A FEATURE EXTRACTION METHODOLOGY FOR AUTOMATIC SPEECH RECOGNITION IN NOISE ENVIRONMENTS

J. Gómez-Mena, R. García-Gómez

U.P.M., Department of S.S.R., Ciudad Universitaria s/n
28040-Madrid

1. INTRODUCTION

During the last few years a new generation of automatic speech recognizers have been implemented. These recognizers were designed according to different constraints: many deal with isolated words, others with connected or even continuous speech. Some are speaker dependent while others are speaker independent. The achievements in the speech recognition field are very encouraging. Many techniques have been tested and the designers can choose between very different alternatives for implementing their systems. However, in spite of the success of these recognizers with clean or laboratory environments, they have a limited success in real life situations. Due to this fact, there are now an increasing interest in to obtain robust recognizers capable of achieving comparable results in noise environments as the ones obtained in controlled situations. This communication is focused on the problem of increasing the robustness in the speech recognition field. Particularly we study an alternative for feature extraction.

We are going to refer to the problem of speech recognition inside cars, with the recognition of medium size vocabularies of isolated words. We describe a method for feature extraction and the scores obtained with a HMM based recognizer used to recognize isolated words in both speaker independent and speaker dependent ways. We are going to study the influence of different factors on the feature vectors, consequently although for different tasks, as large vocabularies or fluent speech, the behaviour could be different we try to obtain general conclusion on the characteristics of this methodology.

2. FEATURE EXTRACTION

The feature extraction block, in a general scheme of a speech recognizer, measures the short-time spectral envelope of speech. This spectrum could be represented in different ways. One very interesting alternative is to use LPC analysis [4] in order to obtain the spectral envelope. There are efficient algorithms to obtain the LPC parameters from the autocorrelation of speech signal. From the LPC parameters the cepstrum coefficients are obtained. Additionally from the cepstra vectors of the current frame, the previous and the next ones, we obtain the derivatives of each cepstrum. The feature vector will consist on the first p-cepstra, their p derivatives and the derivative of the logarithm of the short-time energy. These features have been used in different speech recognizers and prove to render good results. Unfortunately the above method has some important drawbacks. Perhaps the more important is its high sensivity to the background noise.

We are interested in recognizing speech with a low signal-to-noise ratio, SNR, environments, near 10 dB or even less. Consequently the feature extraction algorithm must be designed to have low sensitivity to noise and to be capable of subtracting it from speech. We like to have a method capable of the subtraction of time variant noises and also deal with impulsive ones.

Another effect which degrades the recognition scores is that speech signal passes through different filters, one during training and another different during the recognition. This situation can happen if the microphone has been changed or if the speaker speaks in a different room. Consequently we would like to have some capabilities in the extraction block for equalizing the speech spectra in order to overcome or reduce this problem. Additionally the equalization can introduce some degree of speaker normalization which is desirable in speaker independent systems.

Our feature extraction block ideally must have the following characteristics: it has low noise sensitivity, it is capable of noise subtraction and is capable of equalization. These two last characteristics can be carried out even with time varying noises and speech channels. It is based on the use of the Short-time Modified-Coherence, SMC, function, that is reviewed in the next section.Then a SMC model of noisy speech is discussed. We carried out noise subtraction and channel equalization in the SMC model. We are going to explore the efectiveness of this feature extraction methodology that combines SMC with spectral subtraction, equalization and cepstrum and its derivatives obtained from LPC.

This section ends with a description of an isolated word recognizer based on continuous density HMM and the experimental results obtained.

2.1. THE SMC REPRESENTATION

The SMC representation of the speech signal is carried out according to the diagram of Figure 2.

We assume that the input signal, x(n), to the feature extraction block has two components:

$$x(n) = L\{s(n)\} + r(n)$$

The first one is a linear transformation of the speech signal and the other a noise signal uncorrelated with the s(n).

1.- $z(n)=x(n)w(n-n_0)$; w(n) being a rectangular window of length N and n_0 the centre of the window. We use N=320 speech samples (40 msec. at 8 kHz of sampling rate). Succesive windows have an overlap of 30 msec.

2.- Short-time coherence function is defined according to [1]:

$$c(m) = \sum_{l=0}^{\frac{N}{2}-1} z(l)z(l+m); \qquad 0 \le m \le \frac{N}{2}$$

3.- Short-time modified coherence, SMC:

$$ch(m) = c(m)hw(m),\ 0 \le m \le \frac{N}{2}$$

where hw(m) is a Hamming window.

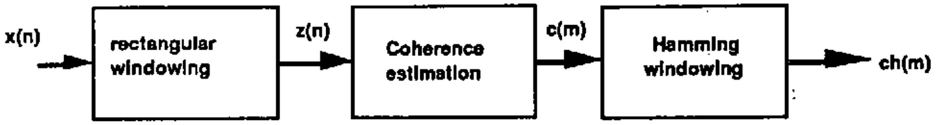

Figure 2: The SMC representation

The coherence function is a cross-correlation function between the first half on the windowed signal and the full window. If noise is present and this noise is independent of the speech signal then c(m) can be viewed as having two aditive components one is bassically the short-time correlation of the speech signal and the other of the noise. If this noise is white, only the first terms or c(m) are corrupted by the noise, ideally only c(0). In practice we have noises which are not white but its autocorrelation function is significatively more affected for small lags than the larger ones. In other words, the additive noise is not white but has a continuous spectrum, decreasing with frequency. The reason for the selection of the coherence function instead of the autocorrelation is related with the fact that for the largest lags the coherence is basically a crosscorrelation function between two signals, which are contaminated by uncorrelated noises. Of course, we pay the price of decreassing the time resolution of the spectral analysis but this is not an important drawback as we can see below. For small lags the coherence function behaves as the autocorrelation although the average is taken in a half length, this effect is diminished in ch(m) by using a Hamming window.

4.- Zero-padding of the SMC:

$$z(m) = ch(m); \qquad 1 \le m \le \frac{N}{2}$$

$$z(m) = 0; \qquad \frac{N}{2} + 1 \le m \le 256$$

5.- FFT computation:

$$Z(k) = FFT\{z(m)\}; \qquad 0 \le k \le 255$$

6.- Evaluation of the absolute value of this FFT:

$$D(k) = |Z(k)|; \qquad 0 \le k \le 255$$

7.- Evaluation of the pseudo-autocorrelation of speech signal

$$\rho(k) = FFT^{-1}\{D(k)\}; \qquad 0 \le k \le p$$

8.- Evaluation of LPC coefficients for a predictor of order p using the pseudo-autocorrelation $\rho(k)$.

In order to gain some insight in the above methodology we consider the linear prediction adjustment of the autocorrelation function. If the speech signal has a p-order LPC model then its autocorrelation has a 2p model order. In other words, if the speech signal is modelled with an all-pole filter: 1/A(z) then its autocorrelation can be modelled by

$$\frac{1}{A(z)A(z^{-1})}$$

In order to avoid the use of a 2p order predictor the square root of the spectra can be computed. Figure 3 shows a diagram of two alternatives for LPC prediction of ch(m). We estimate the autocorrelation through a FFT. In the first line of this Figure we show a method for autocorrelation estimation of the coherence function, R(k), that could be used as input to the Levinson-Durbin recursion but with the mentioned inconveniences. In the second line we include the square root transformation in order not to duplicate the order of the LPC filter. As can be observed the same transformation can be achieved by the method of the third line, reducing the number of FFTs. Then in SMC representation the central idea is to fit the predictor to the correlation function instead of to the speech signal.This idea has been used for reducing bias in spectral analysis using all-pole models [9].

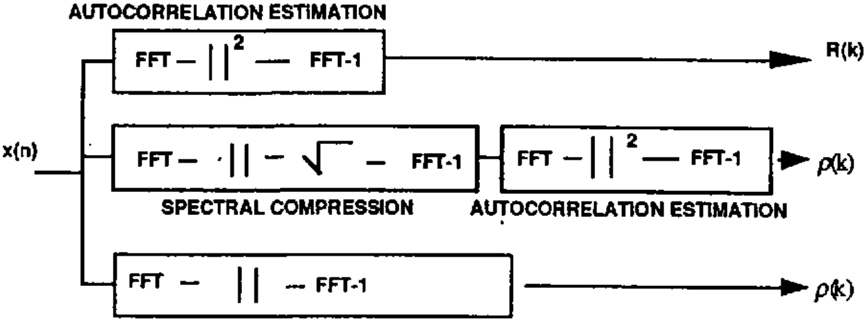

Figure 3: Interpretation of the SMC method

Of course, as proposed in [1], the method of the third line is the prefered one for reason of computational efficiency.This is the method we are going to use for pseudo-autocorrelation estimation. Since in the SMC representation we have a spectral domain representation of the speech signal it seems attractive to carry out spectral subtraction, to reduce the noise, and equalization in this domain. This is the subject of the next paragraphs.

2.2.NOISE SUBTRACTION AND SPECTRAL NORMALIZATION

Since we have a spectral domain representation of the speech signal it seems natural to take advantage of it in order to reduce the noise by means of a spectral subtraction technique. The basic assumption is that D(k) is the sum of two terms, one is speech dependent and the other noise dependent:

$$D(k) = D_s(k) + D_n(k)$$

The noise subtraction needs an estimation of the noisy part and subtraction of it from D(k). Assuming that the estimation of the noise component is:

$$NC(k)$$

If we subtract it from D(k) we can obtain negative values which do not correspond to a Fourier transform of a correlation function. In order to avoid this problem we establish a threshold T(k) and some spectral normalization as it is show in Figure 4.

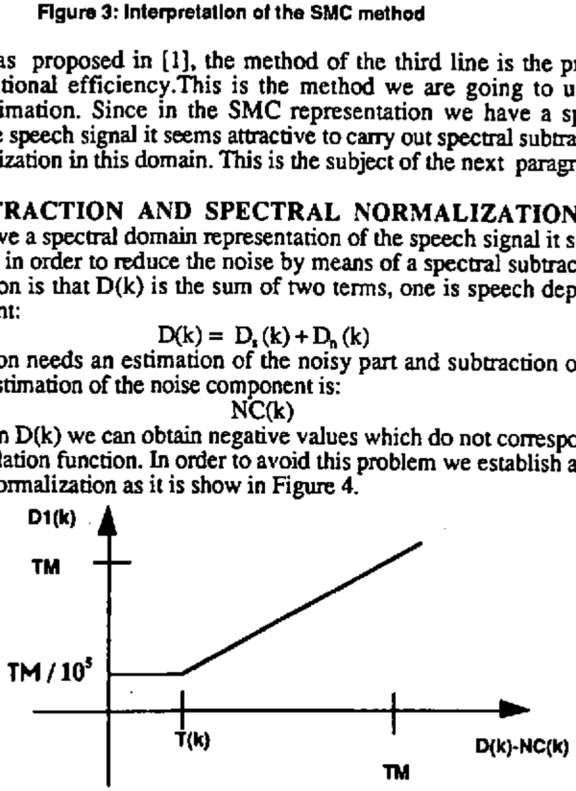

Figure 4: Spectral compression and noise subtraction

The spectral threshold T(k) has been estimated by averaging speech spectra and attenuating it by a fix quantity. It is maintained fixed. This non-linear transformation includes a dynamic range compression for the D(k) which contributes to a normalization of the representation for different environments. The dynamic range is estimated for a specific environment.The problem which remains is how to estimate the noise component.

In order to do that we have to classify the input signal in two classes: one corresponds to only noise and the other includes speech and noise. This classification is not very critical, since what is important is that we are able to detect when only noise is present. If we make a mistake and decide that a particular frame corresponds to speech plus noise although it is really a noise frame this error does not degrade significantly the procedure of estimating the characteristics of noise if it is not too frequent in which case the characteristics of the system can loose the ability of estimating the time-varying characteristics of the noise.

2.3. CHANNEL EQUALIZATION

The purpose of the channel equalization is to compensate for linear distortions introduced by microphones and room characteristics. We assume an average SMC spectrum for speech. Let AS(k) be this spectrum. We estimate AS(k) by averaging the SMC representation of clean speech and is mantained fixed for the recognizer. From a given speech recognition session we estimate the average SMC representation of the speech by averaging D(k) ,after noise subtraction, taken from different phrases, this average is session dependent and could be time varying if the speech channel is also varying.

We assume that there is a gain G(k) which satisfies:

$$AS(k)=G(k)\ AD(k)$$

The value of G(k) is used for equalizing the SMC spectral representation after the subtraction of noise:

$$D2(k)=D1(k)\ G(k)$$

We can estimate G(k) from a few seconds of speech taken in a given car and with the microphone located in a given position making a measure of S(k). We supose that this gain change significatively from one car to another and with the change of the position of microphone. However this hypothesis must be verified.

When low frequency noise are present it is important to carry out band-pass filtering. This is the kind of noise we observe in cars. We do this operation of windowing in the spectral domain by making G(k)=0 for low and high frequency. We select the low and high frecuency indices depending on the contents of interfering noise. Then the SMC representation permits an efficient way of bandpass filtering.

2.3. UPDATING CONTROL FOR SMC SPECTRAL SUBTRACTION AND EQUALIZATION

When noise is present, especially if its level is high, it is not possible to detect with low probability of error the speech and noise conditions. In fact, that is not necessary. What we need is to detect speech at a sufficient frame rate in order to follow the possible time-varying conditions of the channel. In an analogous way, it is necessary to update the noise characteristics by taking enough frames of noise.

The control of both conditions is carried out based on the frame energy of the signal. We define two thresholds. If the energy of the frame is under the lower threshold we decide that this frame is noise. If the energy of the frame is greater than the highest threshold we decide that speech is present.

2.4. LPC ANALYSIS AND FEATURE VECTOR CONSTRUCTION

Using the pseudo-autocorrelation:

$$\rho(k)$$

we calculate the LPC coefficients by means of the Levinson-Durbin recursion [4]. From these coefficients we estimate the cepstrum [4].Then we have a vector of cepstrum coefficients and the logarithm of the short-time energy every 10 msec.

As the next step we include a decimation by a factor of two stage. Every decimated cepstrum and log. energy is obtained by averaging the three obtained in consecutive windows. Additionally we calculate differential information of the cepstrum evolution. Following we estimate the regression coefficients. After the decimation stage we have the characteristic vector which include: p cepstrum coefficients, p regression coefficients and the regression of the logarithm of short-time energy.

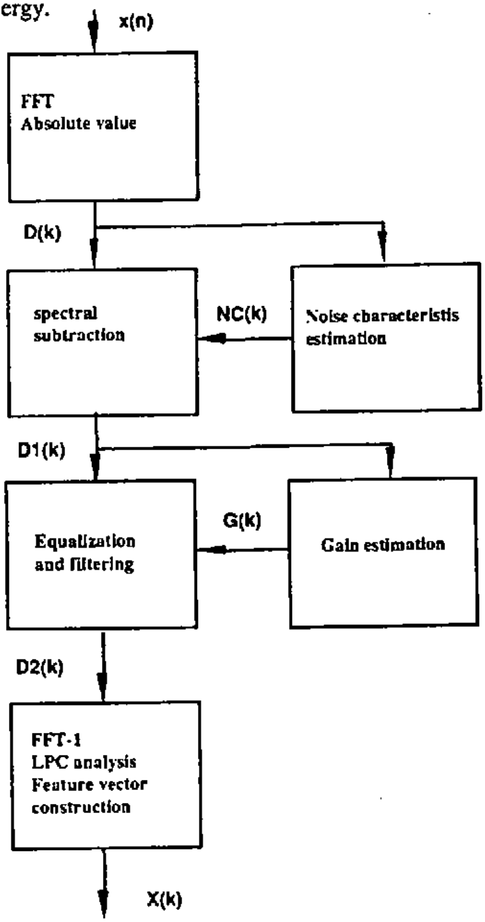

Figure 5: Feature extraction block diagram

3. ARCHITECTURE OF THE FEATURE EXTRACTION BLOCK

In Figure 5 we can see a general overview of the feature extraction block. It includes: SMC spectral representation,noise subtraction, channel equalization, LPC analysis, LPC to cepstrum estimation, decimation and regression coefficients estimation. The output of this block is the characteristic vector X(k).

4. EXPERIMENTAL RESULTS

For obtaining recognition scores we use an isolated word HMM based recognizer.It is left-to-right,without skips, continuous densities HMM.Each feature component is modelled as a Gaussian, different components are assumed to be uncorrelated. Then each state of the HMM is characterized by two vectors, one with the mean value of each component of the characteristic vectors and the other with their variances.

Training is carried out using the Grand-variance method [3]: after segmentation of the sequence of vectors obtained for given word of the vocabulary the means and variance vectors are estimated for each state. The first segmentation is uniform and the following ones are obtained by means of the Viterbi algorithm.This iterative process is carried out few times until convergence. We observe that four or five iterations are enough.

In order to evaluate the efectiveness of SMC spectral representation we carried out several recognition experiments and observed short-time spectra with different noise conditions.

The first experiment compares the percentage of errors obtained when SMC spectral representation are used for different sinal-to-noise ratios, SNR, with the errors obtained when standard LPC analysis is carried out.These results are shown in the first and second row of Table 1. They were obtained with a vocabulary of 35 entries. Each entry is a word or a phrase. Each entry has been modelled with a HMM of 10 states. HMM models were trained with SNR of 40 dB. Different noise conditions were simulated by adding synthetic white noise to clean speech.

SNR	40 db	20 dB	15 dB	10 dB	5 dB	0 dB
LPC	0.6	32	65	71	95	99
SMC	0.6	1.1	4	12	33	65
SMC+SPE SUBTRA	0.6	1.1	2.3	3.4	17	46

TABLE 1: Percentage of errors obtained with LPC, SMC and SMC with spectral subtraction.

We can observe that when we use SMC a lower percentage of errors were obtained for the different SNR used. In terms of SNR, the gain is about 15 dB.

The effect of spectral subtraction can be observed too. For larger SNR the percentage of error does not increase compared with the recognizer with only SMC. However for lower SNR we can observe an aditional gain of about 15 dB in SNR.

5. CONCLUSIONS

After studying different alternatives for feature extraction and noise reduction [10], we developped a feature extraction module whose main characteristics are:

- It is robust against additive noise. SMC representation is significantly more robust than the classical LPC analysis.

- We included in this block noise subtraction and equalization due that we have a spectral representation of signal. This reduced the computational cost compared with methods of spectral subtraction used outside of this block.

- We incorporated regression representation of features because these parameters tends to be more robust when noise is present since they are estimated by adjusting straight lines to a sequence of features.

- This block incorporated different ideas that have been used by other systems but, to our knowledge, its complete design is new.

6. REFERENCES

[1].- D. MANSOUR, B.H. HWANG JUANG. The short-Time Modified Coherence Representation and Noisy Speech Recognition. IEEE Trans. on Acoustic, Speech and Signal Processing. Vol. 17, No. 6, pag. 795-804. June 1989.

[2].- D.V. COMPERNOLLE. Noise Adaptation in a Hidden Markov Model Speech Recogniton System. Computer, Speech and Language, Vol 3, 1989, pag. 151-167.

[3].- E.A. MARTIN, R.P. LIPPMANN, D.B. PAUL: Dynamic Adaptation of Hidden Markov Models for Robust Isolated-Word Recognition. Procee. Int. Conf. on Acoustic, Speech and Signal Processing, pp 52-54. April 1988.

[4].- J.D. MARKEL, A.H. GRAY: Linear Prediction of Speech. Springer Verlag, Berlin, 1976.

[5].- S.E. LEVINSON, L.R. RABINER, M.M.SONDHI: An Introduction to the Application of the Theory of Probabilistic Functions on a Markov Process to Automatic Speech Recognition. The Bell System Technical Journal. April 1983.

[6].- L.R. RABINER, B.H. JUANG: An Introduction to Hidden Markov Models. IEEE ASSP Magazine. pp 4-16, January 1986.

[7].- L.R. RABINER: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. IEEE Proceeding, 1988.

[8].- K.F. LEE: Automatic Speech Recognition. Kluwer Academic Publishers. Boston 1989.

[9].- D. P. McGINN, D. H. JOHNSON: Reduction of all-pole parameter estimation bias by sucessive autocorrelation. ICASSP-83. Boston. pp1088-1091.

[10].- Deliverable 3200/1 of European SPRIT II project 2101, "Adverse environment Recognition of Speech".