Using Recursive Least Squares Algorithm for Adaptive Filter and Wavelets to Increase Automatic Speech Recognition Rate in Noisy Environment

José Luis Oropeza Rodríguez[1]; Sergio Suárez Guerra[1]

[1] Center for Computing Research, National Polytechnic Institute,
Juan de Dios Batiz esq Miguel Othon de Mendizabal s/n, P.O. 07038, Mexico
joropeza@cic.ipn.mx, ssuarez@cic.ipn.mx

Abstract. This paper shows results obtained in the Automatic Speech Recognition (ASR) task for a corpus of digits speech files with a determinate noise level immerse. In the experiments, we used several speech files that contained Gaussian noise. We used HTK (Hidden Markov Model Toolkit) software of Cambridge University in the experiments. The noise level added to the speech signals was varying from fifteen to forty dB increased by a step of 5 units. We used the Recursive Least Squares Algorithm for an Adaptive Filtering (RLSAAF) to reduce the level noise and two different wavelets (Haar and Daubechies). With RLSAAF we obtained an error rate lower than if it was not present and it was better than wavelets employed for this experiment of Automatic Speech Recognition. For decreasing the error rate we trained with 50% of contaminated and originals signals to the ASR system. The results showed in this paper are focused to try analyses the ASR performance in a noisy environment and to demonstrate that if we are controlling the noise level and if we know the application where it is going to work, then we can obtain a better response in the ASR tasks. Is very interesting to count with these results because speech signal that we can find in a real experiment (extracted from an environment work, i.e.), could be treated with these technique and we can decrease the error rate obtained. Finally, we report a recognition rate of 99%, 97.5% 96%, 90.5%, 81% and 78.5% obtained from 15, 20, 25, 30, 35 and 40 noise levels, respectively when the corpus mentioned before was employed and RLSAAF algorithm was used. Haar wavelet level 1 reached up the most important results as an alternative to RLSAAF algorithm, but only when the noise level was 40 dB and using original corpus.

Keywords. Automatic Speech Recognition, Haar wavelets, Daubechies wavelet, Recursive Least Squares Algorithm for an Adaptive Filtering and noisy speech signal, noisy reduction.

Computer Engineering.
Research in Computing Science 30, 2007, pp. 13-25

1 Introduction

Speech recognition systems generally assume that the speech signal is a realisation of some message encoded as a sequence of one or more symbols. To effect the reverse operation of recognising the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform on the basis that for the duration covered by a single vector (typically 10ms or so), the speech waveform can be regarded as being stationary. Although this is not strictly true, it is a reasonable approximation.

Typical parametric representations in common use are smoothed spectra or linear prediction coefficients plus various other representations derived from these.

The role of the recogniser is to effect a mapping between sequences of speech vectors and the wanted underlying symbol sequences. Two problems make this very difficult. Firstly, the mapping from symbols to speech is not one-to-one since different underlying symbols can give rise to similar speech sounds. Furthermore, there are large variations in the realised speech waveform due to speaker variability, mood, environment, etc. Secondly, the boundaries between symbols cannot be identified explicitly from the speech waveform. Hence, it is not possible to treat the speech waveform as a sequence of concatenated static patterns.

The second problem of not knowing the word boundary locations can be avoided by restricting the task to isolated word recognition. As shown in Fig. 1.2, this implies that the speech waveform corresponds to a single underlying symbol (e.g. word) chosen from a fixed vocabulary. Despite the fact that this simpler problem is somewhat artificial, it nevertheless has a wide range of practical applications. Furthermore, it serves as a good basis for introducing the basic ideas of HMM-based recognition before dealing with the more complex continuous speech case. Hence, isolated word recognition using HMMs will be dealt with first.

The different sources of variability that can affect speech determine most of difficulties of speech recognition. During speech production the movements of different articulators overlap in time for consecutive phonetic segments and interact with each other. As a consequence, the vocal tract configuration at any time is influenced by more than one phonetic segment. This phenomenon is known as coarticulation. The principal effect of the coarticulation is that the same phoneme can have very different acoustic characteristics depending on the context in which it is uttered [Farnetani 97].

Speech recognition-system performance is also significantly affected by the acoustic confusability or ambiguity of the vocabulary to be recognized. A confusable vocabulary requires detailed high performance acoustic pattern analysis. Another source of recognition-system performance degradation can be described as variability and noise.

State-of-the-art ASR systems work pretty well if the training and usage conditions are similar and reasonably benign. However, under the influence of noise, these systems begin to degrade and their accuracies may become unacceptably low in severe environments [Deng and Huang 2004]. To remedy this noise robustness issue in ASR due to the static nature of the HMM parameters once trained, various adaptive techniques have been proposed. A common theme of these techniques is the utilization of some form of compensation to account for the effects of noise on the speech characteristics. In general, a compensation technique can be applied in the signal, feature or model space to reduce mismatch between training and usage conditions [Huang at el. 2001].

2 Characteristics and Generalities

Speech recognition systems 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

Robustness in speech recognition 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. Some of these sources of variability are illustrated in Figure 1. Speaker-to-speaker differences impose a different type of variability, producing variations in speech rate, co-articulation, context, and dialect, even systems that are designed to be speaker independent exhibit dramatic degradations in recognition accuracy when training and testing conditions differ [Cole & Hirschman 92].

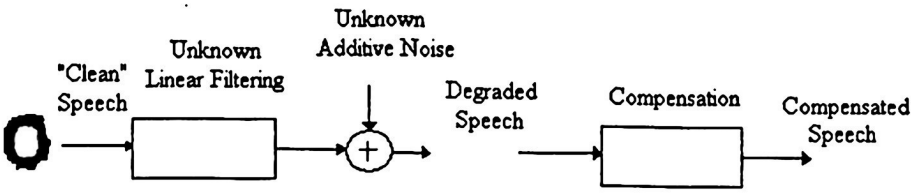

Fig. 1 Schematic representation of some of the sources of variability that can degrade speech recognition accuracy, along with compensation procedures that improve environmental robustness.

3 Automatic Speech Recognition Systems

Automatic Speech Recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols. The ASR is constitutive by: training and recognition stages. Voice is a static procedure that can to have a duration time between 80-200 ms. a simple but effective mathematical model of the physiological voice production process is the excitation and vocal tract model.

The excitation signal is assumed periodic with a period equal to the pitch for vowels and other voiced sounds, while for unvoiced consonants, the excitation is assumed white noise, i.e. a random signal without dominant frequencies. The excitation signal is subject to spectral modifications while it passes through the vocal tract that has an acoustic effect equivalent to linear time invariant filtering. The model is relevant because, for each type of excitation, a phoneme (or another structural linguistic) is identified mainly by considering the shape of the vocal tract. Therefore, the vocal tract configuration can be estimated by identifying the filtering performed by the tract vocal on the excitation. Introducing the power spectrum of the signal $P_x(\omega)$, of the excitation $P_v(\omega)$ and the spectrum of the vocal tract filter $P_h(\omega)$, we have:

$$P_x(\omega) = P_v(\omega)P_h(\omega) \qquad [1]$$

The speech signal (continuous, discontinuous or isolated) is first converted to a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform on the basis that for the duration covered by a single vector (typically 10-25 ms) the speech waveform can be regarded as being stationary. Although it is not strictly true, it is a reasonable approximation. Typical parametric representations in common use are smoothed spectra or linear predictive coefficients plus various other representations derived from these. The database employed consists of ten digits (0-9) for the Spanish language. Many of the operations performed by HTK (Hidden Markov Model Toolkit) which involve speech data assumes that the speech is divided into segments and each segment has a name or label. The set of labels associated with the speech data will be the same as corresponding speech file but a different extension.

4 Hidden Markov Models

As we know, HMMs mathematical tool applied for speech recognition presents three basic problems [Rabiner and Biing-Hwang, 1993] y [Zhang 1999]. For each state, the HMMs can use since one or more Gaussian mixtures both to reach high recognition rate and modeling vocal tract configuration in the Automatic Speech Recognition.

Gaussian mixtures

Gaussian Mixture Models are a type of density model which comprise a number of functions, usually Gaussian. These component functions are combined to provide a multimodal density. They can be employed to model the colors of an object in order to perform tasks such as real-time color-based tracking and segmentation. In speech recognition, the Gaussian mixture is of the form [Bilmes 98] [Resch, 2001a], [Resch, 2001b], [Kamakshi et al., 2002] and [Mermelstein, 1975].

:

$$g(\mu,\Sigma)(x)=\frac{1}{\sqrt{2\pi^d}\sqrt{\det(\Sigma)}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \qquad [2]$$

Equation 2-3 shows a set of Gaussian mixtures:

$$gm(x)=\sum_{k=1}^{K}w_k * g(\mu_k,\Sigma_k)(x) \qquad [3]$$

In 4, the summarize of the weights give us

$$\sum_{i=1}^{K}w_i=1 \quad \forall \quad i\in\{1,.....,K\} \quad : w_i \geq 0 \qquad [4]$$

Viterbi Trainning

We used Viterbi training, in this a set of training observations O^r, $1 \leq r \leq R$ is used to estimate the parameters of a single HMM by iteratively computing Viterbi alignments. When used to initialise a new HMM, the Viterbi segmentation is replaced by a uniform segmentation (i. e. each training observation is divided into N equal segments) for the first iteration.

Wavelets Transform

This section shows an introductory description about wavelet analysis, includes a discussion of different wavelet functions

Windowed Fourier Transform

The WFT represents one analysis tool for extracting local-frequency information from a signal. The Fourier transform is performed on a sliding segment of length T from a time series of time step δt and total length $N\delta t$, thus returning frequencies from

$T-1$ to $(2\delta t)-1$ at each time step. The segments can be windowed with an arbitrary function such as a boxcar (no smoothing) or a Gaussian window.

As discussed by Kaiser (1994), the WFT represents an inaccurate and inefficient method of time–frequency localization, as it imposes a scale or "response interval" T into the analysis. The inaccuracy arises from the aliasing of high- and low-frequency components that do not fall within the frequency range of the window. The inefficiency comes from the $T/(2\delta t)$ frequencies, which must be analyzed at each time step, regardless of the window size or the dominant frequencies present. In addition, several window lengths must usually be analyzed to determine the most appropriate choice. For analyses where a predetermined scaling may not be appropriate because of a wide range of dominant frequencies, a method of time–frequency localization that is scale independent, such as wavelet analysis, should be employed [Torrence Christopher and Compto Gilbert, 1198].

Wavelet Transform

The *wavelet transform* can be used to analyze time series that contain nonstationary power at many different frequencies (Daubechies 1990). Assume that one has a time series, x_n, with equal time spacing δt and $n = 0 \ldots N - 1$. Also assume that one has a *wavelet function,* $\psi_0(\eta)$, that depends on a nondimensional "time" parameter η. To be "admissible" as a wavelet, this function must have zero mean and be localized in both time and frequency space (Farge 1992). An example is the Morlet wavelet, consisting of a plane wave modulated by a Gaussian:

$$\psi_0(\eta) = \pi^{-1/4} e^{j\omega_0 \eta} e^{-\eta^2/2} \qquad [5]$$

where ω_0 is the nondimensional frequency, here taken to be 6 to satisfy the admissi bility condition (Farge 1992). This wavelet is shown in Fig. 2a.

The term "wavelet function" is used generically to refer to either orthogonal or nonorthogonal wavelets. The term "wavelet basis" refers only to an orthogonal set of functions. The use of an orthogonal basis implies the use of the *discrete wavelet transform,* while a nonorthogonal wavelet function can be used

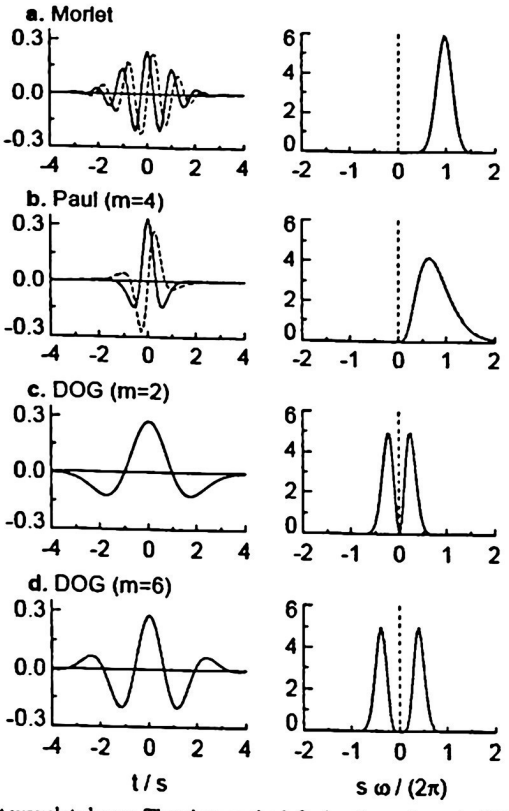

Fig.2. Four different wavelets bases. The plots on the left give the real part (solid) and imaginary part (dashed) for the wavelets in the time domain. The plots on the right give the corresponding wavelets in the frequency domain. (a) Morlet, (b) Paul (m=4), (c) Mexican Hat (m=2), and d) Mexican Hat (m=6)

The continuous wavelet transform of a discrete sequence *xn* is defined as the convolution of x_n with a scaled and translated version of $\psi_0(\eta)$:

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^* \left[\frac{(n'-n)\delta t}{s} \right] \qquad [6]$$

where the (*) indicates the complex conjugate. By varying the *wavelet scale s* and translating along the *localized time index n*, one can construct a picture showing both the amplitude of any features versus the scale and how this amplitude varies with time. The subscript 0 on ψ has been dropped to indicate that this ψ has also been normalized (see next section). Although it is possible to calculate the wavelet transform using (6), it is considerably faster to do the calculations in Fourier space.

To approximate the continuous wavelet transform, the convolution (6) should be done N times for each scale, where N is the number of points in the time series (Kaiser 1994). (The choice of doing all N convolutions is arbitrary, and one could choose a smaller number, say by skipping every other point in n.) By choosing N points, the convolution theorem allows us do all N convolutions simultaneously in Fourier space using a discrete Fourier transform (DFT). The DFT of x_n is

$$\tilde{x}_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \qquad [7]$$

where $k = 0 \ldots N-1$ is the frequency index. In the continuous limit, the Fourier transform of a function $\psi(t/s)$ is given by $\tilde{\psi}(s\omega)$. By the convolution theorem, the wavelet transform is the inverse Fourier transform of the product:

$$W_n(s) = \sum_{k=0}^{N-1} \tilde{x}_k \tilde{\psi}^*(s\omega_k) e^{j\omega_k n\delta t} \qquad [8]$$

where the angular frequency is defined as

$$\omega_k = \begin{cases} \dfrac{2\pi k}{N\delta t}: & k \le \dfrac{N}{2} \\ -\dfrac{2\pi k}{N\delta t}: & k > \dfrac{N}{2} \end{cases} \qquad [9]$$

Using (8) and a standard Fourier transform routine, one can calculate the continuous wavelet transform (for agiven s) at all n simultaneously and efficiently.

5 Experiments and Results

The evaluation of the algorithm proposed involved clustering a set of speech data consisting of 100 isolated patterns from a digits vocabulary. The training patterns (and a subsequent set of another 200 independent testing pattern) were recorded in a room free of noise. Only one speaker provided the training and testing data. All training and test recordings were made under identical conditions. The 200 independent testing patterns was addition with a level noise, we obtained a total of 1200 new sentences contaminated (200 per noise level, that is because we used 6 noise levels). After that, we used an adaptive filter to reduce that noise level and the results are shown below, then we obtained another 1200 sentences. Finally, we made experiments with a total of 2600 sentences (between noisy, filtered and clean sentences) of

speech signal. Figure 3 shows the RLSAAF employed. For each corpus created, we used three databases test to recognition task: with same characteristics, noisy and filtered. All sentences were recorded at 16 kHz frequency rate, 16 bits and mono-channel. We use MFCCs (Mel Frequency Cepstral Coefficients) with 39 characteristics vectors (differential and energy components). A Hidden Markov Model with 5 states and 1 Gaussian Mixture per state.

$$W_k = W_{k-1} + G_k e_k$$

$$P_k = \frac{1}{\gamma}\left[P_{k-1} - G_k x^T(k) P_{k-1}\right]$$

where

$$G_k = \frac{P_{k-1} x(k)}{\alpha_k}$$

$$e_k = y_k - x^T(k) W_{k-1}$$

$$\alpha_k = \gamma + x^T(k) P_{k-1} x(k)$$

with

$x(k)$ *samples*

γ *forgetting factor* $\gamma = 0.98$

P_k *recursive way to determinate* $\left[X_k^T X_k\right]^{-1}$

Fig. 3 Recursive Least Squares Algorithm for an Adaptive Filtering (RLSAAF)

This algorithm stop when the error is lest than 0.9%.

Table 1 shows the results obtained when we used a noisy corpus to training the ASR. A total of 600 speech sentences were analyzed.

Table 1 Results obtained with noisy corpus created

	speech recognition with noisy corpus created					
	noise level					
Speech signal recognized	15	20	25	30	35	40
Noisy	95,5	96,5	98,5	98	99,5	99,5
Original	57	72,5	83,5	91,5	99	99
Filtered	23	50	76,5	90,5	98	99,5

As we can see, when we used a noisy corpus like we hoped, recognition level with noisy database was adequately. When we used high S/N rate (25, 30, 35 and 40 dB), the recognition rate was increased. It is important because it significance that the

noisy corpus is a good reference. Figure 4 shows a histogram related with the table contents.

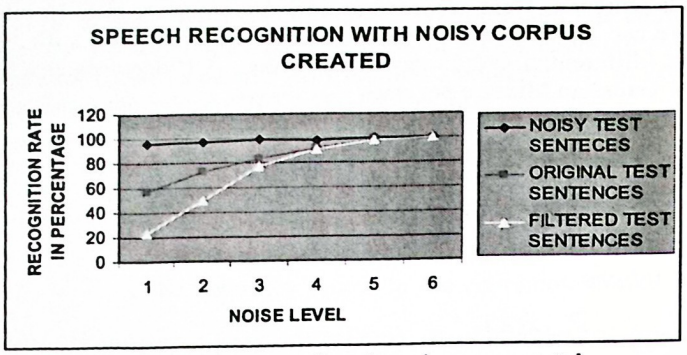

Fig. 4 Graphic representation using noisy corpus created

Table 2 shows the results obtained when we used a noisy and clean corpus to training the ASR. A total of 600 (300 noisy and 300 clean) speech sentences were analyzed.

Table 2 Results obtained with noisy and clean corpus created

	speech recognition with noisy and clean corpus created					
	noise level					
Speech signal recognized	15	20	25	30	35	40
Noisy	98,5	98	99,5	99	99,5	99,5
Original	19	34	84	91,5	96,5	99
Filtered	78,5	81	90,5	96	95,7	99

As we can see, when we used a corpus compound by noisy and original signals, the recognition rate for filtered speech signal was increased considerably. Figure 5 shows that.

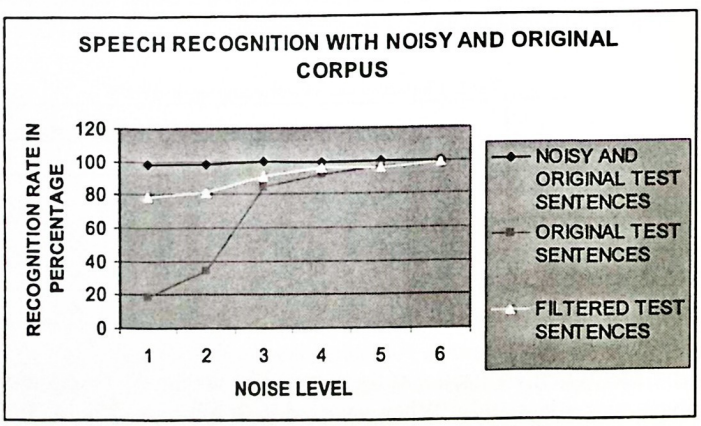

Fig. 5 Graphic representation using noisy and original corpus

Table 3 shows the results obtained when we used a clean corpus to training the ASR. A total of 600 speech sentences were analyzed.

Table 3 Results obtained with clean corpus created

	speech recognition with clean corpus created					
	noise level					
Speech signal recognized	15	20	25	30	35	40
Noisy	99,5	99,5	99,5	99,5	99,5	99,5
Original	16	21,5	18	43	70,5	87
Filtered	18.5	29	33,5	56	99,5	86,5

With the original corpus the results was not satisfactory, although the recognition rate with filtered signals was better than noisy signals, it was poor and not enough to be considered important as figure 6 shows.

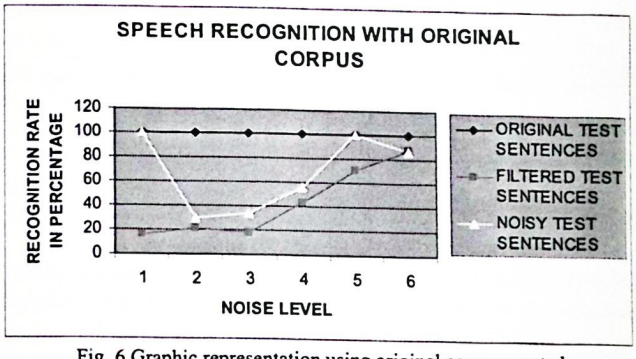

Fig. 6 Graphic representation using original corpus created

Finally, we probed different wavelets to try to determine better results than we obtained above. The results were not that we hoped.

Table 4 Results obtained with clean corpus created and wavelets

Atenuación	Haar1	Haar2	Haar3	db3n3
15 dB	20,5	12	12	15
20 dB	21	12	11,5	10
25 dB	22,5	13,5	12	21,5
30 dB	27,5	16,5	15	27,5
35 dB	43,5	13,5	12,5	28
40 dB	74,5	15	14	36

As we can see in figure 7, only Haar 1 wavelet at 40 dB had a high performance in ASR rate. We consider that results obtained were failed because noisy level selected before to apply wavelet transform must be changed. But we consider that it only can not help us so much.

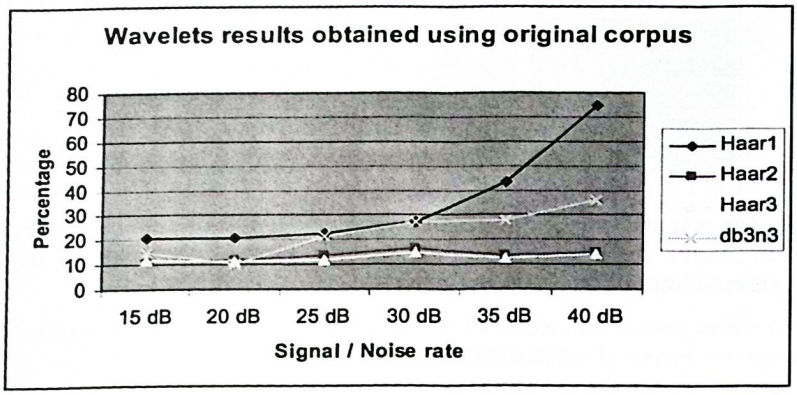

Fig. 7 Graphic representation for ASR using wavelets and original corpus

6 Conclusions and future works

The results shown in this paper demonstrate that we can use an adaptive filter to reduce the noise level in an automatic speech recognition system (ASRS) for the Spanish language. The use of this paradigm is not new but with this experiment we propose to reduce the problems find out when we tread with real speech signals. MFCCs and CDHMMs (Continuous Density Hidden Markov Models) were used for training and recognition, respectively. First, when we used database test with the same characteristics that corpus training a high performance was reached out, but when we used the clean speech database our recognition rate was poor. The most important results extracted of this experiment were when the clean speech was mixed with noisy speech, when we used filtered speech we obtained a high performance in our ASR.

For that, our conclusion is that if we want to construct an ASR immerse in a noisy environment, it is going to have a high performance if we included in our database training clean and noisy speech signal. So, if we known the Signal/Noise ratio and it are greater than 35%, we can use the filtered signal in an ASR without problems. For future works is recommendable try to probe the results obtained using another methods employed to reduce noise into signal (wavelets i. e.), and extract the results.

References

[Bilmes 98] BILMES J.A., "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", International Computer Science Institute, Berkeley, CA. 1998.

[Cole & Hirschman 92] R. A. Cole, L. Hirschman, et al. Workshop on spoken language understanding. Technical Report CSE 92-014, Oregon Graduate Institute of Science & Technology, P.O.Box 91000, Portland, OR 97291-1000 USA, September 1992.

[Deng, Li. and Huang, X. (2004)] Challenges in Adopting Speech Recognition. Communications of the ACM, Vol. 47, No. 1, pp. 69-75.

[Farnetani 97] Farnetani E., "Coarticulation and connected speech processes", in the Handbook of Phonetic Sciences, W. Hardcastle and J. Laver, Eds., Blackwell, pp. 371-404 (1997).

[Huang, C., Wang, H. and Lee, C. (2001)] An ASR Incremental Stochastic Matching Algorithm for Noisy Speech Recognition. IEEE Trans. Speech and Audio Processing, Vol 9, No. 8, pp. 866-873.

[Kamakshi et al. 2002] KAMAKSHI V. Prasad, Nagarajan T. and Murthy Hema A. "Continuous Speech Recognition Using Automatically Segmented Data at Syllabic Units". Department of Computer Science and Engineering. Indian Institute of Technology. Madras, Chennai 600-636, 2002.

[Mermelstein 1975] MERMELSTEIN Paul "Automatic Segmentation of Speech into Syllabic Units". Haskins Laboratories, New Haven, Connecticut 06510, pp. 880-883,58 (4), June 1975.

[Rabiner and Biing-Hwang 1993] RABINER Lawrence and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[Torrence Christopher and Compto Gilbert, 1198] Torrence Christopher and Compo Gilbert P. Program in Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado. Bulletin of the American Meteorological Society, Vol. 79, No. 1, January 1998.

[Zhang 1999]. ZHANG Jialu, "On the syllable structures of Chinese relating to speech recognition", Institute of Acoustics, Academia Sinica Beijing, China, 1999.