

Speaker Characterization Using Principal Component Analysis and Wavelet Transform for Speaker Verification

C. Tadj, A. Benlahouar

École de Technologie Supérieure - Electrical Engineering Department
1100, Notre-Dame Ouest, Montreal, Qc, H3C 1K3 Canada
ctadj@ele.etsmtl.ca

Abstract

In this paper, we investigate the use of the Wavelet Transform for text-dependent and text-independent Speaker Verification tasks. We have introduced a Principal Component Analysis based wavelet transform to perform frequencies segmentation with levels decomposition. A speaker dependent library tree has been built, corresponding to the best structure for a given speaker. The constructed tree is abstract and specific to every single speaker. Therefore the extracted parameters are more discriminative and appropriate for speaker verification applications. It has been compared to MFCC's and other wavelet-based parameters. Experiments have been conducted using corpus, extracted from Yoho and Spidre Databases. This technique has shown robustness and 100% efficiency in both cases.

Keywords: Principal Component Analysis, Wavelet Transform, Frequencies Segmentation, Levels Decomposition, Abstract Tree, Speaker Verification.

1. Introduction

Mel-Frequency Cepstral Coefficients (MFCC) have been the most widely used speech features for speech and speaker recognition in the last decades. The reader can refer to the survey in [1] for a complete introduction to speaker recognition and the use of MFCC features.

However, the use of Discrete Cosine Transform (DCT) of mel-scaled log filter bank energies in the MFCC based features extraction has some drawbacks [2]. This is due to the fact that DCT covers all frequency bands and therefore the corruption of a frequency band of speech by noise affects all MFCC. Subband speech recognition has been used to overcome the problems associated with MFCC-based recognizers [3].

Some extensive work has been conducted using Wavelet Transform (WT) for the extraction of the features for speech recognition [4]. The main result of these studies is the use of Wavelet Transform to overcome the drawbacks of the subband-based recognizers. The use of the WT, which has a good time and frequency resolution instead of the DCT, should solve the problem mentioned above.

In this paper we are interested in the use of the WT for Text-Dependent Speaker Verification (TDSV) and Text-Independent Speaker Verification (TISV) tasks. We will investigate how well the multi-resolution WT can capture the variation of the speech as well as the speaker. An important step in these tasks is to extract sufficient information for a better discrimination between speakers. This process is performed by applying the well-known Principal Component

Analysis (PCA) technique at different levels of the Wavelet Packet decomposition tree.

In the next section, we present an overview and some background for PCA, Fourier and Wavelet Transforms. Section 3 presents our main contribution. It describes the step-by-step construction principle of the *Best Structure Abstract Tree* (BSAT) according to the energy criterion. Extensive experiments are shown in section 4. The results are then discussed and analyzed.

2. Related Work: Background and Review

2.1. Principal Component Analysis

Principal Component Analysis (PCA) [5], is widely used in signal processing, statistics, and neural computing. The basic idea in PCA is to find the components s_1, s_2, \dots, s_n that explain the maximum amount possible of variance by n linearly transformed components.

The basic goal in PCA is to reduce the dimension of the data. It can be proven that the representation given by PCA is an optimal linear dimension reduction technique in the mean-square sense. Such a reduction in dimension has important benefits. First, the computational overhead of the subsequent processing stages is reduced. Second, noise may be reduced, as the data not contained in the n first components may be mostly due to noise. Third, a projection into a subspace of a very low dimension, for example two, is useful for visualizing the data.

2.2. MFCC

The MFCCs are computed by taking the DCT of Mel-scaled log filter bank energies:

$$MFCC_i = \sum x(j) \cos\left(\frac{\pi i}{N} \left(j + \frac{1}{2}\right)\right), \quad i = 1, \dots, M \quad (1)$$

where N is the number of the filter banks and $x(j)$ represents the log-energy of the j^{th} filter.

2.3. Wavelet Transform

The wavelet transform permits the use of short time-windows at high frequencies. This property permits to obtain signal representations with good resolutions in both the frequency domain and the time domain [6]. A wavelet can be seen as band pass filter and the set of its dilated version as a bank of constant filters. To avoid the need for an infinite number of such filters in a given signal representation, a scaling function

can be used. The analysis of a signal with wavelets and a scaling function permits its expression in terms of the wavelets up to a given scale.

The choice of the best tree for a given application can be performed using the wavelet packet approach.

2.4. MFDWC

Mel Frequency Discrete Wavelet Coefficients (MFDWC) give a new feature vector for speaker verification. This method is used by applying Wavelet Packet Transforms (WPT). An admissible wavelet packet tree is proposed by [4] giving a new filter banks structure in which filters have frequency bands spacing approximating the Mel scale used in MFCC, as shown in Figure 1. The purpose of using the WPT is to benefit from its localization in the time and frequency domains [2].

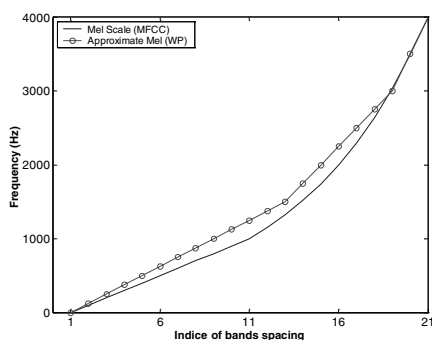


Figure 1: Partitioning of the frequency axis of Mel scale using MFCC and its approximation using MFDWC.

The property of partitioning the frequency axis using WPT in MFDWC is obtained by using recursively a pair of conjugate mirror filters [7], which divides the frequency band into two equal halves at each given scale.

Each scale in the wavelet packet tree is indexed by its depth j and number of subspaces p below it. The two orthogonal bases at each parent node (j,p) are defined by a low pass filter and a high pass filter as in [4].

The features are extracted from the different nodes (j,p) of the wavelet packet tree. The bands spacing of this nodes corresponds exactly to the frequency bands, which are achieved by the approximate Mel scale as shown in Figure 1. The MFDWC coefficients are computed by taking the DCT of log-energy of each nodes (j,p) in the wavelet packet tree. Each couple of filters (low pass filter, high pass filter) in the WP tree is associated with the biorthogonal Daubechies compactly supported wavelets with N vanishing moments [8].

2.5. Selecting the Best Basis

The best basis selection (BBS) method is usually used for signal compression applications. We have applied it successfully for features selection in order to extract the most significant information from the speech signal with a minimum cost [9], the cost being a selected entropy function. This has been done by searching the best basis in which that signal is best represented. The main idea proposed by [10] is to build a library of orthonormal basis relative to a given signal or collection of signals which has the lowest information cost. The best basis relative to the best WPT

binary tree is obtained by eliminating branches according to the selected entropy cost function.

3. Proposed Scheme for Speaker Verification

3.1. Best Structure Abstract Tree (BSAT)

We propose an algorithm that shows a step-by-step construction of the *Best Structure Abstract Tree* (BSAT) according to the energy criterion. This algorithm allows the extraction of the characteristics of the speaker. The BSAT algorithm performs as follows:

1. Preprocess the audio signal.
2. Perform complete wavelet decomposition to each temporal analysis window binary tree for a given number of levels.
3. Compute the parameters using the energy criterion.
4. Connect all these numerical values to obtain a specific level-based tree structure for each speaker.
5. Apply PCA for a given rate of variance and generate a Speaker dependent Abstract tree by Level (SAL). More details are given in the next paragraph.

The general design, which connects all the blocks used in this algorithm, is illustrated in Figure 2.

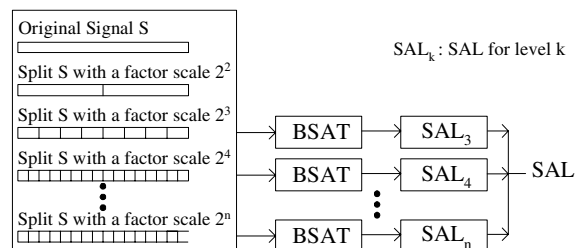


Figure 2: Frequencies segmentation with levels decomposition.

3.2. Information Maximization

We are interested in the extraction of the maximum of the information in the speech signal regarding the characteristic of each speaker. As the cumulated rate of the Principal Component (PC) cannot be equal to the rate of fixed variance, we define two rates of variance:

- Effective Rate of Variance (ERV) is the percentage of total information we want to keep.
- The Rate of Real Variance (RRV) is the percentage of the real information.

Let i be a PC with cumulative value ERV_i . There exists one and only one node so that $ERV_{i+1} \geq ERV$. We denote the difference by

$$T_+ = ERV_{i+1} - ERV$$

The RRV is obtained while increasing the ERV by T_+ . We say that the information is maximized and the $RRV = ERV_{i+1}$.

3.3. Cost Introduced by RRV

The use of TVR introduces a cost in the BSAT algorithm. The node i such as $T_+ = TVE_{i+1} - TVE$ depends on speaker characteristics and ERV. Greater value of T_+ corresponds to greater PC of the node $(i+1)$ and consequently the last node chosen by application of PCA is more representative for a

given speaker. On the other hand, it is possible that this same node corresponds to a smaller PC for another speaker. In this case it is more interesting to maximize the information for the first speaker and minimize it for the second. The ideal case to minimize this cost is the use of an appropriate criterion to every application of CPA. The cost introduced is minimal if: a) T_+ and CP_i are important and b) T_+ is small.

Determination of ERV is very important. An iterative algorithm can be used to determine the optimum value corresponding to the best performance achieved by the SV system.

3.4. Application of PCA in BSAT

Given a decomposition in the WP tree with n levels, we build n sets $\{A_e\}_{e=1,2,\dots,n}$ such as,

- $A_e = \left[\{A_s\}_{s=1,2,\dots,N_s} \right]$, N_s is the number of sessions available for a given speaker.
- $A_s = \left[\{A_p\}_{p=1,2,\dots,N_p} \right]$, N_p is the number of sentences for a given session.
- $A_p = \left[\{A_f\}_{f=1,2,\dots,N_{f,p}} \right]$, $N_{f,p}$ is the number of windows in sentence p .
- $A_f = \left[\{E_e^k\}_{e=1,2,\dots,k_e} \right]$, where $E_n^k = \sum_i |C_n^k(i)|^2$ is the energy of node C_n^k .

The obtained abstract tree after step 5 of the BSAT algorithm is an abstract tree with characteristics related to the speaker.

Note that the number of windows for all the available sentences is $N_f = \sum_p N_{f,p}$.

4. Experimental Setup and Task

4.1. Corpus Description

Two different subsets of corpora have been used: Yoho and Spidre.

4.1.1. Yoho Corpus

A subset of sixty (60) speakers (47 males and 13 females) extracted from Yoho corpus has been used for all the experiments. It consists of 96 sentences uttered by each speaker for the training process and 40 different sentences for the verification task. Each speech signal contains approximately 6 seconds of speech.

4.1.2. Spidre Corpus

A subset of forty-five (45) speakers (27 males and 18 females) extracted from Spidre corpus has been used. Four conversations are available for each speaker. Three (3) conversations (two from match conditions and one from mismatch conditions) are used for training. One (1) conversation (from mismatch conditions) for the test. Each speech signal contains approximately 55 seconds of speech.

The verification process has been done on different lengths of the speech signal.

4.2. MFCC Analysis

The features are a 24-dimensional vector consisting of 12 cepstral coefficients and 12 Δ coefficients. Analysis conditions are listed in Table 1.

4.3. Wavelet-Based Analysis

WP decomposition is applied to each temporal analysis window of 25 ms of duration. a) The log energy in each of the frequency bands is computed giving a total of 20 coefficients in the case of MFDWC. b) The log energy is applied to BBS and BSAT as described in 2.5 and 3.1 respectively. In all cases, DCT is then computed and the first 12 DCT coefficients are selected as static features. Finally, the 12 corresponding Δ coefficients are computed.

Table 1: Fourier analysis

| Parameter | Value | |
|--------------------------------|-------------------|-------------------|
| | Fourier | Wavelet |
| Pre-emphasis | $1 - 0.97 z^{-1}$ | $1 - 0.97 z^{-1}$ |
| Window length | 25.0 ms | 25.0 ms |
| Window shift | 10.0 ms | 10.0 ms |
| Number of features | 24 | 24 |
| Cepstral coefficient liftering | 22 | - |
| Cepstral mean normalization | yes | - |
| Hamming window | yes | yes |
| Order of Daubechies | - | 8 |

4.4. Speaker Model Estimation

We have been interested in two different applications depending on the corpus used: a) Yoho corpus is used for TDSV. b) Spidre corpus is used for TISV. In the first case, three-state left to right with no skip phone-based HMM models were constructed for each speaker. In the second case, GMM-based speakers have been constructed.

In both cases, each model contains 16 mixtures (8 as static and 8 as dynamic). Each of the mixture components has a diagonal covariance matrix. The background model is estimated using all the data available for training.

4.5. Experimental Results

Experiments using Yoho and Spidre are presented in this section. We have set the following parameters:

- Order of Daubechies wavelets is set to 8.
- Initial tree of 5 decomposition levels is used.
- Performance of the system is defined by:

$$Perf = 100 - (FA + FR)$$

FA and FR are the percentages of false acceptations and false rejections.

We have used the method of selecting the Best Structure Abstract Tree (BSAT), which computes the optimal sub tree from an initial tree. The results are as follows.

4.5.1. Experiments with Yoho

Table 2 shows TDSV system performance (SP) using BSAT for different values of ERV and MFCC, MFDWC and BBS.

Table 2: Comparison of TDSV SP between BSAT for different values of ERV and MFCC, MFDWC and BBS using Yoho corpus.

| Type | | FR (%) | FA (%) | Perf (%) |
|-------|---------|--------|--------|----------|
| BSAT | ERV=96% | 0.08 | 12.54 | 87.38 |
| | ERV=80% | 0.00 | 3.63 | 97.21 |
| | ERV=70% | 0.00 | 0.00 | 100.0 |
| MFCC | | 0.69 | 2.49 | 96.82 |
| MFDWC | | 0.88 | 13.66 | 85.46 |
| BBS | | 0.16 | 55.50 | 44.42 |

The results show the efficiency of MFCC parameters compared to MFDWC and BBS for clean data. An appropriate value of ERV provides a 100% performance with BSAT.

4.5.2. Experiments with Spidre

Figure 3 shows a comparison of TISV SP between BSAT using ERV=70% and MFCC, MFDWC and BBS. The verification process has been performed for a) 2 seconds of speech, b) 4 seconds and c) using the entire speech test signal. The results show that MFCC, MFDWC and BBS parameters-based SV performance worsen drastically with reduction of test duration of these noisy data. BSAT parameters have shown that two to four seconds are sufficient to extract speaker characteristics and provide a very high performance.

5. Discussion

From a numerical point of view, we found that there is a significant difference in the output probabilities. For brevity, only Yoho corpus is discussed, similar results have been found for Spidre. Let P_{BS1} and P_{BS2} be the first and second best output probabilities for a given speech signal. We define $\mu P_{BS1} = (1/N) \sum P_{BS1}$ the average best output probability over the N test speech signals. We also define $\Delta P_{BS} = (1/N) \sum |P_{BS1} - P_{BS2}|$, the average difference between the best two scores. Table 3 shows that MFCC, MFDWC and BBS range probabilities are close as well as the ΔP_{BS} , which suggest that the system is vulnerable to errors. Depending on the ERV, BSAT performs quite differently. Simulations have shown that the best μP_{BS1} and ΔP_{BS} can be obtained for ERV=70% and therefore provides more robustness to the SV system.

6. Conclusions

In this paper we have introduced successfully a PCA-based wavelet transform to perform frequencies segmentation with levels decomposition. A speaker dependent library tree has been built. The constructed tree is abstract and specific to each speaker. Therefore the extracted parameters are more discriminative and appropriate for speaker verification applications. This technique has shown robustness and 100% efficiency in both cases Yoho and Spidre. We are currently running more experiments with larger databases.

7. Acknowledgements

The authors would like to thank professor C. Gargour, École de Technologie Supérieure, Canada, for his interesting suggestions related to this work. We would like also to

acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

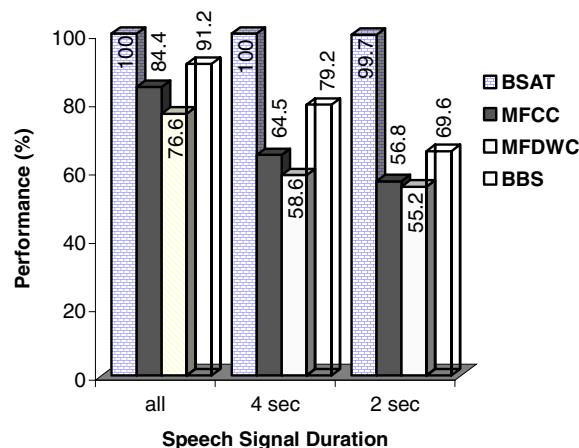


Figure 3: Comparison of TISV SP between BSAT (ERV=70%), MFCC, MFDWC and BBS using Spidre corpus for different lengths of speech signal.

Table 3: Average output probabilities between the first and the second best scores for BSAT and MFCC, MFDWC and BBS using Yoho corpus.

| Yoho Corpus | | μP_{BS1} (log) | ΔP_{BS} (log) |
|-------------|---------|---------------------|-----------------------|
| BSAT | ERV=96% | -16.14 | 0.86 |
| | ERV=80% | -12.93 | 2.14 |
| | ERV=70% | -11.29 | 3.14 |
| MFCC | | -55.63 | 0.28 |
| MFDWC | | -28.37 | 0.53 |
| BBS | | -35.95 | 0.23 |

8. References

- [1] Gish, H. and Schmidt, M., "Text-Independent Speaker recognition", *IEEE Sig. Proc. Magazine*, pp. 18-32, 1994.
- [2] Gowdy, J. N. Tufekci, Z., "Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition", *ICASSP*, 2000.
- [3] Bourlard, H. and Dupont, S., "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands", *ICSLP*, 1996.
- [4] Farooq, O. and Datta, S., "Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition", *IEEE Signal Processing Letters*, 8(7), 196-198, 2001.
- [5] Jolliffe, I.T., "Principal Component Analysis", Springer Verlag, 1986.
- [6] Goswami, J. & all., "Fundamental of Wavelets", Wiley, 1999.
- [7] Mallat, S. G., "A Wavelet Tour of Signal Processing", *New York: Academic Press*, 1998.
- [8] Daubechies, I., "Ten Lectures on Wavelets", *Philadelphia, Pa.: Soc. for Ind. and App. Math.*, 1992.
- [9] Badri, N., Benlahouar, A., Tadj, C., Gargour, C. Ramachandran, V., "On the Use of Wavelets and Fourier for Speaker Verification", *45th IEEE MWSCAS*, 2002.
- [10] Coifman, R. and all., "Entropy-Based Algorithms for Best Basis Selection", *IEEE Trans. on Inf. Th.*, 38, 1992.