Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 6 (2011) 124–129

Procedia
Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent, Filterbank-Based Speaker Recognition

Claude Turner[a], Anthony Joseph[b], Murat Aksu[c], Heather Langdon[d]a*

[ac]*Dept. of Computer Science, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA*

[b]*Dept. of Computer Science,Pace University, 163 William Street, New York, NY 10038*

Abstract

An important step in speaker recognition is extracting features from raw speech that captures the unique characteristics of each speaker. The most widely used method of obtaining these features is the filterbank-based Mel Frequency Cepstral Coefficients (MFCC) approach. Typically, an important step in the process is the employment of the discrete Fourier transform (DFT) to compute the spectrum of the speech waveform. However, over the past few years, the discrete wavelet transform (DWT) has gained remarkable attention, and has been favored over the DFT in a wide variety of applications. This work compares the performance of the DFT with the DWT in the computation of MFCC in the feature extraction process for speaker recognition. It is shown that the DWT results in significantly lower order for the Gaussian Mixture Model (GMM) used to model speech and marginal improvement in accuracy.

"Keywords: Cepstral Coefficients, Speaker Recognition, Wavelets"

1. Introduction

Speaker recognition is the identification of persons from their speech or voice samples (Furui, 1997; Campbell, 1997; Bimbot et al., 2004). Speaker recognition can be divided into three specific tasks: (1) identification, (2)

doi:10.1016/j.procs.2011.08.024

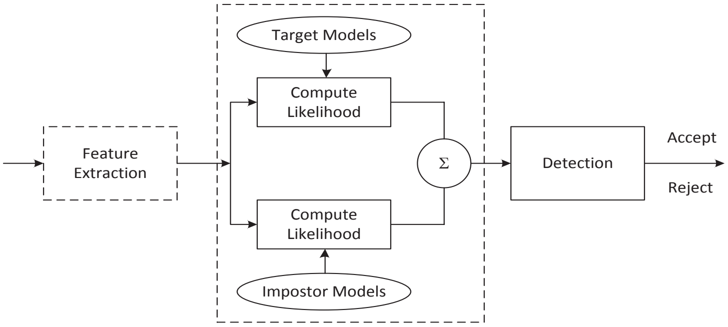

Figure 1 Speaker Detection System Framework

verification, and (3) segmentation and clustering (Solomonoff et al., 1998). Speaker *identification* determines who among a group of speakers provided the input speech to the system (Furui, 1997). It has two modes of operation: closed-set and open-set. In closed-set mode, the input speech is assumed to belong to one of the speakers in the group. The open-set case allows for the situation where the input speech could belong to speakers outside of this group. Speaker *verification* determines if a person is whom he/she claims to be from the person's voice samples. Synonyms for speaker verification include, voice verification or authentication, speaker authentication, talker verification or authentication, and speaker detection (Jin, 2007). In a multiple speaker scenario where the speech is a heterogeneous mixture of speech from various speakers, speaker *segmentation and clustering* partitions the speech into speaker-homogeneous regions. Specific speaker recognition applications include forensics, diarization (attempts to extract speaking turns of different participants from a spoken document), and telephone based systems that recognize the speaker. Speaker recognition systems may be further categorized as text-dependent or text-independent. In *text dependent* recognition, the phrase to be spoken is known to the system. In the *text independent* case, the system does not have any a priori knowledge of the phrase. Text-independent speaker recognition is the more challenging of the two categories.

At its most fundamental level, the speaker recognition problem may be cast as a pattern recognition problem (Jin, 2007). As such, it can be partitioned into two modules: (a) a training module, and (b) a classification module. The classification module can be further divided into two components: pattern matching and decision. The feature extraction module estimates a set of features from the speech signal that represent some speaker-specific information. The speaker-specific information results from complex transformations occurring at multiple levels of the speech production process: semantic, phonologic, phonetic, and acoustic (Atal, 1976;Campbell, 1997). Despite the variety among the categories of speaker-specific information, there are only a small set of criteria that they must satisfy. These are discussed by Nolan and Wolf (Nolan, 2009;Wolf, 1972). The *pattern matching* module is responsible for comparing the estimated features to the speaker models. There are many types of models that could be used in speaker recognition, including Gaussian Mixture Models (GMMs), Hidden Markov Models, and vector quantization (VQ). The decision module analyzes the similarity score(s), which could be either statistical or deterministic, to make a decision. The *decision* process is dependent on the system task. For the closed set identification task, the decision could be to select the identity associated with the model that is most similar to the test sample. In open-set applications, the systems can also require a threshold to verify whether the similarity is valid. Figure 1 depicts the speaker recognition system, and the dashed box represents the pattern matching module.

A typical speaker recognition system usually consists of two phases: an enrollment (or training) phase, and an authentication (or testing) phase. In the enrollment phase, the user speaks an appropriate phrase into a microphone or similar device attached to the system. The system then extracts speaker-specific information from the speech signal to be used to build a model for the speaker. GMM (Reynolds, 1995) is one of the most popular methods for the modeling process. The purpose of the testing phase is to determine whether the speech samples belong to one of the registered speakers. As in the training phase, speech features are extracted from the speech signal presented. The speaker is then determined by finding the speaker model which yields the maximum posterior probability for the input feature vector sequence (Reynolds, 1995).

All audio processing techniques begin with feature extraction—the conversion of the raw speech signal to acoustic vectors that characterize speaker-specific information. Although a variety of filterbank-based methods exist for feature extraction, such as Linear Prediction Cepstral Coefficients (LPCC) (Makhoul, 1975) and Perceptual

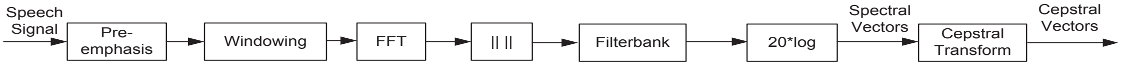

Figure 2 Modular Representation of Feature Extraction

Linear Prediction Cepstral Coefficients (PLPCC) (Hermansky, 1990); the Mel Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980) approach has been the most widely employed for feature extraction (Ganchev et al., 2005). In recent years, numerous variations and improvements of the original MFCC idea have been proposed (Ganchev et al., 2005; Sigurdsson et al., 2005). This is mainly attributable to researchers' efforts to exploit progress made in the area of psychoacoustics (Ganchev et al., 2005).

Wavelets have been employed in speaker recognition applications for over a decade with some success (Phan, 2000; Al-Ani, 2007). The goal of this work is to compare the performance of the discrete wavelet transform with the discrete Fourier transform in the speaker recognition process. Specifically, we will compare performance in terms of accuracy and efficiency between the DFT and the DWT for Daubechies's first ten wavelets at six different decomposition levels.

2. Methods

2.1 Mel Frequency Cepstral Coefficients (MFCC)

A modular representation of a filterbank-based feature extraction model that generates the MFCC is depicted in Fig. 2. The speech signal is first pre-emphasized by applying the filter, $x(t) = y(t) - a \cdot y(t-1)$, where $a \in [0.95, 0.98]$. The goal of the filter is to enhance the high frequencies of the spectrum, which is diminished during the speech production process. Following the pre-emphasis stage is a windowing process, where a window whose size in duration is much smaller than the whole speech signal, is applied starting at the beginning of the signal, and then shifted to the right and applied, successively, until the end of the signal is reached. Two quantities must be set: the width of the window and the shift between consecutive windows. For the width of the window, two values are often used: 20ms and 30ms. These values correspond to the average duration necessary for the stationary assumption to hold. In the case of the delay, a value is chosen so that there is some overlap between consecutive windows. Ten milliseconds is often used. Once the width of the window and the shift between consecutive windows are found, the type of window can then be chosen. The Hamming and Hanning windows are most often used in speaker recognition. Next, for each of the windowed signals emerging from the windowing process, an N-point DFT is computed. Typically, N is chosen as a power of 2 and is classically 512 points, which is greater than the number of points in the window. Next, the modulus of the DFT for each of the spectral vectors is obtained, and from this the corresponding power spectrum for each is taken over 512 points. Since the signal is real valued, the spectrum is symmetric, thus only the first half plus one of it is kept--257 points.

The spectrum consists of much fluctuation. However, in this context, such details are not of interest. It is only the envelope of the spectrum that is of interest. Smoothing removes some of these details. To realize the smoothing and to get the envelope of the spectrum, the spectrum is multiplied by a filterbank. The filterbank is defined by the shape of the filters and by their frequency localization—left frequency, central frequency, and right frequency (Ganchev et al., 2005). Filters may be of a triangular or other shape, and can be differently located on the frequency scale. The Bark/Mel scale is sometimes used for frequency localization of the filters. It is an auditory scale that is similar to the frequency scale of the human ear. A commonly implemented equation for localization of the central frequencies, which is the one used in the experiments of this study, is given by:

$$\tilde{f} = 2595 \cdot \ln\left(1 + \frac{f_{lin}}{700}\right). \qquad (1)$$

The original filterbank of Davis and Mermelstein (Davis & Mermelstein, 1980), FB-20, is the one used here. It proceeds as follows: Given the N-point DFT of the discrete input signal, x_n, $\hat{x}_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi nk}{N}}, k \in \{0, 1, \cdots, N-1\}$, a filter bank with M equal height triangular filters is constructed. Each of the M equal height filters is defined by:

$$\hat{h}_i(k) = \begin{cases} 0\,, & k < f_{b_{i-1}} \\ \frac{k - f_{b_{i-1}}}{f_{b_i} - f_{b_{i-1}}}, & f_{b_{i-1}} \leq k < f_{b_i} \\ \frac{f_{b_{i+1}} - k}{f_{b_i} - f_{b_i}}, & f_{b_{i-1}} \leq k < f_{b_i} \\ 0\,, & k \geq f_{b_{i-1}} \end{cases} \tag{2}$$

$i \in \{1, 2, \cdots, M\}$, where i is the filter index, f_{b_i} is the boundary point for the filter, $\hat{h}_i$, and $k \in \{1, 2, \cdots, N\}$ corresponds to the *k th* coefficient of the N-point Discrete Fourier Transform (DFT). Each boundary points f_{b_i} depends on the sampling frequency, F_s, and the number of points, *N*, in the DFT, and is given by:

$$f_{b_i} = \left(\frac{N}{F_s}\right) \cdot \tilde{f}^{-1}\left(\tilde{f}(f_{low}) + i \cdot \left(\frac{\tilde{f}(f_{high}) - \tilde{f}(f_{low})}{M+1}\right)\right). \tag{3}$$

The values f_{low} and f_{high} are, respectively, the low and high boundary frequencies for the entire filter bank, *M* is the number of filters, and $\tilde{f}^{-1}$ is the inverse of (1) given by: $\tilde{f}^{-1} = f_{lin} = 700\left[e^{-\left(\frac{\tilde{f}}{2595}\right)} - 1\right] \cdot$ $f_{low}, f_{high},$ and f_{lin} are in Hertz (Hz), while $\tilde{f}$ is in mels. The filter bank of Davis and Mermelstein is comprised of 20 equal height filters, which cover the frequency range [0,4600] Hz. The center frequencies for the first ten filters are linearly spaced between 100 Hz and 1000 Hz, and the next ten have center frequencies logarithmically spaced between 1000 Hz and 4000 Hz. The next step computes the logarithm of windowed signal followed by the discrete cosine transform. The process may be summarized compactly as follows:

$$c_t = \sum_{k=1}^{M} X_k cos\left(t \cdot (k - 1/2)\frac{\pi}{M}\right), \quad t \in \{1,2,\cdots,J\}, \tag{4}$$

where

$$X_i = \log_{10}\left(\sum_{k=0}^{N-1} |\hat{x}_k|\, \hat{h}_i(k)\right). \tag{5}$$

2.2 Wavelets

Given low pass and high pass filters, g and h, and associated scaling and wavelet functions, φ and ψ, respectively, approximation and detail coefficients of *x* are obtained via the DWT and are given by $a_x(j+1,k) = ((a_x(j,\cdot) * g) \downarrow 2)(k) = \sum_{m \in Z} g_{2k-m}\, a_x(j,m)$, and $d_x(j+1,k) = ((a_x(j,\cdot) * h) \downarrow 2)(k) = \sum_{m \in Z} h_{2k-m}\, a_x(j,m)$, respectively, where $j \in \{1, 2, \cdots, L\}$ and $k \in \{0, 1, \cdots, n_j - 1\}$, * and ↓ represent the convolution and downsampling operations, respectively, and n_j represents the number of approximation (or detail) coefficients at level *j*. Also, $a_x(j,\cdot) = \{a_x(j,0), a_x(j,1), \cdots, a_x(j, n_j - 1)\}$. We assume that $j = 0$ yields the pre-emphasized signal, x, itself. That is, $a_x(0,\cdot) = x$. We replace the DFT in Fig. 2 by the DWT, and take as its output the following detail signal obtained from the detail coefficients, $d_x(j,k)$:

$$D_j(t) = \sum_{k \in Z} d_x(j,k)\, \psi_{j,k}(t). \tag{6}$$

To obtain feature extraction filterbank coefficients using DWT, we substitute Eq. 6 for $\hat{x}_k$ in Eq. 5, to get: $z_k = \log_{10}\left(\sum_{i=0}^{N-1} \left|D_j(i)\right| \hat{h}_k(i)\right)$, for some $j \in \{1, 2, \cdots, L\}$. The coefficients that results when the DWT is substituted for the DFT in Eq. 5 is then obtained by substituting z_k for X_k in Eq. (4), to get:

$$w_t = \sum_{k=1}^{M} z_k cos\left(t \cdot (k - 1/2)\frac{\pi}{M}\right), \quad t \in \{1,2,\cdots,J\}. \tag{7}$$

3. Experimental Setup & Results

We used six Region 1 speakers from the TIMIT database—three males and three females—and the following single utterance from each: “She had your dark suit in greasy wash water all year.” Each speaker has a copy of this utterance stored in a file name sa1.wav. The following six speakers were used—three males and three females—from the TIMIT database: FECD0, FJSP0, FKFB0, MKLS0, MPGH0, and MPGR0. The first letter of the speaker designation tells us the gender. The next three letters following it are the first, middle and last initial of the speaker’s name. The last position makes it possible to distinguish multiple speakers with the same gender and initials--a zero indicates the first such speaker, a 1 for the second, etc. Fig. 4 gives the speech signal for FJSP0.

In the training phase of our experiment, Eq. 7 was implemented for each of the speaker signals, and a GMM was used to model the features obtained. In the testing phase, for each speaker, Eq. 7 was again used to extract the features, and then a maximum likelihood function was used to determine the model that best matched the input speech. The process was repeated for the first ten Daubechies's wavelets—db1, db2, ..., db10—and for six decomposition levels of the DWT. These results were compared with the FFT approach given by Eq. 4. The value of a used for the preempahsis was $a = 0.95$. The window size, and overlap used in the windowing module was, 320 and 160 samples, respectively. The filterbank was the original filterbank design of Davis and Mermelstein (Davis & Mermelstein, 1980), with 20 filters, $M = 20$, as discussed in Section 2.1. The order of the GMM (number multivariate Gaussian distributions used) was optimized using the Akaike Information Criteria (Akaike, 1974).

The results for the DWT are provided in Table I. Each value in the table is a measure of the number of times a speaker is misidentified. The best results are provided on Level 2 of the wavelet decomposition, with db1, db4, db6, and db10 identifying each speaker without error. This is an improvement over the DFT, which misidentified one out of the six speakers (error=1/6). Table II compares GMM order for the DWT versus the DFT. The row marked "DFT" gives the number of models used in the training phase for each speaker. There are four speakers, FECD0, FJSP0, MKLS0 and MPGH0, that have order 20. The other two speakers FKFB0 and MPGR0 have order 18 and 16, respectively. The results for the DWT, given on the row labeled "DWT," show that the DFT require an order that is three to five times that of the DWT. A smaller optimal order is preferred because it leads to a GMM that is less computationally intensive to generate and use. Therefore, the DWT approach seems to provide marginal improvement over DFT in terms of its accuracy for speaker recognition with the MFCC. Further, the DWT provides significant improvement in terms of the optimal order required to generate the GMM.

TABLE I Speaker Identification Error out of 6

LEVEL	**WAVELET**									
	db1	**db2**	**db3**	**db4**	**db5**	**db6**	**db7**	**db8**	**db9**	**db10**
1	2	2	2	2	2	2	3	3	3	3
2	0	1	1	0	1	0	1	1	1	0
3	5	3	4	3	2	3	3	3	3	3
4	1	1	2	1	2	2	0	2	4	4
5	2	2	3	5	3	4	4	4	5	3
6	1	5	3	2	4	2	5	4	3	4

TABLE II Order of Gaussian Mixture Model – DFT vs. DWT

	Speakers					
	FECD0	FJSP0	FKFB0	MKLS0	MPGH0	MPGR0
DFT	20	20	18	20	20	16
DWT	5	5	5	5	4	5

4. Conclusion

This work compares the performance of the DFT with the DWT in the computation of the MFCC for feature extraction in speaker recognition. It showed that the speech features derived through the DWT resulted in a more efficient representation, in terms of order, for the GMM that used in the statistical modeling of features. It also showed marginal improvement in accuracy of the DWT over the DFT. Specifically, it was shown that the GMM order required when the DFT is used in the MFCC feature extraction process was approximately three to five times that required for the DWT. Finally, in terms of accuracy, the wavelet approach matches the DFT at decomposition Level 2. However, the DWT outperforms the DFT with an error rate of zero when the following wavelets are used: db1, db4, db6, and db10.

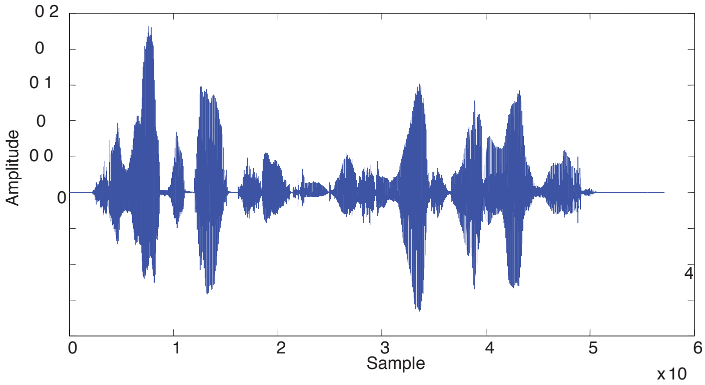

Figure 3: Speech samples sa1.wav for FJSP0

References

Al-Ani, M. S., Mohammed, T. S. and Aljebory, K. M., (2007). Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform. *Journal of Computer Science*, *3*(5), 304-309.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*(4), 460-475.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravieer, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, *2004*(4), 430-451. Hindawi Publishing Corporation.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85*(9), 1437-1462.

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

Furui, S. (1997). *Speaker Recognition in Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

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)* (p. pages). Patras, Greece. Retrieved from 10.1.1.83.7049.pdf.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, *87*, 1738-1752.

Jin, Q. (2007). *Robust Speaker Recognition*. Carnegie Mellon. Retrieved from http://www.lti.cs.cmu.edu/Research/Thesis/QinJin.pdf.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63, 561-580.

Nolan, F. (2009). *The Phonetic Bases of Speaker Recognition (Cambridge Studies in Speech Science and Communication)*, 5-25. Cambridge University Press.

Phan, F., Micheli-Tzanakou, E. and Sideman, S., (2000). Speaker identification using neural networks and wavelets. *IEEE Engineering in Medicine and Biology Magazine 19*(1), 92-101.

Sigurdsson, S., Peterson, K. B., & Lehn-Schiøler, T. (2005). Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. *Proc. Int. Conf. Music Inf. Retrieval,* 286–289.

Solomonoff, A., Mielke, A., Schmidt, M., & Gish, H. (1998). Clustering speakers by their voices. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)* (*2)*, 757-760.

Wolf, J. J. (1972). Efficient Acoustic Parameters for Speaker Recognition. *Journal of the American Statistical Association*, *51*, 2044-2056.