 Hz, with the results presented in the third column of the same table.

Table 1: The EER for the 8 disjoint frequency subbands along with the proposed resolutions

Frequency subband [Hz]	Center freq. [Hz]	EER at resolution 31.25 Hz [%]	Proposed resolution [Hz]
0-500	250	17.59	31.25
500-1000	750	16.09	31.25
1000-1500	1250	21.38	62.5
1500-2000	1750	26.22	125
2000-2500	2250	23.03	62.5
2500-3000	2750	26.08	125
3000-3500	3250	23.27	62.5
3500-4000	3750	26.16	125

As it can be observed from the above Table 1, the smaller EER is attained in the lowest two frequency bands {0 ÷ 500 Hz and 500 ÷ 1000 Hz} which means that most speaker-specific information is revealed in these two frequency bands, as expected. The frequency bands {1000 ÷ 1500 Hz, 2000 ÷ 2500 Hz, and 3000 ÷ 3500 Hz} have lower EER and thus appear to contain more speaker-specific information than frequency bands {1500 ÷ 2000 Hz, 2500 ÷ 3000 Hz, and 3500 ÷ 4000 Hz}, which demonstrated the worst results.

Apart from revealing useful speaker-specific information of the different subbands, our approach exploits these results in a constructive way to improve the wavelet packet analysis. Having in mind that the wavelet packet nodes at depths 7, 6, and 5, result in 16, 8, and 4 frequency intervals of width 31.25, 62.5 Hz, and 125 Hz, respectively, and considering the analysis of the speaker verification performance of each frequency subband, we built a wavelet packet tree as follows: The frequency bands were clustered into groups according to the observed EER, i.e., {0 ÷ 500 Hz and 500 ÷ 1000 Hz}, {1000 ÷ 1500 Hz, 2000 ÷ 2500 Hz, and 3000 ÷ 3500 Hz} and {1500 ÷ 2000 Hz, 2500 ÷ 3000 Hz, and 3500 ÷ 4000 Hz}. Due to the different relative importance measured in the experiments, these groups should not contribute equally into the final vector of features. Thus, we propose their frequency reso-

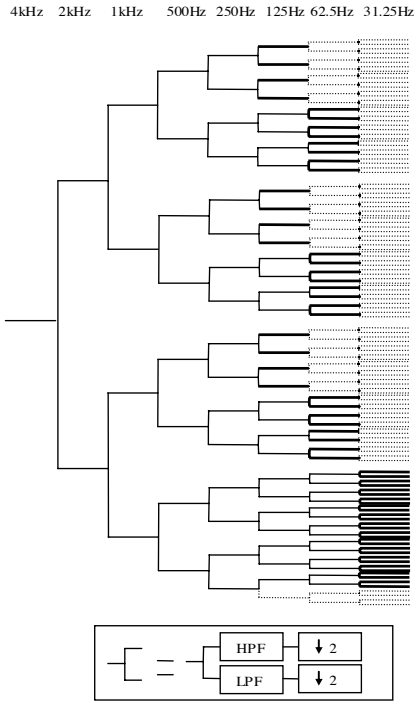


Figure 2: The wavelet-packet tree for the proposed features. Only the frequency bands drawn with solid bold line are utilized in the computation of the features.

lutions to rely on their relative importance as it is shown in the last column of Table 1. This means that in the feature extraction stage, presented in the next Section 4, 16 coefficients are taken into account for each of the frequency subbands in the first group, 8 coefficients for the frequency subbands in the second group, and 4 coefficients for the frequency subbands in the third group. This, in turn, results in the representation of the frequency subbands according to their EER: more coefficients for the lowest EER, less coefficients for higher EER and even less coefficients for the highest EER. That step is referred as to 'Objective subband energy estimation' into the block diagram presented in Figure 3. Due to their lowest EER, and thus a relatively higher importance, the lowest two frequency subbands receive the highest frequency resolution of 31.25 Hz. The subbands with EER between 21% and 24% receive a frequency resolution of 62.5 Hz, and the subbands with EER higher than 24% receive the lowest resolution of 125 Hz.

4. Feature Extraction

Based on the frequency resolutions defined in Section 3, we constructed a wavelet packet tree that renders an account of the relative importance of each frequency subband. The lowest four bins of width 31.25 Hz, covering the frequency range $0 \div 125$ Hz were excluded from the wavelet tree because that frequency interval is very much affected by the characteristics of transmission channels. Thus, the wavelet packet tree shown in Figure 2 was constructed providing a total of B=64 frequency subbands.

The proposed speech features, OWPFs, are computed as depicted in the block diagram of Figure 3: The speech signal is filtered by a fifth order Butterworth filter with pass-band from 80 Hz to 3800 Hz, followed by framing in intervals of 32 milliseconds, with a skip rate of 16 milliseconds. Due to

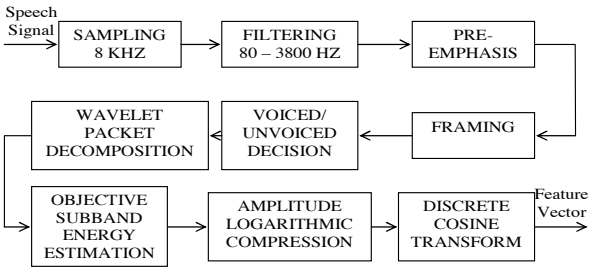


Figure 3: Schematic representation of the proposed wavelet packet features (OWPFs) computation.

the compact support of wavelets, no Hamming or other complex window is required, and therefore a rectangular one is considered. A pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ is employed. A voiced / unvoiced frame decision is obtained using a pitch estimation based on the modified autocorrelation method with clipping [8]. Only those feature vectors representing voiced speech frames are further used to represent the speaker's identity. Next, wavelet packet decomposition is applied at a maximum depth of j=7, corresponding to a frequency resolution of 31.25 Hz in accordance to the wavelet tree referred to in Section 3. In order to avoid appearance of false large amplitude coefficients at the margins of every speech frame, boundary wavelets were utilized in the computation of the WPT. The normalized energy in each frequency band is computed as:

$$E_j = \frac{\sum_{i=1}^{N_j} [W_j^p f(i)]^2}{N_j}, j = 1,...,B \qquad (5)$$

where $W_j^p f(i)$ is the i-th coefficient of the wavelet packet transform of a signal f at node W_j^p of the wavelet packet, B is the total number of nodes used, and N_j is the total number of coefficients consisting node j.

Finally, logarithmic compression is performed and Discrete Cosine Transformation is applied on the logarithmic subband energies to decorrelate the parameters, allowing reduction of the dimensionality of feature vectors:

$$F(i) = \sum_{n=1}^{B} \log_{10} E_n \cos(\frac{i(n-1/2)}{B}), i = 1,...,r\ , \qquad (6)$$

where r is the number of feature parameters. In our case, we consider r=*35*, since these coefficients represent 99.99% of the energy of the complete set of 64 subbands.

5. Experiments and Results

The speaker verification system [12] was used as a platform to evaluate several sets of speech features. In the comparative experiments, we have utilized the male part of the 2001 NIST SRE corpus [11]. Approximately 40 seconds of voiced speech for training of each user model were available. The common reference model was built from one hour and forty minutes of voiced speech, exploiting the male training data offered in the 2002 NIST SRE database [13]. Each speaker verification experiment included 850 target and 8500 impostor trials with lengths between 0 and 60 seconds, covering the entire diversity of transmission channel types, defined in the complete one speaker detection task [11].

In the experiments with the MFCC features, we adhered to an approximation of the Mel-scale with 32 filters, covering the frequency diapason 133 ÷ 3954 Hz. It was found that MFCC computed from 32 filters are the most successful for speaker verification applications, when compared to other implementations – with a filter-bank of 13, or 20 filters. The MFCCs were computed as described in [12]. The Farooq-Datta's features and the Sarikaya's features were estimated by following the methodology of the corresponding author: [2] and [1], respectively. The proposed wavelet packet based features set, OWPFs, has been computed as described in Section 4.

Table 2 reveals a comparison among the wavelet packet based features and MFCCs. Column 2 of Table 2 presents the actual coefficients for each type of speech features, included in the corresponding experiment, while column 3 provides the obtained EERs. The best set for each kind is denoted by the symbol "*". In the case of MFCCs, excluding the first cepstral coefficient of the feature vectors reduced the influence of transmission channel mismatch between train and test, and thus decreased the EER. It was found that removing the next cepstral coefficients after the first one deteriorates the speaker verification performance. However, for all wavelet packets based features studied here, namely OWPFs, WP2, and WP3, it was observed that excluding the first three coefficients from the feature vectors leads to significant reduction of the EER.

Table 2: The EER in percentage for the evaluated feature sets

Speech Features	Set of Coef.	EER [%]
OWPF – proposed	{1,2,…,35}	17.51
OWPF – proposed	{2,3,…,35}	16.44
OWPF – proposed	{3,4,…,35}	14.99
*** OWPF – proposed**	**{4,5,…,35}**	**13.98**
WP2 – Sarikaya	{1,2,…,24}	18.15
WP2 – Sarikaya	{2,3,…,24}	17.14
WP2 – Sarikaya	{3,4,…,24}	15.96
*** WP2 – Sarikaya**	**{4,5,…,24}**	**15.39**
MFCC FB-32	{1,2,…,32}	17.45
*** MFCC FB-32**	**{2,3,…,32}**	**16.35**
WP3 – Farooq-Datta	{1,2,…,13}	19.61
WP3 – Farooq-Datta	{2,3,…,13}	19.08
WP3 – Farooq-Datta	{3,4,…,13}	17.74
*** WP3 – Farooq-Datta**	**{4,5,…,13}**	**16.98**

When we compare the best members of each kind of speech features, shown in Table 2, the Farooq-Datta's set, WP3, exhibits the highest error rate, while the proposed set, OWPF, expresses the lowest one. The Sarikaya's features, WP2, were confirmed to perform better than the MFCCs, but are outperformed by the proposed features, OWPFs.

In conclusion, we can generalize that the experimental results confirmed the advantage of proposed speech features, OWPFs, which were optimized in a systematic way to emphasize human voice individuality, and by that reason are able to provide a better speaker verification performance than the MFCCs, and the Farooq-Datta's and Sarikaya's wavelet packets based speech feature sets.

6. Conclusion

A novel, wavelet packet based speech features set, appropriate for speaker verification, was proposed. Our contribution is mainly in the wavelet packet tree design that was constructed according to results from an objective study of the relative importance of each frequency subband in demarcation of human voices. Thus, the speech feature set we propose is fine-tuned to emphasize the relatively more important spectral subbands for voice differentiation. A comparative experimental evaluation of the proposed features, performed on a well-known speaker recognition corpus, proved the practical significance of our approach. The proposed speech features demonstrated a superior performance, when contrasted to other wavelet packet-based features and to the Mel-frequency scaled cepstral coefficients, due to a better representation of the speaker-specific variations of the speech signals.

Acknowledgements

This work was supported by the "Infotainment management with Speech Interaction via Remote microphones and telephone interfaces" - INSPIRE project (IST-2001-32746).

References

[1] Sarikaya, R., Hansen, H.L., "High resolution speech feature parameterization for monophone-based stressed speech recognition", *IEEE Signal Processing Letters*, vol.7,no.7, 2000.

[2] Farooq, O., Datta S., "Mel-scaled wavelet filter based features for noisy unvoiced phoneme recognition", *Proc. of ICSLP 2002*, Denver, Colorado, USA, pp.1017-1020, Sept.16-20, 2002.

[3] Sarikaya, R., Pellom, B.L., Hansen H.L., "Wavelet packet transform features with application to speaker identification", *Proc. of IEEE Nordic Signal Processing Symp.*, Visgo, Denmark, pp. 81-84, June 1998.

[4] Orman, O.D., Arslan, L.M., "Frequency Analysis of Speaker Identification", *Proc. of The Speaker Recognition Workshop 2001: A Speaker Odyssey,* Crete, Greece, pp.219-222, June18-22, 2001.

[5] Fletcher, H., "Auditory patterns", *Reviews of Modern Physics*, no. 12, pp. 47-65, 1940.

[6] Zwicker, E., "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", *The J. of Acoustical Society of America*, vol.33, pp.248-249, 1961.

[7] Moore, B.C.J., "An introduction to the psychology of hearing", *Academic Press*, London, 5th ed., 2003

[8] Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A., "A Comparative Performance Study of Several Pitch Detection Algorithms", *IEEE Transactions on ASSP*, Vol. ASSP-24, No.5. pp. 399-418, 1976

[9] Mallat S., "A wavelet tour of signal processing", *Academic Press*, San Diego, USA, 1998.

[10] Petrovska, D., Hennebert, J., Melin, H., Genoud, D., "Polycost: A Telephone-Speech Database for Speaker Recognition," *Speech Communication*, vol.31, no.2-3, pp.265-270, June 2000.

[11] "The NIST Year 2001 Speaker Recognition Evaluation Plan". The NIST of USA, 2001. Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf

[12] Ganchev, T., Fakotakis, N., Kokkinakis, G., "Text-Independent Speaker verification Based on Probabilistic Neural Networks", *Proc. of Acoustics 2002*, Patras, Greece pp.159-166, 2002

[13] "The NIST Year 2002 Speaker Recognition Evaluation Plan", The NIST of USA, 2002. Available: http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf