# Robust Speech Recognition in GSM Codec Environments

Ph.D. Thesis Prospectus

**Juan M. Huerta**

Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

April 2, 1998

# TABLE OF CONTENTS

## 1.- Motivation

Speech recognition performance degrades rapidly under adverse conditions (*e.g.* [12], [25]). Approaches proposed previously that aim to alleviate the problem frequently assume a linear transmission channel, stationary noise, or similar assumptions. The main disadvantage of methods that follow a predefined model of the environment is that these methods degrade rapidly as the underlying assumptions fail to be met. Such is the case of CDCN [1] and VTS [22] when the channel fails to follow a strict linear model.

The evolution of digital communication systems and their application to speech communication (*e.g.*, cellular and mobile communications) is resulting in the introduction of new speech compression and coding methods. These methods transform speech data by means of numerical algorithms, purposely employing very limited arithmetic precision and aggressive quantization. As a result, a new virtual channel is obtained in which the recovered uncompressed signal is distorted in a nonlinear way. The overall effects of these speech coding techniques (*e.g.*, GSM. CELP, G.723, LPC-10) can be modelled as a lossy channel with substantial nonlinearities.

In this work we will focus on the GSM communication standard for speech. This is the European Community standard for cellular telephony (GSM 06.01) with a bit rate of 13 kbps. It has been shown in the literature [5, 17] that even though the GSM standard produces a signal of good intelligibility, GSM has a considerable impact on the recognition accuracy obtained by modern HMM speech recognizers. This problem is observed for all speech coding techniques: recognizers performance deteriorates as the bit rate lowers. The following table extracted from [5] illustrates this problem. The error rate more than doubles for 4.8-bit coding when models are trained on unprocessed speech.

| Technique | Bit rate | Word Accuracy |
|---|---|---|
| PCM | 64 kbps | 96.61% |
| LD-CELP | 16 kbps | 94.52% |
| GSM | 13 kbps | 94.08% |
| CELP | 4.8 kbps | 91.95% |

*Table 1: Word Accuracy as a function of bit rate (after Euler and Zinke).*

The objective of this thesis is to come up with a family of robustness techniques that will reduce the gap in performance observed between uncoded speech and speech processed by the GSM algorihm with and without background noise. Specifically, this work will focus on the three following areas:

- Robust front-end features
- Autoregressive compensation algorithms
- Compensation techniques based on GSM-specific models of degradation

## 2.- The GSM codec and its effect on speech recognition systems

The GSM standard is a lossy compression algorithm based on Regular Pulse Excited Long Term Prediction (RPE-LTP coding, *cf*. Kroon *et al.* [15]). For a more detailed description of the standard see [4]. GSM converts 13-bit uniform PCM digital signals sampled at 8 kHz into encoded blocks composed of 260 bits for every 160 original samples. Hence, the GSM coding algorithm produces a gross bit rate of 13.0 kbps (although the actual GSM transmission rate is higher due to error recovery and packet information added). Together with the Multi Pulse Excited Linear Prediction and Code Excited Linear Prediction algorithms, the RPE-LTP coding scheme belongs in the family of LPAS (*linear-prediction based analysis-by-synthesis*) algorithms.

### 2.1.- The analysis-by-synthesis principle

The principle of analysis by synthesis is illustrated in the figure below (after Kroon and Kleijn [14]). The idea behind analysis by synthesis is to find and quantize the set of parameters (*e.g.*, vectors, pulses) that if used to reconstruct the excitation signal would minimize the error between the original speech signal and the reconstructed version of it. In order to find these parameters, it is necessary for the decoder to incorporate a decoding block similar to that used in the decoded stage. Because the search for the best set of parameters requires the calculation of a large number of possible excitation signals this family of techniques receive the name of *analysis-by-synthesis* coding.

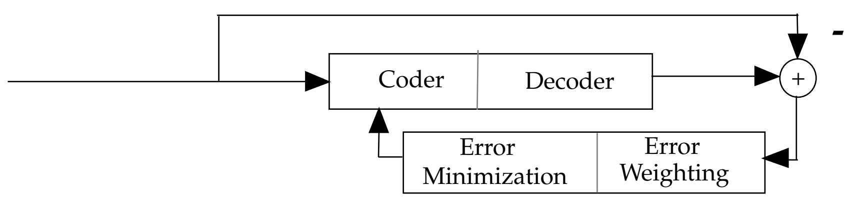


*Figure 1: A simplified analysis-by-synthesis coder (after Kroon and Kleijn)*

### 2.2.-Block level description of the GSM codec

A simplified block diagram of the RPE-LTP encoder is shown below. The 160-sample input speech frame is removed from its offset and preemphasized. The signal is then analyzed to determine the *short term analysis filter* (LPC analysis of order $p$=8). From this filter the *short term residual* signal (1) is obtained. The filter parameters are transformed to log-area ratios, quantized, and transmitted at 36 bits/20ms. The *short term residual* is subdivided into 4 sub-frames with 40 samples each.

Using each 40-sample long sub-block and the stored 120 samples of the *reconstructed short term residual* (4), the *long term prediction parameters* are obtained and transmitted. These parameters are the lag and gain of the long term prediction filter. They are obtained by cross-correlation (9 bits/5 ms). We refer to this section as the LTP analysis.

A block of 40 *long term residual samples* (2) is obtained by subtracting 40 samples of the *short term residual* *estimate* (3) from the *short term residual* signal (1) itself. The resulting block of *long term residual* samples is fed to the Regular Pulse Excitation analysis section which performs the basic compression function of the algorithm. The *long term residual* samples are represented by one subsequence of 13 pulses each which is encoded using Adaptive Pulse Code Modulation (APCM) with estimation of the sub-block amplitude; this is achieved by low-pass filtering the *long term residual* and dropping 2 out of 3 samples (47 bits/5ms). These LTP parameters are then used to reconstruct the *quantized long term residual* (5) that, when added to the *short term residual estimate* (3), produces the *reconstructed shot term residual* (4) that closes the feedback loop.The decoder replicates the feedback loop section of the coder. Specifically, the subblocks present in the decoder are the RPE grid decoding and the long-term synthesis filter, which together generate the *short term residual estimate*. These samples are then applied to the short term synthesis filter (LPC) followed by the deemphasis filter

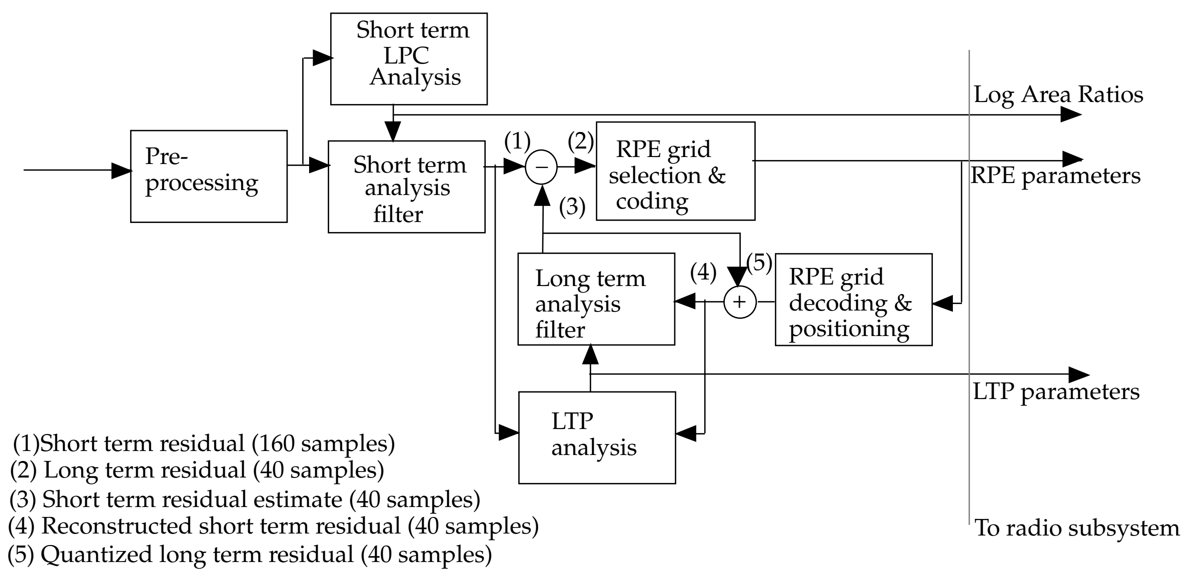


*Figure 2: Block diagram of GSM coder*

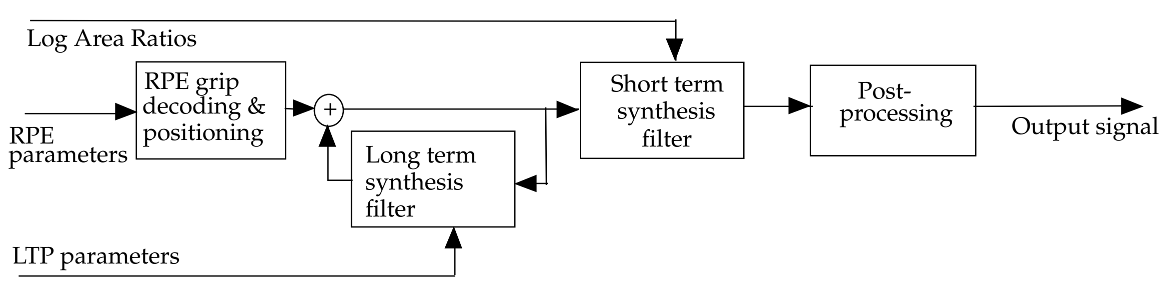


*Figure 3: Block diagram of GSM decoder*

**2.2.- Effects of the GSM codec: impact on recognition performance**

The combined effects of quantization, linear interpolation of LP parameters, sample rate reduction (elimination of data samples) in the GSM coding and decoding process, affect the speech signal in a way that is not describable as a linear time invariant system. Furthermore, as speech can be broadly classified as voiced and unvoiced, one can notice that the effect of the GSM compression over these two broad phonetic classes is different in nature. In the figure below, we illustrate the distortion introduced in the time signal for a segment of unvoiced speech. Even though signal distortion in unvoiced speech regions might be small from a perceptual point of view, it is large enough to produce an drop in the recognition acuracy.

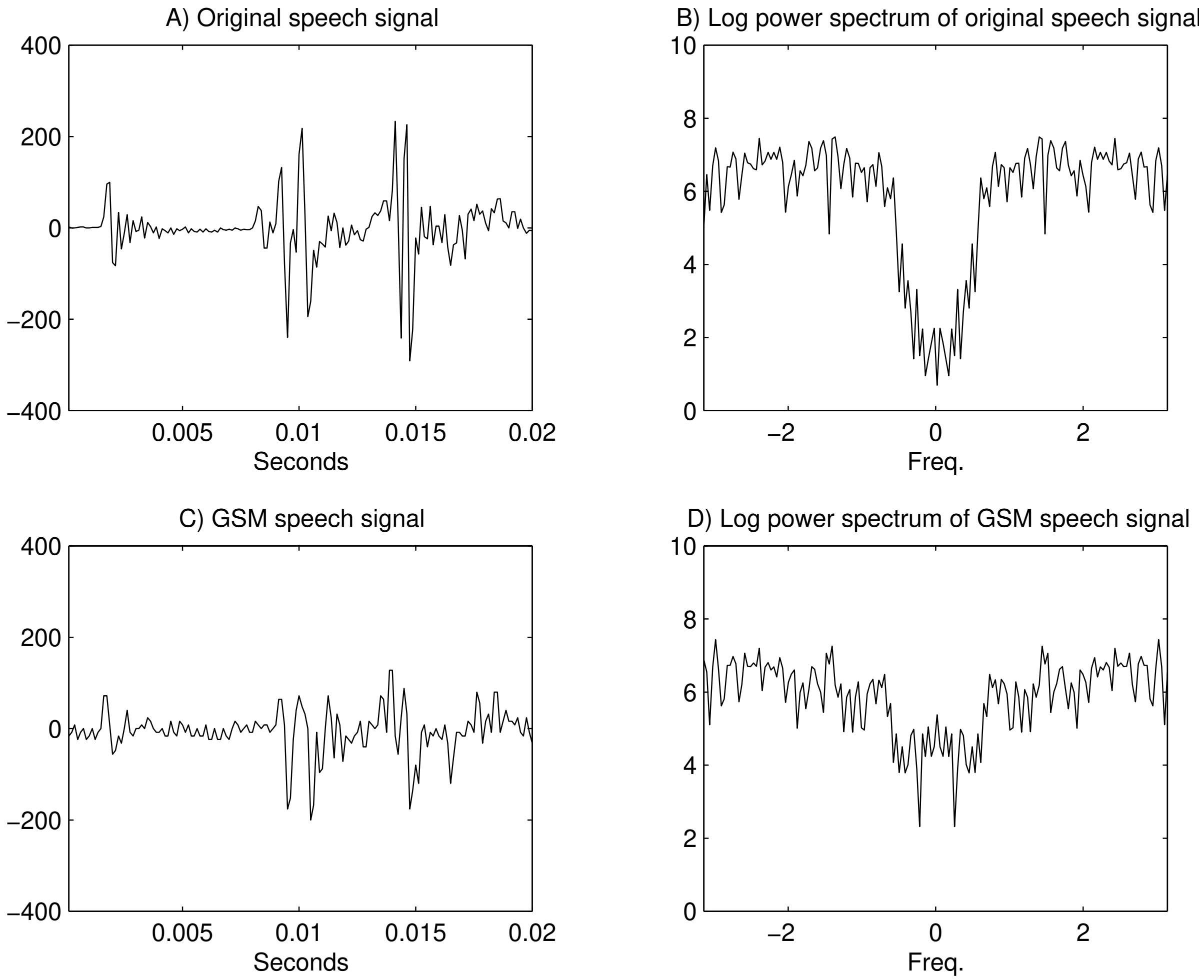


*Figure 4: Segment of unvoiced speech: original and GSM processed and their log power spectra.*

Table 2 illustrates this degradation in recognition accuracy. Experiments were conducted using the Resource Management Database (with reduced language model weight), the SPHINX-III recognizer using continuous density HMMs, and the standard Mel Frequency Cepstral Coefficients (MFCC) front-end (to be detailed in section 4). Results are reported for no additive noise, moderate noise (20 dB), and severe noise (10 dB) conditions. All the recognition experiments used HMM models trained on clean speech that had not undergone GSM distortion.

| | Baseline (Unprocessed) | GSM signal |
|---|---|---|
| **No noise** | 75.5% | 73.1% |
| **20 dB** | 47.2% | 40.3% |
| **10 dB** | 2.0% | -1.8% |

*Table 2: Degradation in recognition accuracy introduced by the GSM codec under different noise-coding conditions.*

From table 2 we can observe that the relative degradation in word error rate of the performance by recognizing GSM processed signals with unprocessed-speech models is 9.8% relative to the error rate obtained for unprocessed speech. Because we are testing on various noise and coding conditions using models that were trained using clean undistorted speech exclusively, we will refer to this kind of conditions as *mismatched conditions*. This thesis will focus on reducing the relative degradation in speech recognition performance introduced by the GSM codec under clean and additive noise acoustical environments in *mismatched* conditions.

Under moderate noise conditions (20 dB), the relative degradation in word error rate introduced by GSM is 13.07%. In this case, of ourse, recognition accuracy is affected not only by the degradation introduced by the noise, but also by the coding algorithm. Many of the speech coding techniques currently employed are based on the speech production model assumed a linear predictive analysis model ([20]), meaning that the speech signal is assumed to be the result of the convolution of a slowly changing filter (representing the effects of the vocal tract) by a spectrally flat excitation signal (representing the glottal waveform). When additive noise is present, the coding algorithm operates on the assumptions that are not valid anymore introducing distortion and artifacts into the reconstructed signal. This results in even further reduction of the recognizer's performance.

## 3.- Relevant compensation techniques: overview and limitations

In this section we present a brief overview of methods previously employed to approach the problem of compensating channel and noise distortion. We will also present an overview of previous attempts to compensate against speech coding degradation in speech recognizers, and based on the inherent limitations of the current techniques presented, introduce some ideas regarding overcoming such limitations.

### 3.1 Overview of existing compensation techniques

Previously proposed methods have attempted to reduce the degradation of performance in speech recognition systems introduced by additive noise and convolutional filtering. Many of these techniques assume that the filtering effect introduced by the channel is linear. The CDCN algorithm (Acero [1]) is a technique that considers jointly the channel and the noise signal, more recent techniques (VTS, Moreno [21]) make use of the same environmental model.

Other family of methods attempt to reduce the degradation by modifying the HMM models. *Maximum Likelihood Linear Regression* MLLR (Leggetter and Woodland [16]), is a technique that estimates a regression matrix to modify the means of the HMM models in the recognizer. Even though this technique does not assume an explicit linear degradation of the signal, it is limited to a set of corrections to the means of an set of HMM models in a way specified by a single regression matrix (for the case of single-class MLLR) or a set of regression matrices (multiple-class MLLR). This means that even though MLLR does not assume a specific model of degradation it can only compensate the model parameters in a linear way determined by the regression matrix.

*Cepstral Mean Normalization* (CMN) is a simple technique that aims at making the means of the training data equal to the means of the testing data. By subtracting the corresponding cepstral mean from an utterance, one is in practice removing the effect of linear time invariant convolutive filtering effectively performing a *blind deconvolution*. Mokbel *et al* [18] have use this method explicitly to remove the linear bias in PSN & GSM experiments. Currently this method is used as a standard feature of most recognizers because its simplicity and efficiency.

*RASTA* processing [8] is a technique that is based on the idea that the linguistic message of a speech signal is manifested in relatively slow movements of the vocal tract. By filtering out the components of the signal that change significantly faster or slower than the vocal tract rate of change, one can effectively remove the non-linguistic elements of the signal.

*Spectral Subtraction* has been used in the areas of speech recognition [3] and speech enhancement [2]. It consists of estimating the power spectrum of the noise and then subtracting it from the power spectrum of the noisy signal. To prevent the resulting power spectrum from having negative values, a rule such as following rule is applied:

$$X(\omega) = max\{0, Y(\omega) - N(\omega)\}$$

Where $Y(\omega)$ is the spectrum of the noisy speech, $N(\omega)$ is the noise estimate and $X(\omega)$ is the estimate of the spectrum of the clean speech. The above rule is in effect a nonlinear compensation of the clean spectrum and relies principally in the quality of the estimate of the noise.

### 3.3.- Previous work on speech recognition using digital speech coding techniques

In [18] Mokbel *et al.* described a series of experiments directed towards obtaining robustness on speech applications over the PSN (Public Switched Network) and GSM networks. The techniques applied consist of a collection of previously used general robustness techniques: cepstral normalization, blind equalization on adaptive filtering, RASTA filtering [8], spectral subtraction and Bayesian HMM adaptation. The experiments were conducted using real field data recorded by a wide variety of speakers. The results obtained show that a degradation in recognition accuracy is observed when GSM test data is recognized in both matched and mismatched model conditions compared to the recognition accuracy obtained for uncoded speech over the PSN network. In addition, most of the compensation techniques listed above yielded a gain in performance in both GSM and PSN channels, but these techniques are also correcting for differences in speaker styles and background noises. Although some improvement was obtained by applying general compensation techniques, no work on compensation techniques that explicitly consider the nature of the GSM degradation was performed.

The work of Haeb-Umbach [6], focussed on the robustness of speech applications that were actually deployed. In real-world applications the speech coding problem is just one small component of the set of sources of degradation. Furthermore, other implementation issues have to be taken into account such as speed and hardware requirements. Haeb-Umbach stressed the idea that under severe degradation, emphasis should be placed on dynamic features (*e.g.*, delta features) and techniques that emphasize the dynamics of speech (*e.g.,* RASTA processing). He also enumerated the problems that real applications encounter: additive car noise (for automobile applications), Lombard effect, reverberation. He did not propose a technique that addressed the problem of distortion introduced by speech coding.

We have been unable to locate oter significant approaches toward improved robustness for recognition under CELP or any other analysis-by-synthesis codec .

### 3.4.- Performance of CDCN under the GSM codec

Table 3 summarizes the performance of CMN and CMN+CDCN under GSM coding in several noise conditions. The Cepstral Mean Normalization technique is applied usually as front-end default processing due to its simplicity and its consistent performance enhancing properties. In all our experiments we use CMN as a default even though we know that its main purpose is to reduce *linear channel* filtering effects.

After applying CDCN a reduction in the error rate is observed under every noise condition. In spite of the substantial nonlinearities introduced to the speech signal by the GSM codec and in spite of the fact that CDCN does not take into consideration the particular model of degradation that the GSM introduces to the signal, the CDCN algorithm is able to achieve significant performance improvement. We believe that by considering the GSM particular structure and model of degradation we can get further reduction of the error rate.

| SNR | CMN | GSM+CMN | GSM+CMN + CDCN |
|---|---|---|---|
| ∞ | 75.5% | 73.1% | 73.6% |
| **20 dB** | 47.2% | 40.3% | 51.6% |
| **10 dB** | 2.0% | -1.8% | 2.5% |

*Table 3: Effect of compensation techniques under different channel-coding situations*

### 3.5.- Overcoming the limitations of current techniques

The work that has been done on achieving robustness under speech coding distortion has only focused on applying general compensation techniques to the problem. No work has focused on reducing the sensitivity of the speech feature to coding artifacts, or to model the degradation introduced by the coding. The results obtained by Mokbel and Haeb-Umbach partially reduce this degradation. The gap in performance between GSM and non-GSM applications seems to persist under the current approaches to the problem.

One way to reduce the sensitivity of the recognizers to coding artifacts is by employing speech features that are robust to these artifacts. In the next chapter we will propose the use of techniques that make use of autoregressive spectral estimates to calculate the cepstra. This not only presents us with the advantage of allowing us to work with an LPC-smoothed version of the spectrum but also it presents the advantage that it uses information already contained in the GSM packet: the log-area ratio coefficients that can easily be converted to an LPC vector. We will refer to this kind of cepstral representation as *autoregressive cepstra*.

Another way to overcome the limitations and problems of techniques that modify the power spectra based on a noise estimate (Spectral Subtraction) is by modifying the LPC vector: this has the advantage that can be done in a structured way and generally will not yield LPC parameters that are unstable or close to unstable. We will propose a special way of Spectral Subtraction that operates on the autoregressive domain and for the particular case of GSM coding, it can be applied at the LAR level. In chapter 5 we will present our ideas in this direction.

Another limitation of the techniques and methods mentioned above is that some of them follow linear models of the degradation (CDCN, CMN), or they assume a linear displacement of the means of the acoustic models (MLLR); we will try to determine and model the effect of quantization, downsampling and coding of the speech parameters in the GSM codec so that we can manipulate and possibly compensate this parameters. This will permit characterize and compensate these coding effects before we convolve the signals back and thus regenerate the speech signal. Chapter 6 will detail our expectations in this area.

In the next chapters we will describe the proposed approaches to overcome the GSM induced degradation by means of a robust front-end (chapter 4), autoregressive compensation (chapter 5), and compensation based on the GSM model of degradation (chapter 6).

## 4.- Proposed work: robust speech features

One way to improve recognizer's performance under coding distortion is to employ features that are sensitive to spectral variations of speech but robust to channel and coding variations. Even though most modern recognizers use cepstral features, there are different ways to obtain the cepstrum resulting in important differences in performance. The two most common methods of feature extraction in mainstream front ends are Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) (Hermansky [7]). The CMU SPHINX-III system has traditionally used MFCC features. In this section we present some preliminary work that indicate the feasibility of utilizing robust features that use ideas from the family of autoregressive front ends (PLP) to increase the system performance under GSM coding and present our ideas for future work in this area.

### 4.1.-Cepstral features for speech recognition

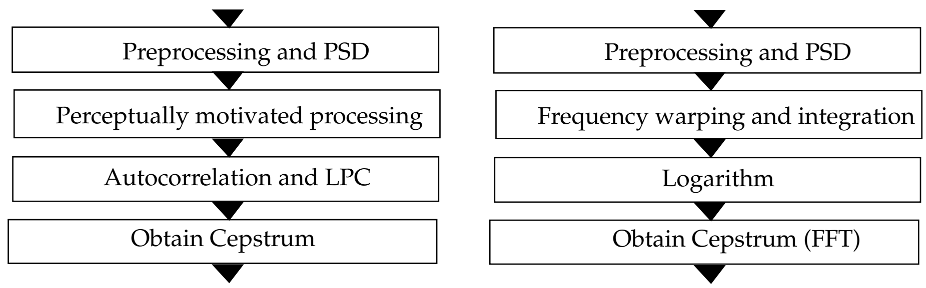


*Figure 5: Block level representation of the PLP (left) and the MFCC (right) front ends*

The figures above illustrate the basic building blocks that comprise the two most important and widely used types of cepstral features: PLP and MFCC. The MFCC front end computes the PSD (Power Spectral Density) and then reduces it to a warped and integrated version; the PLP front end obtains the PSD and then performs some perceptually motivated processing on it. After this, both front ends obtain the cepstrum from the processed version of the PSD: PLP uses an autoregressive method while MFCC computes the logarithm of the PSD and then performs an Inverse Fourier Transform. Front ends that utilize autoregressive spectral estimates to compute the cepstra will be referred to as *autoregressive cepstral* front ends.

### 4.2.- Proposed work: minimizing front-end sensitivity to speech coding artifacts

In section 4.1 we highlighted the differences across cepstral features. In this section we will highlight the differences in performance and point to the importance of robust features. Figure 6 shows a detailed diagram of a front end that has provided better recognition accuracy. This system is based on the PLP and the MFCC front ends but with parameters that have been adjusted for performance in our evaluation conditions. The proposed autoregressive (AR) front end computes a power spectral density for each frame of 20 ms. of speech and performs spectral warping and spectral integration. It then obtains the autocorrelation from the warped and integrated spectra and from the autocorrelatio, and finally obtains the LPC coefficients and the cepstra. This autoregressive front end achieves a 5.8% relative error reduction (Table 4), which outperforms the gain provided by the standard MFCC front end with CDCN compensation. We believe that the result obtained so far, even though modest, are a good indication that further progress can be achieved in reducing the error rate under GSM by customizing the characteristics of the front end to the GSM degradation.

| | **MFCC** | **PLP 13-order** | **AR Cepstrum** |
|---|---|---|---|
| **Clean** | 75.5% | 75.1% | 77.1% |
| **Clean GSM** | 73.1% | 72.1% | 75.1% |
| **20 dB** | 47.2% | 49.9% | 56.5% |
| **20 dB GSM** | 40.3% | 42.4% | 51.2% |
| **10 dB** | 2.0% | 10.0% | 17.0% |
| **10 dB GSM** | -1.8% | 3.4% | 11.5% |

*Table 4: Performance of various cepstral font-ends under different noise-coding conditions*

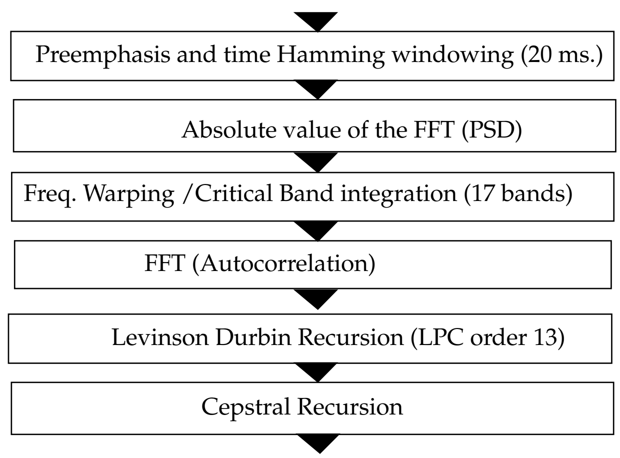


*Figure 6: Autoregressive cepstrum front end*

Ways to increase the robustness of a front end and minimize the sensitivity to coding artifacts include methods that consider different levels of time and frequency resolution at different regions of the spectrum. To achieve this front end robustness we propose to apply the following modifications to our autoregressive front end:

- **Spectral resolution:** Figure 7 shows the frequency response of the GSM codec ([4]). It is evident that the frequency response of the codec is not completely flat, particularlly ar the high part of the frequency spectrum. In order to reduce the distortion introduced by GSM in high frequency range we will develop new front end configurations which emphasize low frequency resolution in the upper part of the spectrum and high frequency resolution in the lower part of the spectrum using a relatively small number of spectral subbands. The purpose of this operation is not to motivated by a perceptual reason (Bark or Mel scales) as much as by an attempt to minimize the effects of the distortion of the GSM codec. Our starting point in this area is to adopt the spectral resolution style of the PLP front end which uses a low number of analysis sub-bands and refine this setting according to the GSM frequency response chart.

- **Spectral integration functions:** We will utilize spectral integration functions with increasing widths as a function of critical band number. We will also experiment with triangles for the lower part of the spectrum and trapezoids for the upper part of the spectra and increase their areas of overlap in the upper part of the spectrum. This should increase the averaging (smoothing) that is performed at the upper part of the spectrum and should help us minimize the non-flat frequency response of the GSM codec in that spectral region.
- **Time resolution/time windowing:** Using longer windows in time yields better frequency resolution. The current use of a single time window (20 ms hamming window) introduces a convolution operation in the spectrum with the frequency response of the window smoothing uniformly across the spectrum. We will control this smoothing operation and the frequency resolution by increasing the length of the time window (currently 20 ms long, every 10 ms) equally for all the frequency subband and subsequently utilize a time window spanning different lengths for different subbands. In this way we will have more control of the effective time/frequency resolution in each subband.

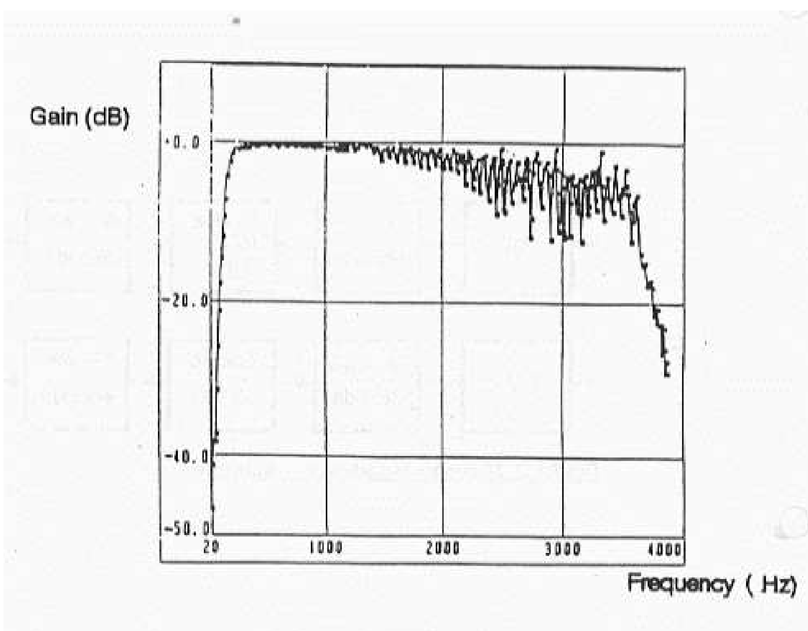


*Figure 7: GSM codec frequency response* [OBTAINED HOW?]

## 5.- Proposed work: autoregressive compensation algorithms

In [1] Acero describes a technique that jointly compensates the effect of additive noise and channel distortion (CDCN). As we showed in table 3 this method gives some improvement in the word error rate performance of the recognizer under GSM with noise. This technique considers the effect of noise on cepstra. In this section we propose a method to directly address the problem of additive noise. The direction that we propose is to act directly on the autoregressive parameters (*i.e.*, poles, LPC coefficients, autocorrelation coefficients) in such a way that we will reduce the effect of the additive noise.

#### 5.1.- GSM autoregressive parameters: Log Area Ratio, reflection coefficients and autocorrelation

The GSM codec transmits the LPC vector as quantized log area ratio coefficients because these parameters are robust for both interpolation and quantization. The reflection coefficients are derived from the LAR coefficients. From them, the LPC vector can as well be derived and the poles of the polynomial obtained. Finally, from the set of LPC coefficients we can obtain the first few lags of the autocorrelation function by the Levinson-Durbin recursion ([20]). In brief, we can convert the set of LAR (directly observed from the GSM packet) into any other form of autoregressive information.

#### 5.2.- Preliminary work: compensation of pole locations

In [10] we proposed a method to reduce the effect of the environment by modifying the pole locations based on the statistics of the poles of clean and noisy data. Even though this method achieve some performance improvement in the case of moderate to low SNRs, it does not provide very much benefit for moderate to high SNRs. The spectral distortion introduced by this method is high with respect to the potential benefit due to normalization at high SNRs. Figure 8 summarizes the performance of this method at various SNRs for the compensation of angle, radius, and angle and radius.:

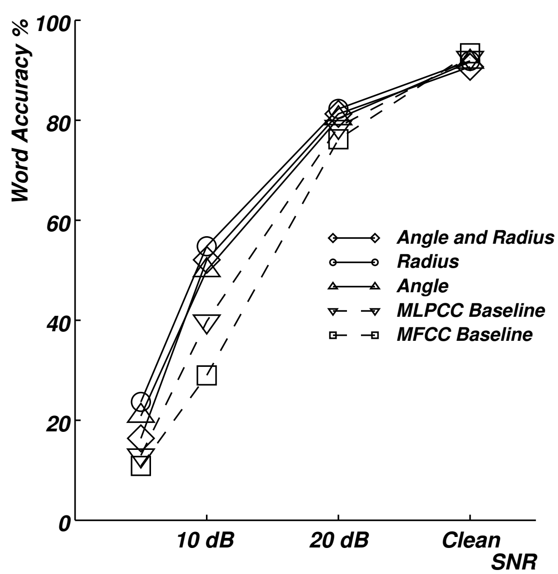


*Figure 8: Performance of the pole compensation method*

The pole compensation technique allowed us to experiment with compensation of autoregressive parameters, specifically with the poles. We also noticed that our algorithm does not help at that SNR range. Several ideas have been brought up to increase the performance of this method, but we believe that the most significant change we can adopt is to base our compensation technique on a model of the perturbation introduced by noise. We discuss on way to achieve this in the next subsection.

### 5.3.- Proposed Work: Extension of ARDOSS (Autoregressive Domain Spectral Subtraction)

In this section we present a method proposed by Van Hamme [27] for compensating the autoregressive LPC vector in the presence of additive noise and we extend it to the case of autoregressive cepstra. This method has the advantage of being based on a first-order noise perturbation analysis of the autocorrelation function, which can be a more robust and effective method than the pole compensation algorithm.

To understand the principles of the algorithm, let $s_t$, $n_t$ and $x_t$ denote samples of the noise-free speech, the stationary noise and the noisy speech respectively

$$x_t = s_t + n_t$$

Let us consider the Yule-Walker equation. It relates the correlation vector $r_s$ to the LPC vector $a_s$ through the matrix of correlations $R_s$

$$\underline{a}_s = -\underline{R}_s^{-1}\underline{r}_s$$

In the presence of noise, we have no direct access to $r_s$, but if the speech signal is independent from the noise signal we could estimate the autocorrelation of the noise and subtract it from the autocorrelation of the observed signal. This subtraction can result in losing the positive definiteness of the Toeplitz matrix $R_s$ and also losing the stability of the polynomial $a_s$ [26][27]. For this reason we will adopt the strategy proposed by Van Hamme in [27] which modifies the prediction polynomial $a_s$ based on a linearized perturbation analysis.

If the correlation vector $r_s$ is perturbed by $\Delta \mathrm{r}_s$ , the effect of such perturbation on the vector $a_s$ is approximated by:

$$\Delta\underline{a}_s = \sum_i \frac{\partial \underline{a}_x}{\partial r_{s,i}} \Delta r_{s,i}$$

After assuming independence of the speech and the noise signals, a perturbation of the autocorrelation can be seen as being produced by the additive noise signal. After some algebraic manipulation ([27]) the vector $\hat{a}_s$ corresponding to a compensated LPC vector (estimation of the clean signal LPC vector) can be expressed as:

$$\hat{\underline{a}}_s = \underline{a}_x + \underline{R}^{-1}{}_s \underline{B}_s E\{\bar{\underline{r}}_n\}$$

where the matrix $B_s$ depends on the coefficients of $a_s$. For high SNRs $R_s$ and $B_s$ can be approximated from $R_x$ and $B_x$([27]), thus:

$$\hat{\underline{a}}_s = \underline{a}_x + \underline{R}^{-1}{}_x \underline{B}_x E\{\bar{\underline{r}}_n\}$$

The expression above says that the compensated estimate of the autoregressive vector of the speech signal can be obtained by adding the calculated vector of the observed noisy signal plus a term that is a function of the $R_x$ and $B_x$ matrices of the observed signal times the estimate of the autocorrelation of the noise signal. The equation above makes it necessary to obtain an estimate of the autocorrelation of the noise. This is generally done by obtaining the autocorrelation of the signal during portions of speech inactivity.

In our extension of the work of Van Hamme to the GSM codec, we will work in the following directions:

- **Detection and handling of regions of *comfort noise*** (introduced in regions without voice activity): The GSM standard for cellular telephony besides of specifying the codec algorithm, also includes provisions for packet handling, error recovery and regions of speech inactivity. The insertion of comfort noise is performed in order to avoid sending GSM packets when there is no voice activity saving bandwidth while making more natural for the receiving end the transitions between voice active segments and voice inactive segments. Estimating the autocorrelation of the noise in this case poses a severe problem. We propose to overcome the problem by using estimates from areas where we know there is voice activity but the SNR is low leaving us with the problem of determining such areas.
- **Use of the *perceptual-autocorrelation*:** The compensation method proposed by Van Hamme is based on the autocorrelation of the speech function. During the computation of the autoregressive cepstrum the LPC vector is obtained from an autocorrelation vector which is derived from the frequency integrated and perceptually warped power spectrum of the signal. This spectral integration of the power spectrum, while theoretically should not affect the assumption of independence of the noise and speech, might result in a less effective algorithm (*e.g.*, the resolution of the spectrum is blurred by the spectral integration process).
- ***Iterative formulation*:** An interesting idea to increase the effectiveness of the algorithm at low SNRs is to apply the algorithm iteratively. First start with the assumption that $a_s$ is very similar to $a_x$ (high SNR assumption) then compute the value of $\hat{a}_s$ and use that value to recalculate the matrix $B_s$ (and possibly $R_s$, by inverting the Levinson-Durbin recursion). Use the new matrices $B_s$ and $R_s$ to recompute the new value of $a_s$, until some convergence criterion is reached. We will study the conditions in which the above iterative procedure might be helpful, as it might be the case that convergence is achieved only when the perturbation of the correlation is small.

## 6- Proposed work: techniques based on the GSM codec

The proposed work in Sections 4 and 5 describes our work towards increasing the performance of recognizer without explicitly considering the nature of the degradation introduced by the GSM coder. Because the GSM codec decomposes the speech signal into a quantized representation of the LPC vector and the residual, we can exploit the GSM coding scheme in two ways. The first form of explicit use of the GSM model is to recognize from features derived directly from the GSM packet. The second form is to model the degradation of each section of the coding algorithm and devise model-based compensation techniques.

### 6.1.- Recognition using features derived directly from GSM parameters

The LPC section of the GSM coding algorithm is at one of the earliest stages. This implies that the LPC analysis is acting directly on the observed speech signal (plus possibly noise). No aliasing, quantization noise, or any other distortion or coding artifact has been introduced to this signal yet. Therefore the transmitted LAR coefficients contain information of the spectral envelope of the undistorted signal. The LAR coefficients represent an LPC filter of relatively low order and with considerable quantization noise. It is possible to derive a cepstrum from this LAR filter information. This cepstrum, however, will have low-order LPC information and will not be based on a perceptually-warped frequency scale. Another problem is that of frame rate: GSM transmit LPC information with a rate of 50 frames per second, while our current recognizers work with twice that frame rate. Finally, we should also expect additional degradation arising from the fact that the transmitted LAR coefficients are quantized versions of the original LAR coefficients: this quantization noise can be modeled as additive zero-mean uniformly distributed noise added to each of the LAR coefficients.

For all these reasons, it is reasonable to expect such cepstra to provide relatively low recognition accuracy when compared to the traditional 100 frames per second, 13 dimensional Mel-frequency warped cepstrum. Table 5 depicts the performance of cepstrum derived exclusively from the LAR parameters. It also shows that as we include additional information obtained in the residual signal, the recognition accuracy increases.

In our work, we will continue working on the derivation of features that provide high-accuracy recognition. We will accomplish this by extending the information contained in the LAR parameters with a judicious selection of information contained in the rest of the GSM parameters (residual information), using frame interpolation, smoothing the data, and applying warping of the frequency axis. We will consider the following issues:

- **Subframe generation by parameter interpolation:** Interpolating frames at the LAR, reflection coefficient, cepstral, and LPC levels results in different SNRs for each set of parameters ([14]). From this SNR perspective some parameters produce less distortion that others, but these SNR metrics might not translate into proportional effects on speech. We will seek the best way to interpolate frames maintaining high recognition accuracy as our principal criteria.
- **Low-pass filtering of the parameter trajectories:** In [9] Hermansky *et al* proposed a method to

increase the quality of speech through cellular channels. Their method applied filtering to the cubic root compressed trajectories of the power spectra. We will determine the extent to which this kind of method presents an advantage to our problem, particularly for the parameters that suffer the greatest quantization.

- **Generate and recombine of LAR cepstra and residual cepstra:** The cepstra obtained from an autoregressive vector are redundant after the coefficient *p* (where *p* is the order of the autocorrelation), thus rendering the higher-order cepstral information obtained from autoregressive parameters less useful. The residual is expected to be spectrally flat ([20]) thus rendering the low cepstral coefficients obtained from residual information less useful. With these 2 ideas in mind we expect to find methods like cepstral concatenation more useful for recognition that cepstral addition. In Table 5, fourth row, we can see that the feature derived from concatenated cepstral information from the residual and from the LAR cepstra outperforms a 13 order LPC cepstra vector.
- **Frequency warping implementation at the LPC level:** It is possible to warp the frequency axis at the LPC level as long as the warping function can be expressed as a function of the complex variable $z$ (*e.g.*, bilinear warping). This has been performed previously at the spectral and cepstral level [1]. Because of we have access to the GSM LPC vector we are interested in identifying the family of frequency warpings that can be applied at the LPC level or to identify if some approximation can be done for the Mel or Bark frequency scales.
- **Frame Power (c[0]) Estimation:** This information is not directly available from the GSM packet data. This parameter should be obtained without having to resynthesize the whole speech signal.

| | **Clean** | **20 dB** | **10 dB** |
|---|---|---|---|
| **8th order LPC cepstra** | 73.1% | 60.9% | 31.1% |
| **13th order LPC cepstra** | 74.7% | 59.5% | 34.0% |
| **GSM LAR Cepstra** | 72.8 | 63.6 | 41.0 |
| **GSM LAR Cepstra plus GSM Residual Cepstra (concatenated)** | 75.7 | 68.2% | 47.0% |

*Table 5: Performance of different residual-LPC cepstral features*

### 6.2.-Towards a model of degradation

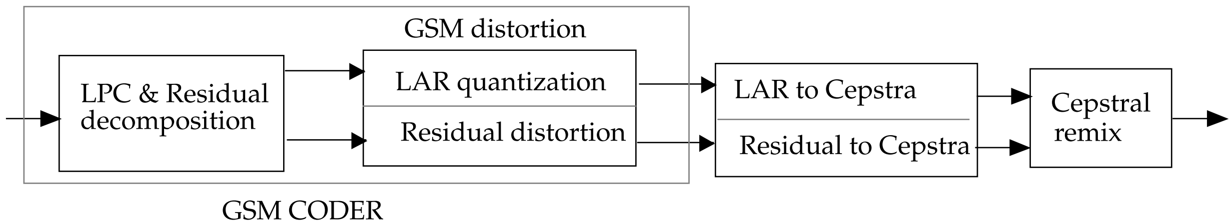


*Figure 9: Model of the GSM degradation*

The decomposition of the signal into an LPC vector and a residual, the compression and quantization of these two signals separately, and the corresponding generation of cepstra from those 2 processed signals leads to a *degradation model*. This proposed degradation model is depicted in the figure above. The GSM coding consists of the generation of the LPC and residual information and subsequent degradation introduced to each of these two signals by GSM quantization and coding.

The major sources of degradation produced in isolation by the GSM block are:

- **Quantization of the log area ratios:** Quantization can be seen as an introduction of additive noise to a signal. Adding noise to the log area ratios, translates as a distortion (not completely additive, though) of the LPC vector. However, LAR coefficients are more resilient to quantization than other representations of the autoregressive information ([14]).
- **Aliasing in the RPE section:** The LPF and downsampling process in the RPE section might introduce some aliasing in the spectrum of the restored residual signal.
- **Quantization of the RPE parameters:** Again, quantizing a signal is equivalent to adding noise to that signal.

We propose to compensate the GSM parameters directly and to compensate the cepstra derived from these parameters. The nature of the compensation will depend on the degradation: for example, for the particular case of parameter quantization, we know that quantization can be modeled as additive noise ([14]). A general technique to reduce such additive noise is Wiener filtering. Thus, we can introduce Wiener filtering of the parameters as a way to minimize the effect of quantization before generating the cepstra.

Finally, we should point out the fact that the GSM may produce different kinds of distortions for the different kinds of speech classes. Speech can be broadly classified as voiced or unvoiced depending on the nature of the of the glottal excitation signal. The RPE section of the coder attempts to exploit the periodicity of the excitation signal. So when the excitation signal is not periodic at all (as in the case of the unvoiced segment of speech in figure 4) one should expect the GSM codec to produce a poor result. We will try to keep this in mind when applying the compensation algorithms proposed so far. Previous compensation techniques have made use of similar subclassifications of the speech signal (*e.g.*, PCDCN, Multiple Class MLLR).

## 7.- Thesis Plan, Goals and Timeline

### 7.1.- Resources: database, system and GSM software

The database in which we will be initially focusing our experiments is the Resource Management Database, which has a 1000 word vocabulary. We expect to have access to a small-vocabulary actual GSM database provided by “Telefónica de Investigación y Desarrollo” in the future. When working with the RM database we will use a suboptimal low language-model weight in order to minimize the effect of constrained grammars. The recognition system is SPHINX-III (continuous density). The GSM software is the freeware written by Jutta Degener and Carsten Bormam at the “Technische Universitaet Berlin”.

### 7.2.- Expected results and contributions of the Thesis

- Evaluation of speech coding on speech recognition performance: Effect of bit rate on recognition accuracy; impact of coding rate on spectral and cepstral distortion. Evaluate the joint effect of additive noise and GSM codec. Evaluate noise conditions and simulate codec induced degradation.
- Analysis of GSM coding on log-spectra and cepstra: LPC section, LTP section, RPE section and the impact on various phoneme classes and noise conditions. Effect of quantization on parameters.
- Robustness against coding algorithms based on robust features: Autoregressive Cepstrum, time-frequency resolution. Power spectral density smoothing techniques.
- Robustness against noise conditions: Autoregressive domain spectral subtraction, extension to frequency warped power spectra, extension to GSM parameters, iterative algorithm.
- Generation of Autoregressive Cepstrum from Autoregressive GSM parameters: interpolation, smoothing frequency warping and compensation of GSM parameters.
- Incorporating the residual information: cepstra generation and compensation from the residual, characterizing and minimizing the coding artifacts in the residual. Techniques based on the model of degradation: block-based analysis of the residual, cepstra of the residual and cepstra of the LPC.

### 7.3.- Preliminary timetable

| Task | Start date | End date | Duration |
|---|---|---|---|
| **Evaluation of speech coding on recognition, evaluate noise conditions, simulate codec induced degradation** | Feb 98 | April 98 | 2 months |
| **Analysis of GSM coding on log-spectra and cepstra, class based distortion, effect of quantization** | April 98 | June 98 | 2 months |
| **Autoregressive Cepstra from GSM: interpolation, frame smoothing, warping** | June 98 | Oct 98 | 4 months |
| **Robustness against noise conditions (Extended ARDOSS)** | Oct 98 | Feb 99 | 4 months |
| **GSM parameter compensation: techniques for residual signal** | Feb 99 | May 99 | 3 months |
| **Overall** | Feb 98 | Aug 99 | 18 months |

| Task | Start date | End date | Duration |
|---|---|---|---|
| **Write report** | June 99 | Aug 99 | 3 months |
| **Overall** | Feb 98 | Aug 99 | 18 months |

## 8.- References


**1.-** A. Acero "Acoustical and Environmental Robustness in Automatic Speech Recognition" Ph.D. Thesis 1990 CMU.

**2.-** M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise". In J.S. Lim (editor), Signal Processing. Volume 1: Speech Enhancement, pages 69-73 1983.

**3.-** S.F. Boll "Suppression of Acoustic Noise in Speech Using Spectral Subtraction" IEEE Trans. Acoustics, Speech and Signal Processing No. 2 pp. 113-120 April 1979

**4.-** European Telecommunication Standards Institute "European digital cellular telecommunications system (Phase 2): Full rate speech processing functions (GSM 06.01)", ETSI 1994.

**5.-** S. Euler and J. Zinke, "The Influence of Speech Coding Algorithms on Automatic Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing. Vol ASSP- No. 1994

**6.-** R. Haeb-Umback, "Robust Speech Recognition for Wireless Networks and Mobile Telephony" EUROSPEECH 97

**7.-** H. Hermansky "Perceptual Linear Predictive (PLP) Analysis of Speech", J.Acoust. Soc. Am. Vol.87 No. 4 April 1990

**8.-** H. Hermansky and N. Morgan "RASTA processing of Speech", IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4 pp. 578-589, 1994.

**9.-** H. Hermansky, Eric A. Wan, and Carlos Avendaño "Noise Suppression in Cellular Communications" 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA 1994).

**10.-** J.M. Huerta, R.M. Stern "Compensation for Environmental and Speaker Variability by Normalization of Pole Locations" Eurospeech 1997

**11.-** M.Y. Hwang "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition", Ph.D. Thesis, School of Computer Science, Carnegie Mellon University. Pittsburgh, PA 1993.

**12.-** B. Juang, "Speech Recognition in Adverse Environments", Computer Speech and Language, Vol. 5 pp. 275-294, 1991.

**13.-** S. M. Kay "Noise Compensation for Autoregressive Spectral Estimates", IEEE Transactions on Acoustics, Speech and Signal Processing Vol. ASSP-28, No. 3, June 1980

**14.-** W.B. Kleijn and K.K.Paliwal (editors), Speech Coding and Synthesis, Elsevier Science B.V., Amsterdam 1995

**15.-** P.Kroon E.F. Deprettere and R.J. Sluyter, "Regular-Pulse Excitation- A novel Approach to Effective and Efficient Multipulse Coding of Speech", IEEE Transactions on Acoustics Speech, and Signal Processing, Vol. ASSP-34 No. 5 October 1986 PP.1054-1063

**16.-** C.J. Leggetter and P.C. Woodland "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models". Computer Speech and Language vol. 9, pp. 171-185

**17.-** B.T. Lilly and K.K Paliwal, "Effect of Speech Coders on Speech Recognition Performance" ICSLP 1996

**18.-** J.S. Lim, A. V. Oppenheim "All-Pole Modeling of Degraded Speech",IEEE Transaction on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 3, June 1978

**19.-** C. Mokbel, L. Mauuary, D. Jouvet, J. Monne, C. Sorin, J. Simonin and K. Bartkova, "Towards Improving ASR Robustness for PSN & GSM Telephone Applications" 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications(IVTTA1994).

**20.-** J.D. Markel and A.H. Gray Jr., "Linear Prediction of Speech". Berlin, Heidellberg, New York: Springer-Verlag 1976

**21.-** D. Mansour and B.H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing Vol. 37 No 11, November 1989

**22.-** P.J. Moreno, "Speech Recognition in Noisy Environments" Ph.D. Thesis 1996 CMU.

**23.-** A.V. Oppenheim and R.W. Schafer "Discrete-Time Signal Processing" Prentice Hall New jersey.

**24.-** M.R. Portnoff, "Short-Time Fourier Analysis of Sampled Speech", IEEE Transactions an Acoustics, Speech and Signal Processing, Vol. ASSP-29, No. 3, June 1981

**25.-** R.M. Stern, A. Acero, F.H.Liu, Y. Oshima, “Signal Processing for Robust Speech Recognition”, in Automatic Speech and Speaker Recognition, edited by Lee, Soon and Paliwal, Kluwer Academic Publishers, 1996.

**26.-** C.K.Un and K.Y. Choi, “Improving LPC Analysis of Noisy Speech by Autocorrelation Subtraction Method”, ICASSP 1981.

**27.-** H. Van Hamme, “ARDOSS: Autoregressive Domain Spectral Subtraction for Robust Speech Recognition in Additive Noise.”, ICSLP 94, Yokohama, Japan, 1994.