Using Signal Processing Based on Wavelet Analysis to Improve Automatic Speech Recognition on a Corpus of Digits

José Luis Oropeza Rodríguez, Mario Jiménez Hernández,
and Alfonso Martínez Cruz

Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Batiz esq Miguel Othon de Mendizabal s/n, 07038, México DF, Mexico
joropeza@cic.ipn.mx, mjimenezh@ipn.mx, almart001@hotmail.com

Abstract. This paper shows results when we used wavelets in a corpus of digits pronounced by five speakers in Spanish language. One of the most important aspects related to the ASR is to reduce the number of data used. Firstly, we show the results when we used wavelet filters to the speech signal in order to obtain low frequencies only. Secondly, we use the ability of wavelet analysis is to perform data compression, to reduce by half the amount of voice data analyzed. For each of the two previous experiments we obtained new corpus, after that each corpus was used to train an Automatic Speech Recognition System using the technique vector quantization (VQ), being more employed in these corpus, also at final we compare the results obtained with DHMM technique. Finally, we compare our results with respect to the original corpus and found a 3-5% reduction in Word Error Rate (WER) when we use VQ and (1-3%) using CHMM. Daubechies wavelets were used in the experiments, as well as Vector Quantization (VQ) with Linear Prediction Coefficients (LPC) as features to represent the speech signal and DHMM.

Keywords: Wavelets, automatic speech recognition systems (ASRs), Daubechies wavelet, wavelet filters, vector quantization (VQ), discrete hidden Markov models (DHMM), data speech compression.

1 Introduction

Automatic Speech Recognition systems (ASRs) work reasonably well in quiet conditions but work poorly under noisy conditions or distorted channels. For example, the accuracy of a speech recognition system may be acceptable if you call from the phone in your quiet office, yet its performance can be unacceptable if you try to use your cellular phone in a shopping mall. The researchers in the speech group are working on algorithms to improve the robustness of speech recognition system to high noise levels channel conditions not present in the training data used to build the recognizer [Cole et al., 1996]. In this paper we used wavelet filters to avoid high frequencies, where noisy signals regularly are present. That is very important because of one of the most important aspects that have been studied for the last years is that the important information of the speech signals is contained within low frequencies more than high frequencies. “The schools of thought in speech recognition” describe four different approach researched at today [6], we are going to use template-based approach.

ASR has implemented one stage called “speech analysis”. The applications that need voice processing (such as coding, synthesis, and recognition) require specific representations of speech information. For instance, the main requirement for speech recognition is the extraction of speech features, which may distinguish different phonemes of a language. In the template-based approach, the units of speech (usually words, like in this work), are represented by templates in the same form as the speech input itself. Distance metrics are used to compare templates to find the best match, and dynamic programming is used to resolve the problem of temporal variability. Tem-plate-based approaches have been successful, particularly for simple applications requiring minimal overhead. We used this approach in this paper. A variety of techniques have been developed to efficiently represent to speech signals in digital form for either transmission or storage. Since most of the speech energy is contained in the lower frequencies results very important to encode the lower-frequency band with more bits than the high-frequency band. In this paper we based our experiments in the last aspect, we use subband coding which is a method where the speech signal is subdivided into several bands and each band is digitally encoded separately. In subband coding a speech signal is sampled at a rate Fs samples per second. The first frequency subdivision splits the signal spectrum into two equal-width segments, a lowpass signal $0 \leq F \leq F_x / 4$, and a highpass signal $F_s / 4 \leq F \leq F_s / 2$. The second frequency subdivision splits the lowpass signal from the first stage into two equal bands, a lowpass signal $0 \leq F \leq F_s / 8$, and a highpass signal $F_s / 8 \leq F \leq F_s / 4$. Finally, the third fre-quency division splits the lowpass signal from the second stage into two equal band-width signals. Thus the signal is subdivided into four frequency bands, covering three octaves. To perform subband coding we use wavelet analysis as implemented in [7].

1.1 Phonetic and Frequency Analysis

One aspect related with propose mentioned above is that /s/ phoneme, that we can consider as unvoiced fricative sound, contains representative frequencies lower than 3.5 kHz. For that, if the noise components integrated into speech signal were removed using a filter and the response filter were used into Automatic Speech Recognition systems to try to reduce the WER, we can obtain a methodology that can be used in ASR. Now, wavelet filter is newest approach used in digital signal processing, their characteristics are better (in some cases) than techniques employed in classical digital filter.

1.2 Wavelet Compression

Wavelet compression is a form of data compression well suited for image compression (sometimes also video compression and audio compression). The goal is to store image data in as little space as possible in a file. Wavelet compression can be either lossless or lossy (JPEG 2000, for example, may use a 5/3 wavelet for lossless (reversible) transform and a 9/7 wavelet for loss (irreversible transform) [Van Fleet, 2008]. Using a wavelet transform, the wavelet compression methods are adequate for representing transients, such as percussion sounds in audio, or high-frequency components in two-dimensional images, for example an image of stars on a night sky. This means that the transient elements of a data signal can be represented by a smaller amount of information than would be the case if some other transform, such as the more widespread discrete cosine transform, had been used.

1.3 Speech Recognition using Wavelets

Taking into account all that we mentioned above for researcher interested in speech recognition is very interesting analyze how can we use compression properties related with wavelet analysis into a corpus of speech signals that it will be used into Automatic Speech Recognition tasks. The results obtained not only have impact over the task mentioned above, but it could be used for a combination of data compression that it will be used for transmission of information into digital networks, for example. Wavelets was used for one important reason, its computational implementation as digital filter and in compression signal is easy to make and, due to that multi-resolution analysis inherent wavelet systems, a reduced number of coefficients are necessary to be implemented in comparison with digital filters design traditional (FIR) or instability conditions (IIR) [Van Fleet, 2008].

An important amount of works have been realized in this aspect, above all for robust speech recognition task, it refers to the need to maintain good recognition accuracy even when the quality of the input speech is degraded, or when the acoustical, articulatory, or phonetic characteristics of speech in the training and testing environments differ. Obstacles to robust recognition include acoustical degradations produced by additive noise, the effects of linear filtering, nonlinearities in transduction or transmission, as well as impulsive interfering sources, and diminished accuracy caused by changes in articulation produced by the presence of high-intensity noise sources. Even systems, that are designed to be speaker-independent, exhibit dramatic degradations in recognition accuracy when training and testing conditions differ [2]. For these reasons, within others, is necessary to find an alternative that can be used to reduce the conflicts presented here.

2 Wavelet Theory

Fourier analysis, using the Fourier transform, is a powerful tool for analyzing the components of a stationary signal. For example, the Fourier transform is a powerful tool for processing signals that are composed of some combination of sine and cosine signals. The Fourier transform is less useful in analyzing non-stationary data, where there is no repetition within the region sampled. Wavelet transforms (of which there are, at least formally, an infinite number) allow the components of a non-stationary signal to be analyzed. Wavelets also allow filters to be constructed for stationary and non-stationary signals [3].

The statistics of many natural images are simplified when they are decomposed via wavelet transform. Recently, many researchers have found that statistics of order greater than two can be utilized in choosing a basis for images, has shown that the coefficients of frequency subbands of natural scenes have much higher kurtosis than a Gaussian distribution. Daubechies wavelets are a family of orthogonal wavelets defining a discrete wavelet transform and characterized by a maximal number of vanishing moments for some given support [3]. With each wavelet type of this class, there is a scaling function (also called mother wavelet) which generates an orthogonal multiresolution analysis. The selected sub-band was the 3 in all case, because the high level recognition was better for this case. A signal or function *f (t)* can often be better analyzed, described, or processed if expressed as a linear decomposition by [4].

$$f(t) = \sum_{k \in Z} c_{j,k} \phi_{j,k}(t) \tag{1}$$

Where $\phi_{j,k}(t)$ is an integer index for the sum, is the expansion coefficients and is the set of real-valued functions of *t* called the expansion set. If the expansion is unique, the set is called a basis for the functions that could be represented. If the basis is orthogonal, then the coefficients can be calculated by the *inner product.* [Walnut, 2001]

$$c_{j,k} = \left\langle f(t), \phi_{j,k}(t) \right\rangle = \int_{-\infty}^{\infty} \left| f(t) \phi_{j,k}(t) \right| dt \tag{2}$$

A single $\phi_{j,k}(t)$ coefficient is obtained by substituting (1) into (2) and therefore for the *wavelet expansion,* a two-parameter system is constructed such that (1) becomes

$$f(t) = \sum_{k} \sum_{j} c_{j,k} \phi(t) + \sum_{k} \sum_{j} d_{j,k} \psi(t) \tag{3}$$

Where $\{\psi_{j,k}(t) \mid j,k \in Z\}$ and $\phi_{j,k}(t)$ is the wavelet expansion that usually forms an orthogonal basis. The set of expansion coefficients $\psi_{j,k}(t)$ are called the discrete wavelet transform of *f (t)* and (3) is its inverse.

All wavelet systems are generated from a single scaling function or wavelet by simple scaling and translation. This two-dimensional representation is achieved from the function $\psi_{j,k}(t)$, also called the mother wavelet, by

$$\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi\left(2^{j}t-k\right) \qquad j,k \in Z \tag{4}$$

Wavelet systems also satisfy multi-resolution conditions. In effect, this means that a set of scaling functions can be determined in terms of integer translates of the basic scaling function by

$$\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi(2^{j}t-k) \qquad j,k \in Z \tag{5}$$

It can therefore be seen that if a set of signals can be represented by $\phi(t-k)$, a larger set can represented by $\phi(2t-k)$, giving a better approximation of any signal.

Hence, due to the spanning of the space of $\phi(t)$ by $\phi(2t)$, it can be expressed in terms of the weighted sum of the shifted as

$$\phi(t) = \sum_{k\in Z} h(n)\sqrt{2}\phi(2t-k) \tag{6}$$

Where the coefficients *h(n)* may be real or complex numbers called the scaling function coefficients. However, the important features of a signal can better be described, not by $\phi(2t)$, but by defining a slightly different set of functions $\psi_{j,k}(t)$ that span the differences between the spaces spanned by the various scales of the scaling function. These functions are the wavelets and, they can be represented by a weighted sum of shifted scaling function $\phi_{j,k}(t)$ defined in (6) by [Yuan Yan, 2009].

$$\psi(t) = \sum_{k\in Z} \sqrt{2}h(n)\phi(2t-k) \tag{7}$$

The function generated by (7) gives the prototype or mother wavelet $\psi(t)$ for a class of expansion functions of the form given by (4).

$$f(t) = \sum_{k}\sum_{j} c_{j,k}\phi(t) + \sum_{k}\sum_{j} d_{j,k}\psi(t) \tag{8}$$

The coefficients in this wavelet expansion are called the discrete wavelet transform (DWT), of the signal *f(t)*. For a large class of signals, the wavelet expansion coefficients drop off rapidly as *j* and *k* increase. As a result, the DWT is efficient for image and audio compression

3 Template-Based Approach

The frequency bandwidth of a speech signal is about 16 KHz. However, most of speech energy is under 7 KHz. Speech bandwidth is generally reduced in recording. A speech signal is called orthophonic if all the spectral components over 16 KHz are discarded. A telephonic lower quality signal is obtained whenever a signal does not have energy out of the band 300-3400 Hz. Therefore, digital speech processing is usually performed by a frequency sampling ranging between 8000 samples/sec and 32000 samples/sec. These values correspond to a bandwidth of 4KHz and 16KHz respectively. In this work, we use a frequency sampling 11025 samples/sec [1].

4 Experiments and Results

In the proposal developed in the experiments we used the methodology represented in figure 6, this figure shows the experiment developed using wavelets theory to reduce the amount of noise contained in the speech signal, it was obtained taking into account the frequency allocation in a (conventional, decimated) DWT where we have A1 and D1 frequency allocation to represent such Approximations as Details obtained from wavelet filter used, the actual shape depends of the wavelet filters. Also, figure 6 shows the experiment developed using wavelet theory to applied data compression, DWT decomposition were used with the difference that we token the speech signal from different points.

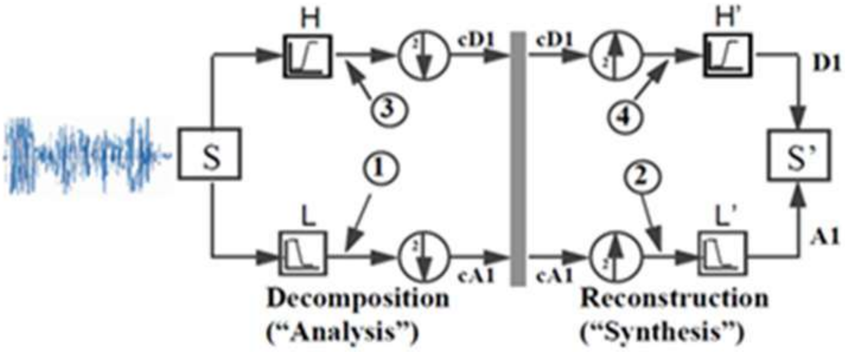


Fig. 1. Analysis wavelet used in the experiments.

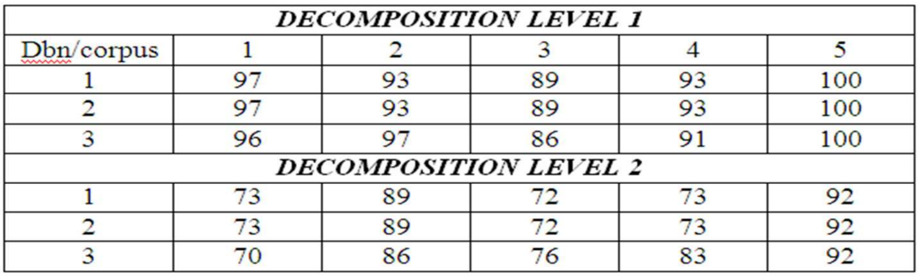
For experiments, a Spanish digits database (0-9) was used, it consists of 5 speakers each one of them pronounced 200 sentences of digits, half of them were used to training system and the rest for recognition. Speech signals were recorded at 11025 kHz with 16 bits by sample, using mono-channel with PCM format. As we mentioned before and after to describe the proposal developed in this paper, we obtained a total of 600 sentences for each speaker obtained of a great total of 3000 sentences processed during experiments reported. As we mentioned before and after to describe the proposal developed in this paper, we obtained a total of 600 sentences for each speaker obtained of a great total of 3000 sentences processed during experiments reported.

The signal flow diagram for a single-level conventional DWT is shown in figure 1. The information used to eliminate the noise was taken after of the block denoted as L, while the information used for compression was taken after of the first label cA1 as we shown in the same figure, where down-sampling for 2 was reached. As we mentioned before and after to describe the proposal developed in this paper, we obtained a total of 600 sentences for each speaker obtained of a great total of 3000 sentences processed during experiments reported. The signal flow diagram for a single-level conventional DWT is shown in figure 6. The information used to eliminate the noise was taken after of the block denoted as L, while the information used for compression was taken after of the first label cA1 as we shown in the same figure.

Table 1 shows results obtained using speech signals after to apply wavelet compression, we can see that performance reached was superior using decomposition level 1, for that in all experiments reported here this decomposition level was selected.

In rows we can see wavelet Daubechies order (form 1 to 3) and in the columns we can see the 5 corpus analyzed, the performance reached also is showed. We can see clearly that the best performance was obtained using decomposition level 1 in comparison with decomposition level 2.

Table 1. Results using wavelet filters.

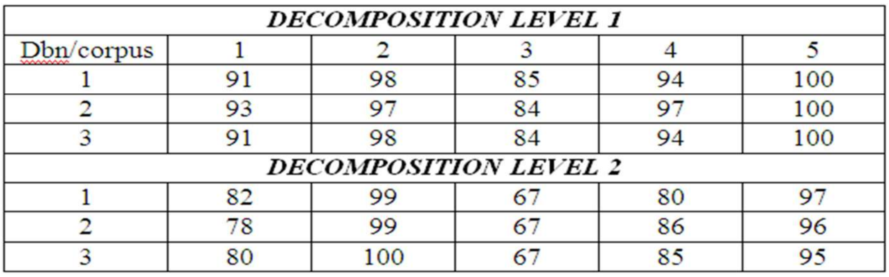


DECOMPOSITION LEVEL 1					
Dbn/corpus	1	2	3	4	5
1	97	93	89	93	100
2	97	93	89	93	100
3	96	97	86	91	100
DECOMPOSITION LEVEL 2					
1	73	89	72	73	92
2	73	89	72	73	92
3	70	86	76	83	92

Another results obtained is illustrated in table 2 which reports obtained using speech signals after to apply lowpass filter, again we can see that the best performance is obtained when we used decomposition level 1.

As we can see, the results obtained using two aspects considered in this paper was successful when we used decomposition level 1, but the best performance occurs and it is most significant using wavelet compression.

Table 2. Results using wavelet filters.

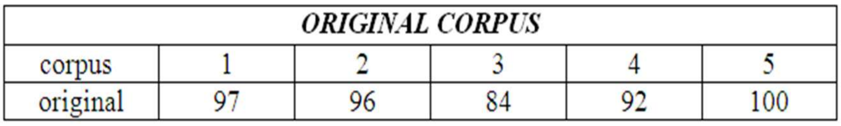


DECOMPOSITION LEVEL 1					
Dbn/corpus	1	2	3	4	5
1	91	98	85	94	100
2	93	97	84	97	100
3	91	98	84	94	100
DECOMPOSITION LEVEL 2					
1	82	99	67	80	97
2	78	99	67	86	96
3	80	100	67	85	95

Finally, we show, in table 3, the performance reached when we used the original corpus for recognition task.

Comparing the results we can conclude that the recognition percentage was better using the compression wavelet with the exception showed for the second corpus because of it had noisy signal.

Table 3 Results using original corpus



ORIGINAL CORPUS					
corpus	1	2	3	4	5
original	97	96	84	92	100

The results previously presented were compared using Discrete Hidden Markov Models (DHMM) with 6 states per model for each word into the corpus. The results obtained demonstrated that DHMM results better than VQ technique in order 3% related with the corpus used.

5 Conclusions and Future Works

Without loss of generality, such as has been demonstrated in ASRs based in VQ, we can to say that the results obtained in these experiments can be extended to others ASRs that employ Continuous Density Hidden Markov Model (CDHMM) and Mel Frequency Cepstrum Coefficients (MFCC) without problem. The main purpose of this paper was to integrate wavelet aspects such as digital filters and compression into ASRs. We showed that the methodology proposed reached out an increase of the 3-5% of the WER for some corpus created. For future works we are going to use the results obtained here to integrate the Automatic Speech Recognition for a system that is focused to speaker and speech recognition wheatear more synthesis in an application that is going to interactively it will respond to anybody people. For that, we must to work in increase the number of the speakers and programming another splitting algorithm. Though the results reported demonstrated a good and better performance.

References

1. Bechetti, C., Ricoti L.P.: Speech Recognition Theory and C++ Implementation, Fundazione Ugo Bordón. Rome, Italy. John Wiley and Sons, Ltd. (1999)
2. Cole, A.R., Mariani,J., Uszkoreit, H., Zaenen, A., Zue, V.:. Survey of the State of the Art in Human Language Technology, National Science Foundation, CSLU, Oregon Institute.
3. Duran, D.I.: The Wavelet Transform, Time-Frequency Localization and Signal Analysis. IEEE Transactions on Information Theory, Vol. 36, No. 5 (1990)
4. Faundez, P., Fuentes, A.: Procesamiento de señales acústicas utilizando wavelets. Instituto de Matemáticas, UACH.
5. Farnetani, E.: Coarticulation and connected speech processes. In: The Handbook of Phonetic Sciences. W. Hardcastle and J. Laver, 12 . Blackwell, pp. 371-404 (1997)
6. Kirschning, A.I.: Automatic Speech Recognition with the parallel Cascade Neural network, PhD Thesis. Tokyo, Japan (1998)
7. Mallat, S.: A wavelet tour of signal processing. Academic Press, ISBN: 0-12-4666606-X (1999)
8. Nicholas, G.R..: Fourier and wavelet representations of functions, Electronic Journal of Undergraduate Mathematics. Furman University. Volume 6, 1-12 (2000)