Using Adaptive Filter to Increase Automatic Speech Recognition Rate in a Digit Corpus

José Luis Oropeza Rodríguez, Sergio Suárez Guerra,
and Luis Pastor Sánchez Fernández

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,
lsanchez@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. The experiments realized treated with 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 an adaptive filtering to reduce the level noise (it was based in the Least Measure Square –LMS- algorithm). With LMS we obtained an error rate lower than if it was not present. It was obtained because of we trained with 50% of contaminated and originals signals to the ASR. The results showed in this paper to analyze the ASR performance in a noisy environment and to demonstrate that if we have 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 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 that we mentioned above was employed. Finally, we made experiments with a total of 2600 sentences (between noisy and filtered sentences) of speech signal.

Keywords: Automatic Speech Recognition, Adaptative Filters, Continuous Density Hidden Markov Models, Gaussian Mixtures and noisy speech signals.

1 Introduction

The science of speech recognition have been advanced to the state where it is now possible to communicate reliably with a computer by speaking to it in a disciplined manner using a vocabulary of moderate size. The interplay between different intellectual concerns, scientific approaches, and models, and its potential impact in society make speech recognition one of the most challenging, stimulating, and exciting fields today. The effect of the noise and filtering on clean speech in the power spectral domain can be represented as:

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 78–87, 2007.

$$P_Y(\omega_k) = |H(\omega_k)|^2 P_X(\omega_k) + P_N(\omega_k) \quad (1)$$

where $P_Y(\omega_k)$ represents the spectra of the noisy speech $y[m]$, $P_Y(\omega_k)$ represents the power spectra of the noise $n[m]$, $P_X(\omega_k)$ the power spectra of the clean speech $x[m]$, $|H(\omega_k)|^2$ the power spectra of the channel $h[m]$, and ω_k represents a particular Mel-spectra band.

To transform to the log spectral domain we apply the logarithm operator at both sides of the last expression resulting in

$$10\log_{10}(P_Y(\omega_k)) = 10\log_{10}(|H(\omega_k|^2 P_X(\omega_k) + P_N(\omega_k)) \quad (2)$$

and defining the noisy speech, noise, and clean speech,

$$\begin{aligned} y[k] &= 10\log_{10}(P_Y(\omega_k)) \\ n[k] &= 10\log_{10}(P_N(\omega_k)) \\ x[k] &= 10\log_{10}(P_X(\omega_k)) \\ h[k] &= 10\log_{10}\left(|H(\omega_k)|^2\right) \end{aligned} \quad (3)$$

results in equations

$$\begin{aligned} 10\log_{10}(P_Y(\omega_k)) &= 10\log_{10}\left(10^{\frac{x[k]+h[k]}{10}} + 10^{\frac{n[k]}{10}}\right) \\ y[k] &= x[k] + h[k] + 10\log_{10}\left(1 + 10^{\frac{n[k]-x[k]-h[k]}{10}}\right) \end{aligned} \quad (4)$$

Where h[k] is the logarithm of $|H(\omega_k)|^2$, and the similar relationships exist between n[k] and $P_N(\omega_k)$, $x[k]$ and $P_X(\omega_k)$, and y[k] and $P_Y(\omega_k)$.

Last expressions shows that is very difficult to intent to find noise and signal separately. For that is a good recommendation to eliminate the noise embedded into a speech signal using other methods. 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. This phenomenon is known as co-articulation mentioned above. The principal effect of the co-articulation is that the same phoneme can have very different acoustic characteristics depending on the context in which it is uttered [1].

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 [2]. 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 [3].

2 Characteristics and Generalities

Speech recognition systems work reasonably well in quiet conditions but work poorly under noisy conditions or distorted channels. The researchers in our speech group (Digital Signal Processing Research Group in the Center for Computing Research) are focused on algorithms to improve the robustness of speech recognition system, so we demonstrated when we employed syllables and Expert Systems for that; we have obtained very good results. Some 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 [4].

Substantial progress has also been made over the last decade in the dynamic adaptation of speech recognition systems to new speakers, with techniques that modify or warp the systems' phonetic representations to reflect the acoustical characteristics of individual speakers [5] [11] [12].

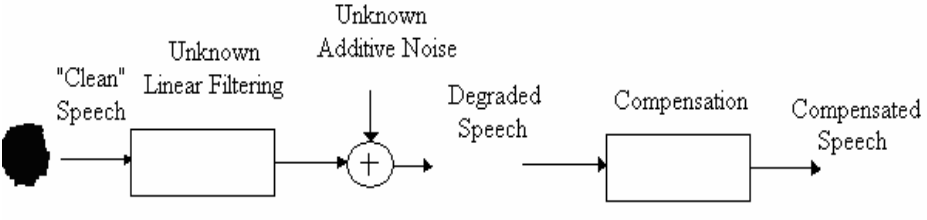

Fig. 1. Schematic representation of some of the sources of variability that can degrade speech recognition accuracy, along with compensation procedures that improve environmental robustness

3 Automatic Speech Recognition Systems

In Automatic Speech Recognition (ASR) systems most of speech energy is under 7 or 7.5 KHz (woman or man voice can change the range mentioned before) dependently. A telephonic lower quality signal is obtained whenever a signal does not have energy out of the band 300-3400 Hz. 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:

$$P_x(\omega) = P_v(\omega)P_h(\omega) \tag{5}$$

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. In our experiments we used the following block diagram for the isolated speech recognition. The database employed consists of ten digits (0-9) for the Spanish language.

4 Hidden Markov Models

As we know, HMMs mathematical tool applied for speech recognition presents three basic problems [6] y [7]. 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.

4.1 Gaussian Mixtures

Gaussian Mixture Models are a type of density model which comprise a number of functions, usually Gaussian. In speech recognition, the Gaussian mixture is of the form [8] [9], [10], [11] and [12].

$$g(\mu,\Sigma)(x) = \frac{1}{\sqrt{2\pi^d}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{6}$$

where μ *and* Σ in equation 6 represent media and standard deviation of a Gaussian function respectively. Equation 7 shows a set of Gaussian mixtures:

$$gm(x) = \sum_{k=1}^{K} w_k * g(\mu_k, \Sigma_k)(x) \tag{7}$$

In 8, the summarize of the w_k weights give us

$$\sum_{i=1}^{K} w_i = 1 \quad \forall \quad i \in \{1,.....,K\} \quad w_i \geq 0 \tag{8}$$

As we can deduce, the sum of all w_i is equal to 1. That is an interesting property of the Gaussian Mixtures employed for Automatic Speech Recognition.

4.2 Viterbi Training

We used Viterbi training, in this work for 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. Apart from the first iteration on a new model, each training sequence O is segmented using a state alignment procedure which results from maximising

$$\phi_N(T) = \max_i \phi_i(T) a_{iN} \tag{9}$$

For *1<i<N* where

$$\phi_j(t) = \left[\max_i \phi_i(t-1) a_{ij}\right] b_j(o_t) \tag{10}$$

With initial conditions given by

$$\begin{aligned} \phi_1(1) &= 1 \\ \phi_j(t) &= a_{1j} b_j(o_1) \end{aligned} \tag{11}$$

For *i<j<N*. In this and all subsequent cases, the output probability $b_j(.)$ is as defined in the following equation:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=!}^{M_{js}} c_{jsm} \aleph(o_{st}; \mu_{jsm}, \sum_{jsm}) \right]^{\gamma_s} \tag{12}$$

If A_{ij} represents the total number of transitions from state i to state j in performing the above maximisations, then the transition probabilities can be estimated from the relative frequencies

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{k=2}^{N} A_{ik}} \tag{13}$$

The sequence of states which maximises $\phi_N(T)$ implies an alignment of training data observations with states. Within each state, a further alignment of observations to mixture components is made.

We can use two methods for each state and each stream

1. use clustering to allocate each observation o_{st} with the mixture component with the highest probability
2. associate each observation o_{st} with the mixture component with the highest probability

In either case, the net result is that every observation is associated with a single unique mixture component. This association can be represented by the indicator function $\psi^r_{jsm}(t)$ which is 1 if o^r_{st} is associated with mixture component m of stream s of state j and zero otherwise.

The means and variances are then estimated via simple averages

$$\begin{aligned} \hat{\mu}_{jsm} &= \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \psi^r_{jsm}(t) o^r_{st}}{\sum_{r=!}^{R} \sum_{t=!}^{r} \psi^r_{jsm}(t)} \\ \hat{\sum}_{jsm} &= \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \psi^r_{jsm}(t)(o^r_{st} - \hat{\mu}_{jsm})(o^r_{st} - \hat{\mu}_{jsm})'}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{t=1}^{M_s} \psi^r_{jst}(t)} \end{aligned} \tag{14}$$

5 Experiments and Results

The evaluation of the adaptive filter implemented to reduce noise 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 (we employed a special software created for us to record the sentences, at 16 kbps; mono-channel and the sentence were normalized). 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 and filtered sentences) of speech signal. Figure 2 shows the adaptive filter algorithm 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 with 39 characteristics vectors (differential and energy components). A Hidden Markov Model with 6 states and 1 Gaussian Mixture per state.

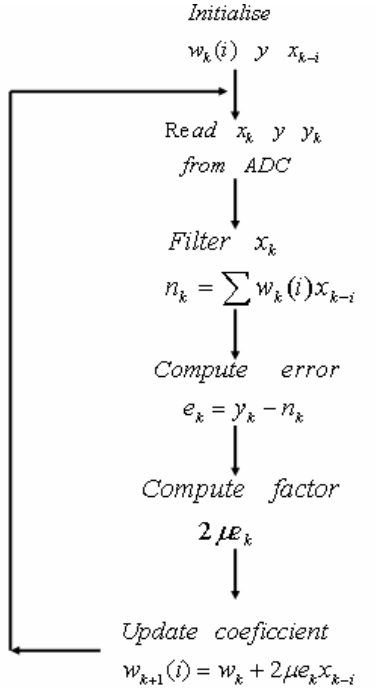

Fig. 2. Adaptive filter algorithm

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

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

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 means that the noisy corpus is a good reference. Figure 3 shows a histogram related with the table contents.

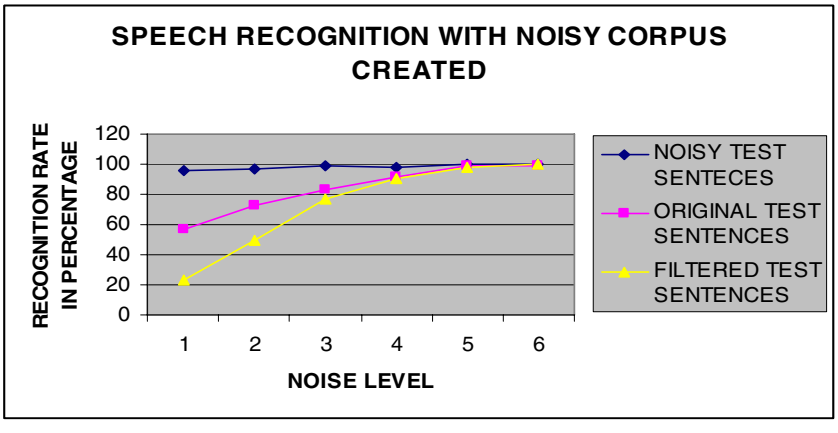

Fig. 3. Block diagram of training step for ASR for isolated words

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

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

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 4 shows that.

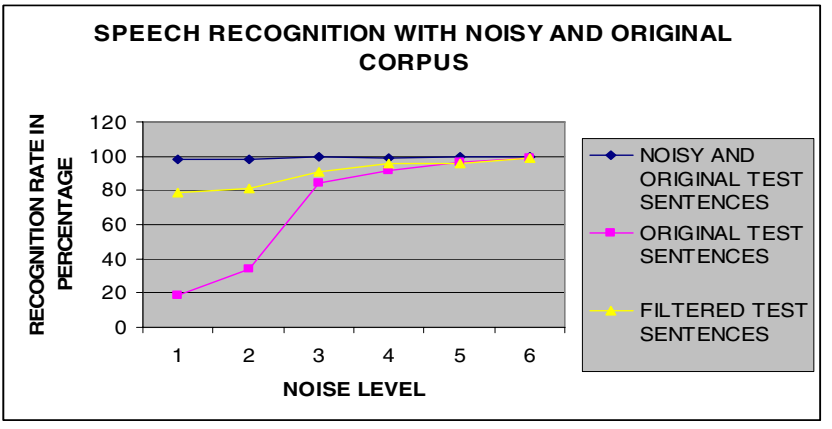

Fig. 4. Block diagram of training step for ASR for isolated words

Table 3 shows the results obtained when we used a noisy 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

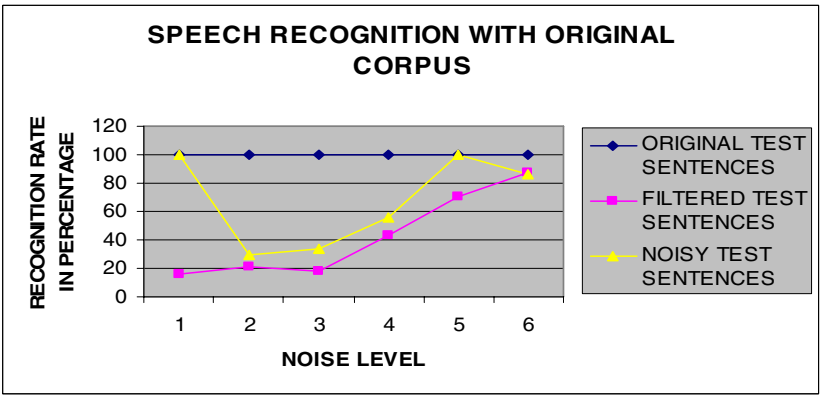

Fig. 5. Block diagram of training step for ASR for isolated words

Finally, 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 5 shows.

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 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 fixed 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's great 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

1. Farnetani, E.: Coarticulation and connected speech processes. In: Hardcastle, W., Laver, J. (eds.) Handbook of Phonetic Sciences, pp. 371–404. Blackwell (1997)
2. Challenges in Adopting Speech Recognition. Communications of the ACM 47(1), 69–75
3. An ASR Incremental Stochastic Matching Algorithm for Noisy Speech Recognition. IEEE Trans. Speech and Audio Processing 9(8), 866–873
4. Cole, R.A., Hirschman, L., et al.: Workshop on spoken language understanding. Tech. Rep. CSE 92-014, Oregon Graduate Institute of Science&Technology, P.O.Box 91000, Portland, OR 97291-1000 USA, (September 1992)
5. Gauvain, J.-L., Lee, C.-H.: Bayesian learning for HMM with GM state observation densities. In: Eurospeech (Eur91), pp. 939–942
6. Lawrence, R., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs (1993)
7. Jialu, Z.: On the syllable structures of Chinese relating to speech recognition. Institute of Acoustics, Academia Sinica Beijing, China (1999)
8. Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for GM and HMM. ICS Institute, Berkeley, CA (1998)
9. Barbara, R.: Gaussian Statistics and Unsupervised Learning. A tutorial for the Course Computational Intelligence Signal Processing and Speech Communication Laboratory (November 15, 2001), www.igi.turgaz.at/lehre/CI

10. Barbara, R.: Hidden Markov Models. A Tutorial for the Course Computational Laboratory. Signal Processing and Speech Communication Laboratory (November 15, 2001), www.igi.turgaz.at/lehre/CI
11. Prasad, K.V., Nagarajan T., Murthy H.A.: Continuous Speech Recognition Using Automatically Segmented Data at Syllabic Units. Department of Computer Science and Engineering. Indian Institute of Technology. Madras, Chennai pp. 600–636 (2002)
12. Paul, M.: Automatic Segmentation of Speech into Syllabic Units. Haskins Laboratories, New Haven, Connecticut 06510 58(4) 880–883 (1975)
13. Huang, X.D., Lee, K.F.: On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. IEEE Transactions on Speech and Audio Processing 1(2), 150–157 (1993)
14. Schwartz, R., Chow, Y., Kubala, F.: Rapid speaker adaption using a probabilistic spectral mapping. In: ICASSP [ICA87], pp. 633–636