Feature Extraction with Discrete Wavelet Transform and Mel Frequency Filters for Spoken Digit Recognition

Andrés Fleiz, Mauricio Martínez

Universidad la Salle, 06140 México City, México
{andresfleiz, mauricio.martinez}@lasallistas.org.mx

Abstract. In this paper we propose a new method for analysis of speech signals based on the discrete wavelet transform and a mel frequency filter bank, in order to extract representative features of a speech signal as one of the most important steps in a speech recognition system. The major issues concerning the design of the proposed system are selecting the optimal wavelets and decomposition level. A comparison of the proposed approach and classic methods as Mel Frequency Cepstral Coefficients (MFCC) and novel proposals that use wavelets shows that our system competes in term of recognition rate.

Keywords: Automatic speech recognition, feature extraction, wavelet transform, speech parameterization, discrete wavelet transform, mel frequency cepstral coefficients, pre-processing, classification.

1 Introduction

The feature extraction process consists in the transformation of a voice signal into a parameter representation. Probably one of the most important representations is the spectral envelop [1] achieved through analysis methods as Mel Frequency Cepstral Coefficients (MFCC) [2]. Ideally, the analysis method should preserve all the meaningful perceptual information in order to identify the phonetic differences; it also should be insensitive to irrelevant variations. Therefore, the meaningful information is given by three variables: amplitude, frequency and time.

Conventional techniques as MFCC make use of time-frequency analysis with fixed window size i.e. uniform resolution that decrease the performance of recognition systems due to unacceptable resolution in time or frequency depending the size of the window. A possible solution is finding an optimal resolution for time and frequency through a multiresolution technique for the analysis, such as wavelet transform.

In the existing work we can find multiple ways of using wavelet for speech recognition tasks, in [3] the speech signal is decomposed into various frequency channels, based on the time-frequency multiresolution property of wavelet transform; in [4] new methods for feature extraction based on wavelet decomposition and reduced order linear predictive coding are proposed for speech recognition; in [5] a sub-band feature extraction technique based on a wavelet transform is proposed for phoneme recognition; in [6] the speech signal is preprocessed by a wavelet transform

Advances in Computer Science and Engineering
Research in Computing Science 67, 2013 pp. 123–132
Paper Recived 16-08-2013 and Acepted 21-10-2013

to increase the accuracy of the recognition system; in [7] an hybrid system based in the discrete wavelet transform and linear predictive coding is proposed for recognition of isolated words; in [8] the performance of the discrete Fourier transform is compared against the discrete wavelet transform in the computation of MFCC in the feature extraction process for speaker recognition; in [9] and [13] the author proposes a new feature vector consisting of coefficients obtained by applying the discrete wavelet transform to the mel-scaled log filter bank energies of a speech frame; in [10] the wavelet transform is used as a part of the front-end processor for feature extraction process in phoneme recognition; in [11] a comparative analysis is performed between the traditional MFCC and a wavelet based feature extraction method for speech recognition; in [12] a speech recognition system is proposed using discrete wavelet transform and artificial neural networks for isolated spoken words; in [14] a system consisting of wavelet based features, linear discriminant analysis and principal component analysis is proposed for recognition of spoken digits; in [22] a method to address the issue of noise robustness using wavelet domain in the front end of an automatic speech recognition (ASR) system; in [23] a paradigm which combines the Wavelet Packet Transform with the MFCC for extraction of speech features is prosed. These approaches can be classified in phoneme recognition, word recognition and speaker recognition. After review each proposed method we found that the design of these systems obeys to an empirical process because several factors affect their performance. The main motivation of this work is to study the behavior of wavelets as multiresolution analysis technique inside of novel recognition system that allows us to identify the main factors that define its performance, this, after delimitating the search space to those values and parameters recommended or used in previous published papers. In order to attain an objective comparison between our technique and others previously proposed, we must modify the recognition systems designed by other authors standardizing the classification and pre-processing steps.

Section 2 of this paper presents the conceptual basis of the method followed for the spoken digit recognition system implementation, for instance the discrete wavelet transform and the mel frecuency cepstral coefficient approaches. The section 3 presents the specifications and details of the experimentation step, and also the main results are presented for discussion in the section 4, where a set of conclusions are obtained in favor of the performance achieved by the proposed system.

2 Method

There have been proved several new approaches to solve speech processing tasks, including new research on the combination of MFCC and wavelet transform, [8] showed that for speaker recognition the DWT outperforms the recognition rate achieved compared to DFT. In [6] the wavelet is used to preprocess the signal and extract features that increase the accuracy for recognition of isolated words.

2.1 Discrete wavelet transform

The wavelet transform is a non-parametric type of analysis which allows multiresolution, understood as localization in time and frequency. This kind of techniques has been used successfully for the analysis of aperiodic and non-stationary signals as speech [15].

The wavelet coefficients are obtained by computing the inner product between an input signal and a function of limited time that has an average value of zero and unit norm [16] so that

$$\int_{-\infty}^{+\infty} \psi(t)\mathrm{d}t = 0\,, \tag{1}$$

$$\|\psi(t)\| = 1\,. \tag{2}$$

Wavelet transform then decomposes the input signal into a set of basic functions called wavelets, which are scaled and translated versions of the mother wavelet ψ. Since the signals treated are discrete we used the Discrete Wavelet Transform (DWT).

DWT is a wavelet transform for which the wavelet is discretely sampled and gives us a compact representation of the signal in time and frequency. It can be computed by an efficient algorithm [16]. The DWT of a signal $s[n]$ is calculated as

$$W[j,k] = \sum_j \sum_k s[k] a^{-j/2}\, \psi(a^{-j}n - k)\,. \tag{3}$$

The DWT can be represented as a filtering process using a low pass filter (scaling) that gives the approximate representation of the signal and a high pass (wavelet) filter that gives the details or high frequency variations. This process is illustrated in the Figure 1.

The filtering process begins when the signal s passes through the first pair of filters producing two coefficients sets: approximation coefficients CA1 and detail coefficients CD1, after the convolution of the filters and a down sampling procedure. The next step or decomposition level splits the CA1 in two other sets CA2 and CD2, this cycle continues until the j-th desired level. We label each wavelet based on the family and decomposition level. We have chosen the wavelet families according to the existing attempts of using the wavelet transform for feature extraction of speech signals [6, 8, 14].

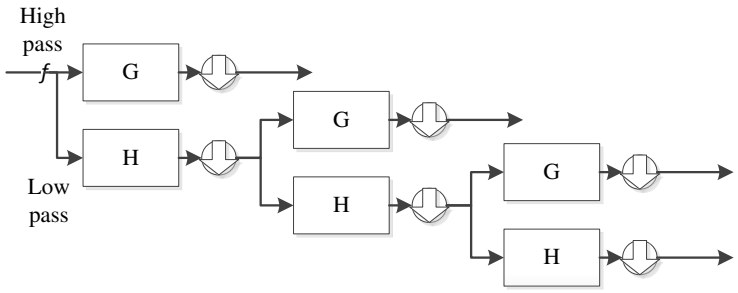

Fig. 1. Filtering representation of a 3 level Discrete Wavelet Transform (DWT) signal decomposition.

The DWT has the main advantage of applying varying widow size, broad for low frequencies and narrow for high frequencies this property allows good resolution in all the frequencies ranges. The coefficients that result of the transformation of the signal are obtained after the concatenation of the last level of the decomposition tree, beginning by the approximation coefficients CAj.

2.2 Mel frequency cepstral coefficients

MFCC's are one of the most popular and studied features used in speech processing applications [17]. MFCC technique obtains the cepstrum of a signal in a set of coefficients that model the human auditory perception through a scale called the mel scale which is linear before 1 kHz and logarithmic above this value. The frequency warping can be implemented through a filter bank of triangular bandpass filters which are centered in the mel scale.

Given the spectrum of a signal

$$S[w] = \sum_{k=0}^{N-1} s[k] e^{\frac{-j2\pi kw}{N}} . \quad (4)$$

The next step is to pass the data of the power spectrum through the mel filter bank, the output is the mel spectrum of the signal. In order to get the cesptrum the logarithmic of the mel spectrum is calculated. The final step involves converting the log mel spectrum back to "time". This could be done by taking the Discrete Cosine Tranform (DCT) of the mel-scaled log spectrum because it is real. The cepstral representation of the signal spectrum provides a good representation of the local spectral properties of the signal in the frame of analysis. This last procedure is represented by

$$c_n = \sum_{k=0}^{M} \log(S_k) \cos\left[\frac{n\left(k-\frac{1}{2}\right)\pi}{M}\right] \quad (5)$$

where $n = 1, 2, \ldots, k$ and S_k is the mel spectrum of the signal under analysis.

2.3 Feature extraction WMFC

In this paper, a multiresolution analysis is performed using the DWT replacing the traditional DFT used to obtain the cepstrum of a speech signal. To obtain a set of features first the signal is broken into 16 ms frames with 50% of overlap, and then is windowed with a hamming window.

$$w[n,\alpha] = (1-\alpha) - \alpha \cdot \cos\left(\frac{2\pi n}{N-1}\right) \tag{6}$$

where $\alpha = 0.46, 0 \leq n \leq N-1$ and N is the length of the window.

$$s_w[n] = s[n]w[n]\,. \tag{7}$$

The DWT of each windowed frame is computed until reach the level of decomposition chosen. The absolute value of the coefficients obtained at j level of decomposition is filtered by a mel scale filter bank and the logarithm of the output of these filters is computed.

$$Z_p = log\left(\sum_{k=0}^{N-1} H_p[k]D_j[k]\right) \tag{8}$$

where H_p is the p^{th} triangular mel filter and $k = 0,1,2,\ldots,N$.

Finally the DCT is applied to decorrelate the log-filter bank energies and a vector of features is obtained.

$$c_t = \sum_{p=1}^{P} Z_p \cdot \cos(\pi t(p-0.5)/P)\,. \tag{9}$$

The block diagram of the analysis system is represented by Figure 2.

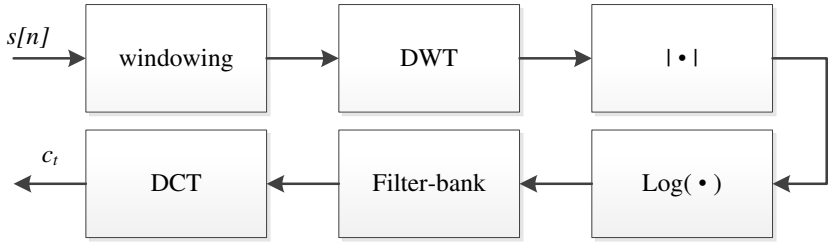

Fig. 2. Block diagram of the proposed analysis system.

3 Experiments and results

3.1 Experimental setup

Our experiments were conducted under ideal conditions, using the TI-46 Speaker Dependent Isolated Word Corpus [21], which consists of ten English digits “zero”

through “nine”, collected from several male and eight female speakers. We selected 360 speech samples from this database; from two male and two female speakers, each one repeating every digit ten times. The sample frequency rate is 11025 kHz, 12-bit PCM and mono channel.

The speech samples are pre-processed to remove the silence using energy threshold criteria, and then the corresponding feature extraction technique is applied. In all the cases the samples are windowed with 16 ms hamming window with 8 ms overlap.

After the preprocessing and windowing the feature vectors are obtained by the corresponding analysis technique. In the DWT case, selection of wavelet families and decomposition level is done accordingly to the conclusions and recommendations stated in [8] (db1, db4, db6, db8 and db10), [6] (db8), [14] (coif5), [18] (sym6)]. In MFCC with DFT and DWT, the number of mel filters in the filter bank is 20 and the number of MFCCs coefficients is 12, both values taken from [19].

The recognition task is achieved with a minimum Euclidian distance criteria based on the computation of the Frobenius norm of the test samples resulting feature vector. In the training phase the mean of 20 feature vectors of random samples is calculated to form the reference patterns for each digit.

The feature extraction techniques used for comparison are: a popular and well proved MFCC method, a new form of improved MFCC (Wavelet MFCC in [6]) and the method proposed in this paper. The comparison is done in term of the percentage accuracy achieved by each method in equal conditions.

In order to evaluate the robustness and performance of the proposed method in noisy environment and the known fact that wavelets performs better than Fourier in recognition tasks with noisy samples, the test patterns were contaminated with additive white Gaussian noise to various signals to noise ratios (SNR).

3.2 Results

To compare performance and robustness of the method here proposed with others (analysis technique in [6] and classic MFCC) experiments are conducted for various wavelets, level decomposition and SNR conditions. Table 1 show the recognition rate at given wavelet and decomposition level with clean speech.

Table 1. Best combination of parameters (wavelet family/order and decomposition level)

Parameter	Wavelet MFCC	WMFC
Wavelet	db1	db10
Decomposition level	4	3
Recognition Rate	83.6%	86.0%

Numerous experiments were conducted in the search of the best combination of wavelet and decomposition level for each of the feature extraction techniques, the result is shown in table 1.

After analyzing the performance of each system in terms of recognition accuracy at a given wavelet and decomposition level, we found that the best recognition rates were obtained with the 2^{nd}, 3^{th} and 4^{rh} level of decomposition with the Daubechies

and Coiflet families, particularly with db1, db10 and coif5. In Figure 3 shows the comparison of the Wavelet MFCC and the WMFC system performance.

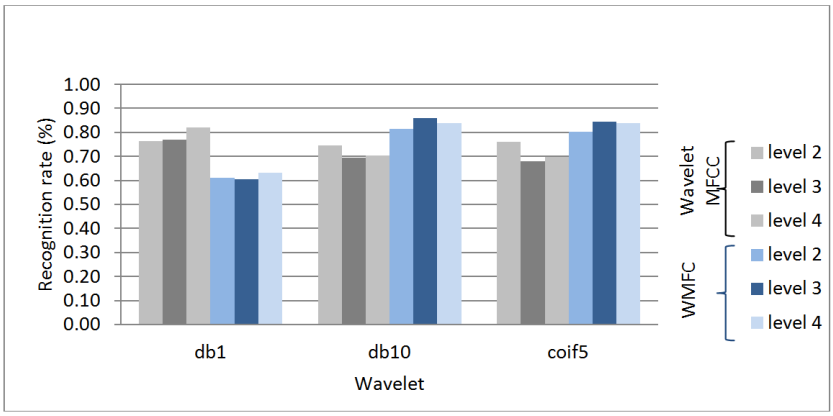

Fig. 3. Recognition rate of Wavelet MFCC system proposed in [6] and WMFC.

The behavior of the proposed system WMFC shows that the recognition accuracy is comparable with the performance of other systems proposed in the bibliography for clean speech.

In order to analyze the robustness of each technique against noisy conditions several experiments were conducted with different signal to noise ratio (SNR). As stated in [6] we found that wavelet techniques have better performance under non ideal conditions, nevertheless techniques that use Fourier as MFCC presented good performance with clean and less contaminated signals.

Table 2. Recognition rate per digit for WMFC in both clean and noisy environment.

Digits	Clean speech	SNR= 20dB	SNR= 10dB	SNR= 0dB
1	62.38	17.00	21.00	09.17
2	86.50	05.63	12.00	09.50
3	92.50	42.38	31.87	09.50
4	80.00	11.13	15.88	06.67
5	88.75	10.38	05.13	09.50
6	92.63	06.63	03.63	16.50
7	87.50	01.63	01.37	13.17
8	88.00	01.62	00.50	16.83
9	89.62	06.13	08.13	08.83
Average	85.32	11.39	11.06	11.07

After comparing Table 2 and Table 3, it is clear that the WMFC performs better with noisy signals in comparison with standard MFCC. Analyzing the recognition rates per digit results in Table 3, which shows an odd behavior as the noise increases, with clean signals having the lower performance for digit “one”, as shown in [20], strangely it improves as the noise increases, while the recognition in all the other digits decay abruptly. This is not the case in WMFC where the recognition rate for all the digits decays as the SNR decreases.

Table 3. Recognition rate per digit for MFCC in both clean and noisy environment.

Digits	Clean speech	SNR= 20dB	SNR= 10dB	SNR= 0dB
1	67.37	13.13	26.75	88.00
2	87.62	05.75	05.00	00.00
3	90.63	00.00	13.00	00.00
4	77.00	00.13	04.50	00.00
5	92.00	02.75	07.62	00.00
6	92.50	20.75	12.50	00.00
7	86.00	00.00	00.25	00.00
8	92.50	05.50	10.25	00.00
9	81.88	09.50	10.88	00.50
Average	85.28	06.39	10.08	09.83

A possible explanation could be that the Euclidian distance in high contaminated environments doesn't perform well for MFCC but still works with WMFC. This result is a clue that allows us to guess that with a more robust and complex classifier such a neural network the WMFC could outperform even more the recognition rate of MFCC.

4 Conclusion

The goal of this work is to develop a feature extraction system for spoken digits by using the multiresolution property of wavelets. The obtained results show that the proposed system WMFC is very competitive with previously published results [6] for different languages, data bases and [8] different tasks. An interesting result is the behavior of a conventional method such MFCC in noisy environments, as it is seen it gives less recognition accuracy as the noise increases, and the worst recognized digit (one) results in the best recognized digit; conversely, WMFC shows better recognition accuracy under noisy environments, and also coherent results.

To achieve a fair comparison of method and systems the conditions and parameters must be standardized, focusing the attention in the analysis method and establishing equal pre-processing and classification techniques. A formal analysis of the implication of using one or another wavelet family or one or another decomposition level is necessary to understand the sensitivity of the method to these factors.

References

1. Rabiner, L.R., Schafer, R.W.: Introduction to Digital Speech Processing. The Essence of Knowledge (2007).
2. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions of Acoustics, Speech and Signal Processing 28(4), 357-366 (1980).
3. Trivedi, N., et al.: Speech Recognition by Wavelet Analysis. International Journal of Computer Applications 15(8), 27-32 (2011).
4. Nehe, N. S., Holambe, R. S.: DWT and LPC based feature extraction methods for isolated word recognition. EURASIP Journal on Audio, Speech, and Music Processing, 1-7 (2012).

5. Datta, S., Farooq, O.: Wavelet Based Robust Sub-band Features for Phoneme Recognition. IEE Proceedings: Vision, Image and Signal Processing 151(3), 187–193 (2004).
6. Anusuya, M.A., Katti, S.K.: Comparison of different speech feature extraction techniques with and without wavelet transform to Kannada speech recognition, International Journal of Computer Applications 26(4), 19–23 (2011).
7. Ranjan, S.: A Discrete Wavelet Transform Based Approach to Hindi Speech Recognition. In: International Conference on Signal Acquisition and Processing, ICSAP 2010, pp.345-348 (2010).
8. Turner, C., Joseph, A., Aksu M., Langdon H.: The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent Filterbank Based Speaker Recognition. Complex Adaptive Systems, vol. 1, 124-129 (2011).
9. Tufekci, Z., Gowdy, J.N.: Feature extraction using discrete wavelet transform for speech recognition. In: Proc. of IEEE Southeast con 2000, pp. 116–123 (2000).
10. Tan, B.T., Fu, M., Spray, A., Dermody, P.: The Use of Wavelet Transforms in Phoneme Recognition. In: ICSLP 1996: Fourth International Conference on Spoken Language Processing, pp. 148–155 (1996).
11. Modic, R., Lindberg, B., Petek, B.: Comparative wavelet and mfcc speech recognition experiments on the slovenian and english speechdat2. In: Proc. Isca-ITRW NOLISP (2003).
12. Sunny, S., Peter, S., Poulose, J.: Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words. International Journal of Computer Applications 38(9), 9-13 (2012).
13. Tavanaei, A., Manzuri, M.T., Sameti, H.: Mel-scaled discrete wavelet transform and dynamic features for the Persian phoneme recognition. In: Int. Symp. Artificial Intelligence and Signal Processing (AISP), Tehran, pp. 138–140 (2011).
14. Panwar, M., Sharma, R., Khan, I., Farooq, O.: Design of Wavelet Based Features for Recognition of Hindi Digits. In: Intl. Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT 2011), pp. 232–235 (2011).
15. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley, Cambridge (1996).
16. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, London (1999).
17. Anusuya, M.A., Katti, S.K.: Front end analysis of speech recognition: a review. International Journal of Speech Technology, vol. 14, 99-145 (2010).
18. Sanchez, F.L., et al.: Wavelet-Based Cepstrum Calculation. Journal of Computational and Applied Mathematics 227, 288-293 (2009).
19. Adam, T.B., Salam, M.D.: Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks. International Journal of Computer Applications 42(12), 21-27 (2012).
20. Martinez-Garcia, M.A.: Métodos de Reconocimiento de Palabras Aisladas Usando Segmentación Acústica y Cuantización Vectorial. M.S. Thesis, Faculty of Engineering, National Autonomous University of Mexico (1998).
21. TI 46 Word Speaker-Dependent Isolated Word Corpus, NIST Speech Disc 7-1.1 (1991).
22. Rajeswari, N. P., Sathyanarayana, V.: Robust Speech Recognition Using Wavelet Domain Front End and Hidden Markov Models. Emerging Research in Electronics, Computer Science and Technology, 435 (2014).
23. Srivastava, S., Bhardwaj, S., Bhandari, A., Gupta, K., Bahl, H., & Gupta, J. R. P.: Wavelet Packet Based Mel Frequency Cepstral Features for Text Independent Speaker Identification. In Intelligent Informatics, 237-247 (2013).