Speech Recognition in Telephone Environments

Pedro J. Moreno

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

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

December 1992

List of Figures

Abstract

This work compares the automatic recognition of telephone-quality speech with that of speech recorded in noiseless and acoustically controlled environments.

The report includes a description of the SPHINX speech recognition system, a general overview of the field of robust speech recognition, transmission characteristics of the telephone network, a description of the speech databases used, and the speech recognition experiments that have been done.

Speech recognition systems work very well dealing with high-quality speech; however, when the speech signal has been corrupted by the telephone network, recognition accuracies decrease dramatically. In the databases we use in this research, a relative increase of about 20% in recognition error rate occurred when telephone speech was used.

We study some possible factors that could account for this loss. Bandwidth reduction and the presence of low-frequency tones can explain part of the loss but not all of it. A potentially improper tuning of front-end signal processing parameters is not the problem either. Preliminary results show that the standard model of environmental degradation that has successfully improved the recognition accuracy of SPHINX in office environments does not sufficiently describe the degradation that is introduced by transmission over a telephone channel. Environmental normalization algorithms such as CDCN *(Codeword Dependent Cepstral Normalization)* and RASTA *(Relative Spectral Processing)* that assume this simple model do not provide complete compensation.

Chapter 1 Introduction

Automatic Speech Recognition (ASR) is a very difficult problem. We are still many years away from unconstrained systems able to recognize speech under any circumstances. Early ASR systems obtained reasonable performance by bounding the problem with constraints such as:

- limited vocabulary size
- restrictive grammars
- use of isolated words (rather than continuous speech)
- development of speaker-dependent systems (rather than speaker-independent systems)
- use of speech recorded using a close-talking microphone (rather than speech from natural acoustical environments)

In recent years considerable progress has been made in addressing all of these restrictions. Nowadays continuous speech systems with vocabulary sizes of 5,000 to 20,000 words using sophisticated language models have been demonstrated by research laboratories. However, the performance of all these systems degrades dramatically when they are operated in different environmental conditions from the ones with which they were trained.

Extensive previous work has been carried out in the field of speech enhancement and speech recognition in adverse environments. For example, the work of Lim and Oppenheim [1979] represents a milestone in the general field of speech enhancement. They developed the estimate-maximize approach, which alternatively estimates the parameters characterizing the speech samples, and the additive unknown noise that degrades it. This work was extended by Ephraim *et a*l [1989], who characterized the input feature vector by a statistical model using a mixture of gaussians. The algorithm was successful in making speech more robust to noise.

Researchers have studied different distance metrics (Gray [1976]), developing special distortion measures that are more robust to the additive noise. The use of auditory models, which mimicking the functional organization of the human ear, has been also investigated (Seneff [1988] and Zue [1990]).

Acero [1990] introduced a series of techniques to deal with the dual problem of additive uncorrelated noise and linear channel distortion simultaneously. The work of Acero clearly shows how the degraded speech can be corrected without any *a priori* knowledge of the statistics of the noise, or of the frequency response of the channel.

In Chapter 3 a more detailed review of previous work in speech recognition in adverse environments is provided. Previous research in acoustically-robust speech recognition has focussed on several environments including open-plan offices, aircraft cockpits, motor vehicles, and long distance telephone lines. This project is concerned with the problem of speech recognition over telephone lines.

Attributes of the telephone network that impact adversely on speech recognition include the following problems:

- stationary noise
- non-stationary noise
- low frequency tones
- nonlinear distortion
- band-limitation to a 300 to 3400-Hz channel
- a channel response that changes from call to call

The purpose of this work is to obtain a better understanding of the degradations that adversely affect speech recognition over the long distance telephone network, and to understand how these degradations may be overcome by improved signal processing techniques. While these studies are carried out in the context of research attempting to improve speech recognition accuracy using specific speech databases, the major goal is to obtain a better fundamental understanding of the telephone environment, the particular problems it poses, and to establish an appropriate direction for future work in the field. For this reason we have restricted some of the conditions in which the experiments have been run, such as the use of no grammar, and the use only of context-independent models in the recognition systems. This work is focused on the "front end" of the speech recognition system, or the module that takes a segment of speech and converts it into an n-dimensional feature vector. No attempt has been made to modify the recognition engine to achieve environmental robustness.

The outline of this report is as follows:

- Chapter 2 provides a brief description of the SPHINX recognition system, with special emphasis in the front end
- Chapter 3 describes the speech databases used in this research
- Chapter 4 provides a review of previous work in the area of speech recognition in adverse environments
- Chapter 5 provides a qualitative description of the telephone channel
- Chapter 6 describes a set of experiments and their results, examining the effects of the telephone channel on the accuracy of speech recognition systems
- Chapter 7 summarizes the results of this research

Chapter 2
The SPHINX Speech Recognition System

The SPHINX system was developed at CMU by Lee *et al* [1989]. It was a pioneer in speaker-independent large-vocabulary continuous speech recognition. We briefly describe the different blocks that compose the system.

A block diagram of the front end of the SPHINX system is shown in Figure 2-1.

2.1 Signal Processing

All speech recognition systems use a parametric representation rather than the waveform itself as the basis for pattern classification. These parameters carry information about the spectrum. SPHINX[1] uses the frequency-warped LPC *(Linear Predictive Coding)* cepstrum as features for speech recognition. The cepstral coefficients are computed as follows:

- Speech is digitized at a sampling rate of 16 kHz
- A Hamming window of 20 ms (320 samples) is used every 10 ms (160 samples)
- A preemphasis filter $H(z) = 1 - 0.97z^{-1}$ is applied
- 14 autocorrelation coefficients are computed
- A Pascal window is applied to the autocorrelation sequence
- 14 LPC coefficients are derived from the Levinson-Durbin recursion
- 32 LPC cepstral coefficients are computed using a standard recursion
- The cepstral coefficients are frequency warped by using a bilinear transform producing 12 warped LPC cepstral coefficients

Although adjacent frames of speech are indeed correlated with each other, the SPHINX system assumes that every frame is statistically independent. In addition to the static information provided by the cepstrum, SPHINX also uses dynamic information represented by first-order and second-order differences of the cepstral vector.

2.2 Vector Quantization

Once the incoming speech has been converted to an *n*-dimensional vector, a data reduction technique known as vector quantization (VQ) is used to map each vector into a discrete symbol. A vector quantizer is defined by a codebook and a distortion measure. The codebook contains L vectors, and it is a quantized representation of the vector space. The distortion measure estimates the degree of proximity between two vectors. An input vector is mapped into a symbol in the alphabet by choosing the closest codebook vector according to the distance metric. In the SPHINX system L is fixed to 256, so a codeword can be represented with just one byte. The distortion measure used is the Euclidean distance.

[1] This description refers to the baseline or default SPHINX system, which is the platform used for the experiments described in Chapter 6.

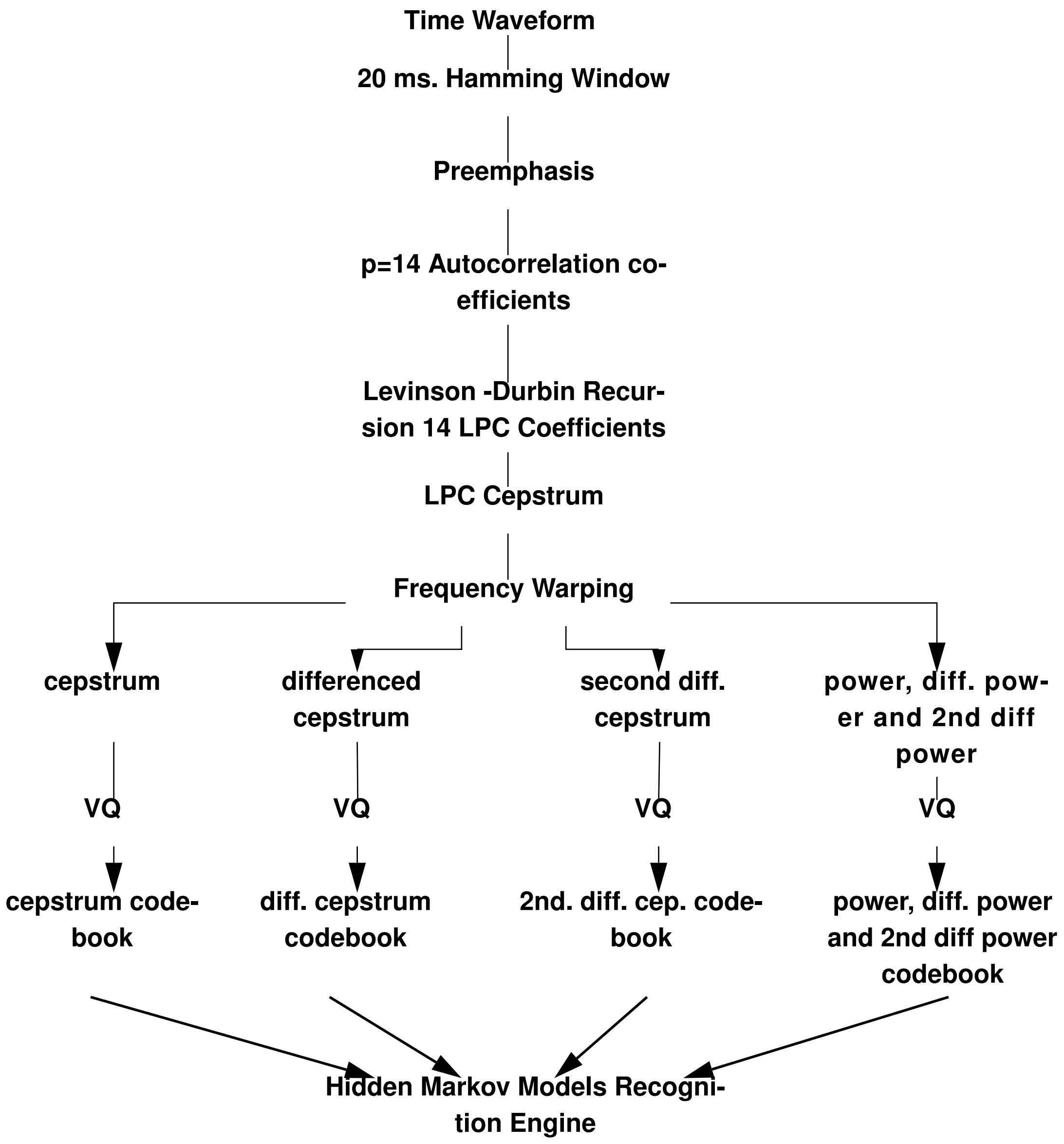

Time Waveform
20 ms. Hamming Window
Preemphasis
p=14 Autocorrelation coefficients
Levinson -Durbin Recursion 14 LPC Coefficients
LPC Cepstrum
Frequency Warping
cepstrum
differenced cepstrum
second diff. cepstrum
power, diff. power and 2nd diff power
VQ
VQ
VQ
VQ
cepstrum codebook
diff. cepstrum codebook
2nd. diff. cep. codebook
power, diff. power and 2nd diff power codebook
Hidden Markov Models Recognition Engine

Figure 2-1: Block Diagram of the SPHINX System Front End

Strictly speaking, there is no need to use VQ in speech recognition systems, and there are several arguments in favor and against the use of these techniques in Hidden Markov Model-based systems. In the SPHINX system VQ is used for two main reasons:

- It achieves great data compression, reducing a frame of speech (320 numbers) to a 12-dimensional vector, and in turns reducing this vector to one byte (assuming that a 256 codebook size is used).
- It reduces the computational load in the recognition engine (HMM), allowing it to work with discrete symbols as opposed to a continuous space of vectors.

The *L* prototype codebook vectors are calculated via a hierarchical clustering algorithm, which provides and estimate that approximates the maximum-likelihood estimate.

2.3 Hidden Markov Models

The Hidden Markov Model (HMM) is the dominant technology in continuous speech recognition, and it is the basic technique used for speech recognition in the SPHINX system. A good introduction to this technique is found in Rabiner and Juang [1986]. Briefly an HMM is a collection of states connected by transitions. Each transition is described by two sets of probabilities:

- A transition probability which describes the probability of a transition from one state to the next
- An output probability density function, which defines the conditional probability of emitting each output symbol from a finite alphabet given that a particular transition takes place.

HMM’s have become a widely-used approach for speech recognition due to the existence of maximum likelihood techniques to estimate the parameters of the models and efficient algorithms to find the most likely state sequence.

Chapter 3
The TIMIT/NTIMIT and AN4 Databases

In this chapter we provide a detailed description of the phonetically-labelled TIMIT and NTIMIT databases which are used to evaluate the effect of the telephone channel on the SPHINX recognition system. Our experiments follow the general form of earlier work of Lee and Hon [1988], who modified the SPHINX system so that it could perform phonetic recognition. (The original SPHINX system had been designed to recognize words rather than phonemes.)

3.1 Description of the TIMIT Task

The TIMIT acoustic/phonetic database (Larner [1986], Fisher [1987]) was originally developed to train and evaluate speaker-independent phone recognizers. It consists of 630 speakers, each saying 10 sentences, making a total of 6300 sentences. From this set 2830 sentences are used for training and 160 for testing.

In our work we used the same subset of the database selected by Lee and Hon [1988]. They performed some modifications on the original labels. In particular they removed all glottal stops from the labels, they merged 15 allophones with their corresponding phones, and finally they identified five groups for which within-group confusions are not counted[1].

3.2 Description of the NTIMIT Task

The NTIMIT database (described by Jankowski *et al* [1990]) was collected by transmitting the TIMIT database over the long-distance telephone network. The database is orthographically and phonetically equivalent to the TIMIT database.

Speech transmission was achieved by creating a “loopback” telephone path to a large number of central offices which were distributed geographically to simulate different telephone network conditions. Half of the database was transmitted over local telephone paths, while half was transmitted over long-distance paths. Transmission involved the use of an artificial mouth to simulate the acoustic frequency response of the channel between the human mouth and a telephone handset.

The entire database was collected using only a single carbon microphone, and the signal levels were adjusted to take advantage of the maximum dynamic range available in the channel. These two features represent the only major flaws of the database. It would normally be preferable to include a number of different microphones, as the microphone is a very important source of variability in the network. Similarly, the artificial adjustment of signal level to obtain maximum dynamic range is also an unrealistic condition.

3.3 The AN4 Database

Since a training and testing cycle in the TIMIT/NTIMIT databases can take up to 24 hours, we also use the smaller AN4 database in some experiments, which reduces the training and testing cycle to only 10 hours. We frequently conducted pilot experiments using the AN4 task and later confirmed our results using

[1] For a more detailed description of the phones see Lee and Hon[1988].

the TIMIT/NTIMIT task, in the belief that results from experiments performed using the AN4 database task are representative of those obtained using the more time-consuming TIMIT/NTIMIT databases.

The AN4 database is composed of 74 speakers(53 males, and 21 female) for training, and 10 speakers (7 male and 3 female) for testing. The training speakers and the testing speakers are different.The database consists of strings of letters, numbers, and a few control words. After discarding some utterances because of bad recording conditions, the training set used for our experiments consists of 1018 utterances and the testing set contains 140 utterances. In all our experiments a reduced version of the database containing 500 utterances was used for training.

The AN4 database was recorded in stereo, so two speech samples are available for every utterance. The primary channel was recorded using a Sennheiser HMD224 close-talking microphone, and a second channel of speech was recorded using the omnidirectional desk-top Crown PZM6fs microphone. Only the Sennheiser HMD224 was used for our experiments.

A downsampled version of this database was used, passing the close-talking utterances through a low-pass filter with a 3600-Hz cutoff frequency.

3.4 Summary

The TIMIT database is not an ideal one, as it was created primarily for phonetic research. It presents several problems, including computational load. Although the use of a telephone database based on a word-based task with a grammar might have been more realistic, the use of the TIMIT/NTIMIT phonetic-based database provides a measure of the degradation introduced by the telephone channel that is independent of any particular task. For researchers interested in telephone speech, NTIMIT is of especial interest because it can function as a stereo pair of the TIMIT database, enabling comparisons between “clean” speech recorded using the close-talking microphone and speech over telephone channels.

Finally we have introduced the AN4 database, which is used as a tool to evaluate candidate hypotheses before testing them using the more time consuming TIMIT/NTIMIT database.

Chapter 4
Speech Recognition in Adverse Environments: Previous Work

In this chapter we summarize a number of the techniques that have been used to make speech recognition more robust in the presence of additive noise and/or an unknown linear filter in the channel. We consider in this chapter modifications to the traditional LPC representation, the use of peripheral signal processing based on the human auditory system, the use of non-lexical models to characterize transient noises and distortions, and the use of various noise subtraction and channel equalization techniques, applied in isolation and jointly.

4.1 Modifications in the LPC

Linear predictive coding (LPC) techniques assume that the signal spectrum can be represented by an all-pole transfer function:

$$A(z) = \frac{G}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}}$$

When the signal is corrupted by additive noise, zeros as well as poles will appear in the transfer function representing the signal. The goal of these techniques is to estimate the LPC parameters of the clean signal when a noisy version of the signal is observed. Lim [1983] describes several possible methods based on this model.

Lim describes an iterative method where the α_{i} coefficients are estimated using a MAP estimate, thus maximizing $p(\alpha | \boldsymbol{y})$, where **y** represents the noisy speech (assuming uncorrelated white noise). After an initial estimate of the α_{i} coefficients, an estimate of the clean speech is obtained. With this new estimate of clean speech a new estimate of the α_{i} is obtained, which in turns helps to obtain a new estimate of the clean speech. The procedure is guaranteed to increase the probability of the α_{i} parameters and clean speech, given the noisy speech.

These techniques help to some extent, but they do not deal with the problem of the frequency response of telephone channel, a great source of variability in the telephone network (see Chapter 5 for more details). For these reasons we didn't consider these techniques in our research.

4.2 Physiologically-Motivated Front Ends

Since the human auditory system is very robust to changes in the acoustical environment, some researchers have tried to develop signal processing schemes that resemble the functional organization of the peripheral auditory system. Seneff [1988] and Zue [1990] used models based in the human auditory system as the front-end processing for their recognition systems.

Although some of these techniques seem to improve the robustness of speech recognition systems, the computational cost is very high when compared with the improvements the techniques bring. In general with less complex techniques better results can be obtained with less complex techniques.

We do believe, however, that some of the ideas introduced by these methods could be incorporated in the front ends used in speech recognition systems, such as the compression of the frequency scale.

4.3 Noise and Noise-Word Models

Ward [1989] and Wilpon [1990] introduced a novel approach for combatting the effects of transient noise. Specifically, they proposed the idea of developing HMM's to characterize transient nonspeech sounds such as filled pauses ("um", "ah", etc.), telephone rings, rustling of paper, and other normal environmental sounds. These non-speech acoustical events were then recognized in the usual fashion, and ignored in subsequent processing. To implement this technique a database of speech with labelled noise segments is needed. At the same time the technique cannot work if the noise conditions are very different in the testing and training, it does not adapt to different environmental conditions.

Gales and Young [1992] present a way of detecting speech and noise simultaneously. Unlike the previous method where noise and speech are not assumed to occur simultaneously, this method recognizes the fact that noise and speech occur at the same time, and tries to match the input speech with a pattern that mixes speech and noise models. HMM patterns are used for the noise and the clean speech words. The method involves the use of a three-dimensional Viterbi search, in which the noise is decoded at the same time than the speech.

While this technique is very powerful, the computational cost is extremely high, since the Viterbi search algorithm has to be solved in three dimensions. This technique does not provide for any compensation of the linear filtering by the channel.

4.4 Noise Cancellation

Noise cancellation or spectral subtraction refers to a family of techniques designed to suppress or reduce the noise in a signal when this noise is stationary. If $x[m]$ is a speech signal, and $n[m]$ is additive noise that is uncorrelated with the speech, then:

$$y[m] = x[m] + n[m]$$

where $y[m]$ represents the noisy speech.

Defining $X(\omega)$, $Y(\omega)$ and $N(\omega)$ as the power spectral densities (PSD) of the signals $x[m]$, $y[m]$ and $n[m]$, we can assume that the additive relation above holds also in the frequency domain. If we can obtain $\hat{N}(\omega)$, an estimate of the PSD of the corrupted signal in regions where it is determined that no speech is present, then an estimate of the PSD of the clean signal is:

$$\hat{X}(\omega) = \hat{Y}(\omega) - \hat{N}(\omega)$$

Acero [1990] provides a good general description of spectral subtraction techniques. Using the AN4 database, he reports good performance in crossed conditions, *i.e.* when the system is trained on clean speech and tested using noisy speech.

4.5 Spectral Equalization

Spectral equalization is a filtering operation used to compensate for spectral tilt and other types of linear filtering. Stockham [1975] introduced the *blind deconvolution* technique in the cepstral domain for reducing the effects of resonances and reverberation introduced by the recording equipment in musical recordings of Caruso. Acero [1990] describes several implementations of the technique in his review of previous work. His results suggest that channel equalization is important, but by itself it is unable to recover completely the degradations introduced by a combination of additive noise and linear filtering.

4.6 The RASTA Method

The RASTA (Relative Spectral Processing) method (Hermansky *et al* [1991]), is a technique in which constant factors in each spectral component of the spectrum are suppressed by passing the running spectral estimate through a band-pass of high-pass filter with a sharp spectral zero at zero frequency. The technique can be applied in the spectral domain to suppress stationary noise, or in the log-spectral domain to suppressing the effects of linear filtering, and the use of a filtering approach to normalization is appealing because of its computational simplicity.

While the RASTA technique produces good results, it has two problems. First, part of the speech spectrum is also removed in the filtering operation, particularly if the $c[0]$ component of the cepstral vector (the frame energy) is included in the filtering operation. In addition, the RASTA method is not able to deal with the combined effects of noise and filtering by the channel.

In chapter 6 we report some results of applying this technique to the TIMIT/NTIMIT database.

4.7 The CDCN Algorithm

Acero [1990] introduced CDCN (*Codeword Dependent Cepstral Normalization*) as a technique for dealing jointly with additive noise and channel equalization. Although the technique is computationally costly, it provides a way of dealing with two main problems in the TIMIT/NTIMIT task, the variation of the telephone channel from sentence to sentence, and the additive noise present in the telephone network.

Given the observed noisy speech, CDCN attempts to estimate for each frame a spectral equalization vector ***q*** and a noise correction vector ***n***. These vectors are chosen to best match the ensemble of cepstral vectors of the incoming speech to the ensemble of cepstral vectors in the training corpus after being subjected through a transformation that compensates for the effects of noise and filtering.

The cepstral vector of the clean speech is estimated with a MMSE estimator:

$$\hat{\boldsymbol{x}}_{MMSE} = \boldsymbol{z} - \boldsymbol{w}$$

where ***z*** is the observed noisy speech vector and ***w*** is a correction vector that itself uses the information represented by the normalization to the training corpus, incorporating the estimates of the noise ***n*** and spectral equalization ***q*** vectors.The estimation of the ***n*** and ***q*** correction vectors is also performed according to

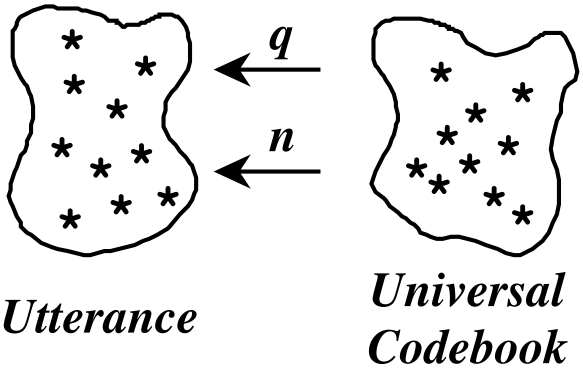


Figure 4-1: CDCN estimates a noise correction vector **n**, and a channel correction vector **q** that best transforms the universal codebook into the set of input frames of the observed speech.

a maximum likelihood approach, given the knowledge of the universal acoustic space of the training corpus and the observed speech.

Since this technique is one of the most promising of previous results, we will attempt to determine the extent to which CDCN is able to recover from the degradations introduced by the telephone channel.

4.8 Summary

In this chapter we have given an overview of most of the techniques already available for the use of automatic speech Recognition in adverse environments. Most of these techniques (with the exception of CDCN) address only one of the two major effects introduced by the telephone network, additive noise and linear filtering.

Chapter 5
The Telephone Network

In this chapter we describe the telephone network in terms of its frequency response, signal-to-noise ratio, noise, and nonlinearities. Most of this information is based on work carried out in Bell Laboratories, and refers to the network as of 1982. The network has changed in recent years, going to a more digital network. Most of the data reported in this chapter refers to end-office to end-office communications. In other words, it does not include either the two customer loops or the telephone microphone. The loop contribution to the parameters discussed here can for the most part be considered negligible except for the frequency response of the channel. The information in this chapter is based on research performed by Carey *et al* [1978].

5.1 Frequency Response

Channel response or frequency response is a measure of loss over the frequency band in a communication channel. Figure 5-1 describes the channel response over short, medium and long distance lines, where short distance lines are those with a range of *0* to *180* miles, medium distance are those with a range of *180* to *720* miles, and long distance lines are those with a range of more than *720* miles. This mileage refers to direct distance, (point to point), and the actual distance covered by the signal can vary.

In Figure 5-1 the channel response[1] for end-office to end-office connections is shown. Surprisingly, short distance lines have a worse response than long or medium distance lines. This is due, according to

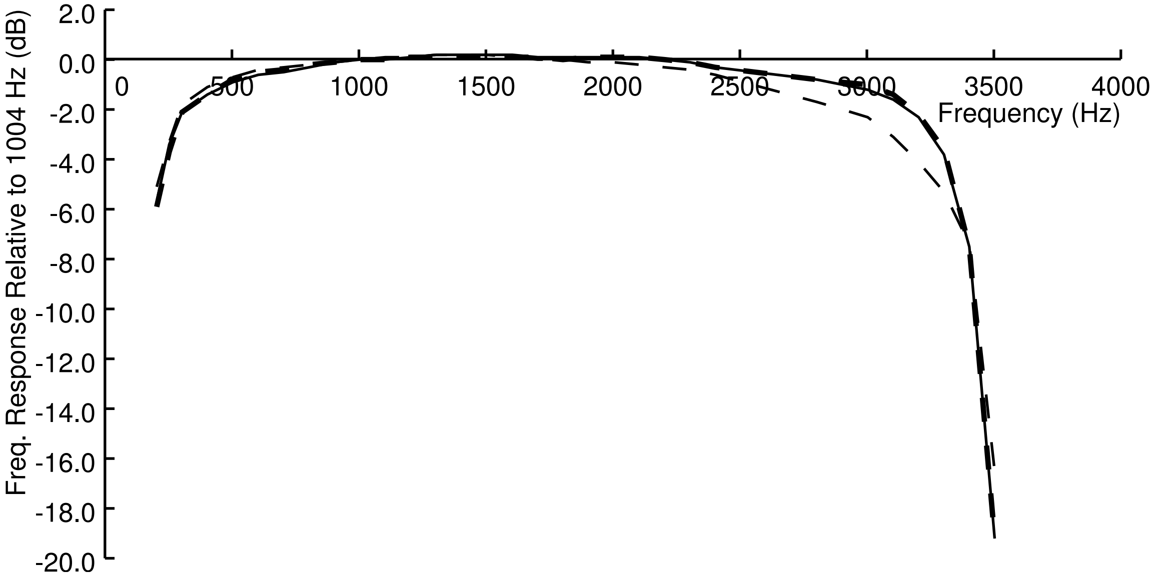


Figure 5-1: Attenuation Distortion Relative to 1004 Hz for short (dashed line), medium (upper dashed line), and long distance (continuous line) mileage bands.

[1] In telephony the channel response is always given relative to 1004 Hz. This frequency is very near the maximum sensitivity point in the human hearing frequency response (1000 Hz). The actual frequency of 1000 Hz is avoided as a reference because if the speech signal is sampled at 8000 Hz, harmonics of a 1000-Hz test signal could beat with the 8000-Hz sampling rate, producing an adverse effect.

Carey *et al* [1984], to the use of analog transmission media for short distance, while long and medium distance connections are carried out on T1 digital carriers.

When the effect of the customer loop is added, the response is similar but with a higher attenuation at higher frequencies, as can be seen in Figure 5-2.

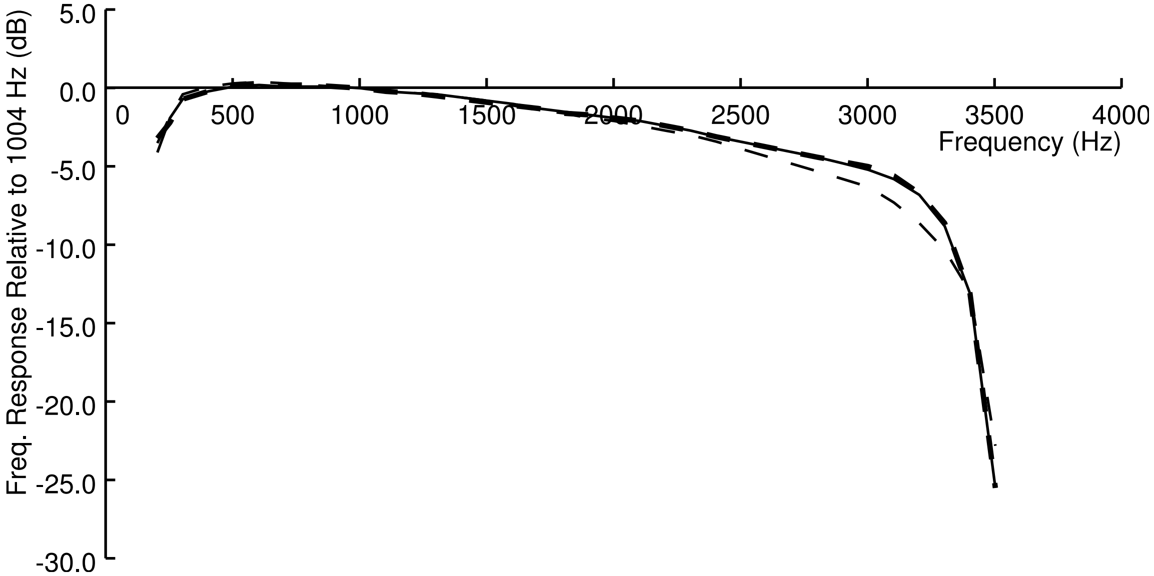


Figure 5-2: Mean customer-to-customer attenuation distortion relative to 1004 Hz for short (dashed line), medium (upper dashed line), and long distance (continuous line).

5.1.1 The carbon microphone

It is important to note that in the previous figures the microphone effect is not included. Modern microphones such as capacitance microphones exhibit a frequency response that tends to be stable across different microphones and quite flat. However carbon microphones, which are still present in the telephone network, exhibit a response that is not flat at all, presenting also some nonlinear effects, such as different sensitivities to different signal levels and also a different response depending on the polarization current. Finally we note that the carbon microphone tends to attenuate low-level signals that are simultaneously presented with high-level signals.

In Figure 5-3 a typical transfer function curve is presented for a carbon microphone.

5.2 Nonlinearities in the Network

The telephone channel is not a perfect linear channel, with sources of nonlinearity including PCM quantization and analog amplifiers working in nonlinear regions. These effects are generally described as distortions. Two types of distortion measurements are used in the telephone literature to characterize the nonlinear effects of the channel, *harmonic distortion*, and *intermodulation distortion.* The first refers to the creation of unwanted frequency components at harmonics of the input signal. The second one occurs when there are two or more frequency components in the channel, creating distortion products at frequencies equal to sums and differences of the harmonics of the input signals.

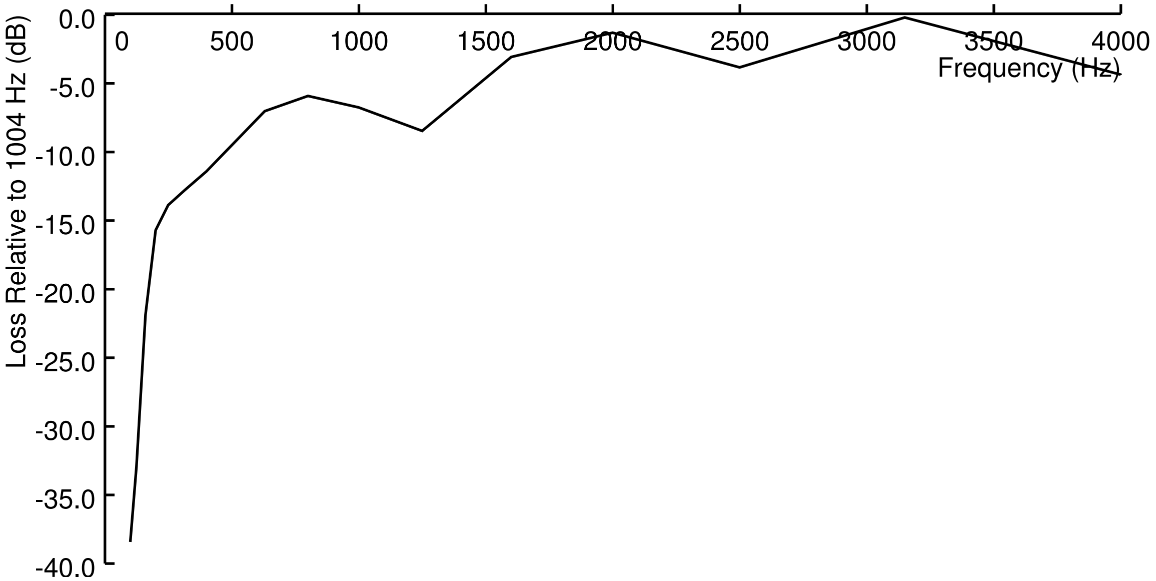


Figure 5-3: Sensitivity curve for a typical carbon microphone.

In telephony, harmonic distortion is expressed as the ratio in dB between the fundamental and the distortion products of interest (in general second and third harmonics only). Intermodulation distortion is generally preferred as a metric to assess nonlinearities in the net since it better describes the subjective distortion a speech signal will suffer.

5.3 Types of Noise

In general the noise in the network can be classified as stationary and non-stationary. Stationary noise appears over the telephone network in the form of low-frequency tonelike signals, or white noise. The white noise is generally produced by thermal and other physical phenomena. The single frequency noise can be produced by harmonics of the 60-Hz power lines, and sometimes by signalling tones that get transmitted by error through the telephone channel. Non-stationary noise over the network is caused by clicks and other transient phenomena caused by intermittent connections (*e.g.* a loose resistor or corroded cable).

Chapter 6
Recognition in the TIMIT/NTIMIT Task

In this chapter we report the results of our experiments using the TIMIT/NTIMIT database and the AN4 database. In Sections 6.1 and 6.2 we describe previous results obtained with the TIMIT and NTIMIT databases. In Section 6.3 we describe our baseline system and the first results in the NTIMIT task. We explore in this section possible explanations for the reduction in performance observed in the NTIMIT baseline system. In Sections 6.4 we discuss the improvements available for telephone speech by manipulating various aspects of the initial LPC signal processing. In Section 6.5 we describe the results obtained with the environmental compensation algorithms CDCN and RASTA. In Section 6.6 we discuss the use of an alternative to LPC cepstral coefficients. Finally, in Section 6.7 an unsuccessful scheme to deal with the problem of the nonlinear distortions present in the telephone channel is described.

6.1 Previous Results on TIMIT

In this section we refer mainly to the results of Lee and Hon [1988], who applied the SPHINX system to the recognition of individual phonemes. Lee and Hon used context-independent and context-dependent[1] HMM's, with and without grammars. A summary of their results is shown in Table 1.

Language Model	Context-Independent Recognition Rate	Context-Dependent Recognition Rate
Bigram	64.07%	73.80%
Unigram	60.91%	70.38%
None	58.77%	69.51%

Table 1: Recognition Results of Hon and Lee

These results were obtained with an LPC analysis of order 14, three codebooks, speech sampled at 16 kHz, a window size of 320 samples, a window step of 160 samples, and a bilinear warping coefficient of 0.6.

6.2 Previous Results on NTIMIT

To our knowledge no other researchers have used the NTIMIT database for phonetic recognition. Some experiments have been done however in phonetic classification[2] using the NTIMIT database. In the most recent such study, Chigier and Leong [1991] compare the performance of a Gaussian classifier with the two databases, TIMIT and NTIMIT. Their results are summarized in Table 2. They found that the first-choice error rate shows a degradation of about 8% in performance when going from the TIMIT to NTIMIT database. They didn't try any compensation or enhancement technique.

[1] For a better understanding of these concepts, Lee's book [1989] provides a good introduction.

[2] Phonetic classification as opposed to phonetic recognition assumes knowledge of the beginning and end of every phone in the sentence. This task is much simpler and a Gaussian Classifier or a neural net can produce good results.

Database	First Choice Error Rate	Top 2 Choices Error Rate	Top 3 Choices Error Rate
TIMIT	25.2%	11.6%	6.4%
NTIMIT	33.5%	17.5%	11.4%

Table 2: Recognition Results of Chigier and Leong

6.3 Baseline System

In our first experiments we tried to replicate the results reported by Lee and Hon [1988]. Their speech recognition experiments with the TIMIT task were slightly different from ours, as they used a front end based on 3 codebooks, and their training set contained 2830 sentences. In our experiments 4 codebooks have been used, and the number of training sentences was reduced to 2000. In pilot experiments this reduction in the size of the training set produced a negligible difference in performance, but it reduced the computational load by one third. On the other hand the addition of one extra codebook produced better recognition accuracy, since this extra codebook increases the discriminatory capabilities of our recognition system.

It is also important to note that Lee and Hon used a different recognition accuracy metric which did not consider insertions of phonemes as errors. They chose a language weight that balanced substitution errors with the insertion rate, always maintaining the insertion rate below 12%. Since the time of the experiments of Hon and Lee the DARPA community has standardized on a different metric that penalizes insertion errors as well as substitution and deletion errors (see formula below). The standard DARPA metric is used in our research:

$$Recognition\ Accuracy = \frac{Total - Substitutions - Deletions - Insertions}{Total}$$

where *Total* refers to the total number of phonemes in the utterance and *Substitutions, Deletions,* and *Insertions* refers to the number of substitutions, deletions, and insertions, respectively.

In contrast, Hon and Lee used the metric

$$DetectionRate = \frac{Total - Substitutions - Deletions}{Total}$$

In Table 3 we compare on the TIMIT database our 4-codebook baseline system with the Lee and Hon baseline system, which used only 3 codebooks. We also report the initial results obtained using the NTIMIT database. Our baseline system uses 16 kHz sampled speech as input, a Pascal window with a τ parameter of 1500, a 14th-order LPC analysis, and an α warping coefficient of 0.6. The language weight used in these

TRAIN	TEST	SYSTEM	DETECTION	RECOG. ACCURACY
TIMIT	TIMIT	Lee	58.77%	46.77%[a]
TIMIT	TIMIT	Moreno	60.50%	48.7%
NTIMIT	NTIMIT	Moreno	49.1%	38.3%

Table 3: Previous versus New Results

[a] Not provided by Lee and Hon, this figure is estimated by assuming a 12% insertion rate

baseline experiments was the one that produced a 12% insertion rate in order to obtain a fair comparison with the results of Lee and Hon.

Our experiments were also run without any grammar, and only with context-independent units. These two conditions will be common to all experiments described in this report. In the rest of this chapter we adopt the more usual recognition accuracy metric, including the effect of substitutions, deletions and insertions. The insertion rate[3] used in every experiment is always the one that yields the best recognition accuracy.

From these baseline experiments, we can observe a degradation of about 10% with the use of telephone speech when compared to clean speech. In the next sections we attempt to find possible reasons for this strong degradation.

6.3.1 Bandwidth Reduction

The NTIMIT database is sampled at 16 kHz, producing, after antialiasing filtering, an effective bandwidth of 6.4 kHz (Jankowski [1990]). Since clean speech is sent through the telephone network, another reduction in bandwidth results. Therefore the 3.4 kHz-6.4 kHz band contains no speech information. To eliminate these components from further analysis a lowpass filtering followed by downsampling was performed on the NTIMIT database, reducing the bandwidth to 4 kHz.

As it can be seen in Table 4, low pass filtering and downsampling clean speech results in a rather small reduction in recognition accuracy. For NTIMIT, however, this downsampling slightly improves recognition re-

[3] Insertion rate is a parameter used to make the transitions between phones more or less likely, in general a small insertion rate produces more insertions, since the transitions from phone to phone are less penalized, in contrast a high insertion rate reduces the insertions.

sults. A possible explanation is that in the case of NTIMIT, all information over 4000-Hz can be considered noise, since the telephone network effectively filters out this band.

TRAIN	TEST	FREQ. RANGE	LPC ORDER	RECOG ACCURACY
TIMIT	TIMIT	8000 Hz	14	52.7%
TIMIT	TIMIT	4000 Hz	14	51.0%
TIMIT	TIMIT	250-3400 Hz	14	49.7%
NTIMIT	NTIMIT	8000 Hz	14	41.3%
NTIMIT	NTIMIT	4000 Hz	14	42.5%

Table 4: Effect of Downsampling on Recognition Accuracy

These results indicate that the degradation in recognition accuracy observed when changing from the TIMIT to NTIMIT data is not a consequence merely of bandwidth reduction. After reducing the bandwidth to 3600 Hz and losing all high-frequency components, only a 1.7% reduction in recognition accuracy is noted. Furthermore, when simulating the bandwidth reduction of a telephone channel (a passband channel in the 250-3400-Hz region) only 3% of accuracy is lost. When these results are compared to actual results using the NTIMIT telephone speech database, it is obvious that there must be other sources of degradation in the telephone network beyond mere bandwidth reduction.

6.4 Fine Tuning of the LPC Parameters

In this section we explore the possibility that the front end configuration, which was optimized for 16-kHz clean speech may not be optimal for 8-kHz telephone speech. Several aspects of the front end are studied, including the LPC order, the analysis window size, the Pascal window size, the effect of low-frequency tones on LPC modelling, and finally the frequency-warping coefficient.

6.4.1 LPC Order

One possible explanation for the reduction in recognition accuracy in the NTIMIT database could be that the LPC model for speech is inappropriate, since it is well known that the best LPC order for speech representation increases with an increase in bandwidth. The speech spectrum can generally be represented as having an average density of two poles per kHz due to the contribution of the vocal tract, plus two more poles to represent the source excitation and radiation load (Rabiner [1978]). In the case of 8-kHz speech this corresponds to approximately eight or nine poles. This has been confirmed in experiments using the AN4 database.

In Figure 6-1 we present a comparison of LPC order versus recognition accuracy for the AN4 database. As we expected the best results are obtained with an 8th-order LPC model. However, tests of statistical significance[4] using the McNemar test and a matched pair test (Gillick [1989]) indicate no significant difference

[4] Statistical significance addresses the possibility that differences in recognition accuracy between two algorithms tested on the same data may be due to chance. In general two tests are used for this purpose. The McNemar test is used for short sentences, and the matched pairs test is used in tasks with long sentences (Gillick [1989]).

between results for the various LPC orders, at the 95% confidence interval. Our experience running this test of statistical significance on the TIMIT/NTIMIT database suggests that a difference in performance of about 2% is evidence of improvement. For the AN4 database the percentage difference needed for statistical significance is around 3.5%, as there are fewer test utterances.

A similar experiment was performed on a downsampled version of the TIMIT database (see Figure 6-2). These results suggest that at least we can obtain as good results with an 8th-order LPC model as with a 14th order LPC model. The recognition accuracy remains more or less constant for an LPC order of 8 to 14. When a test of statistical significance is applied to these results we again obtain the same conclusion; at the 95% confidence interval there is no significant difference in recognition accuracy of all recognition systems with an LPC order of 8 or higher (matched pair by pair).

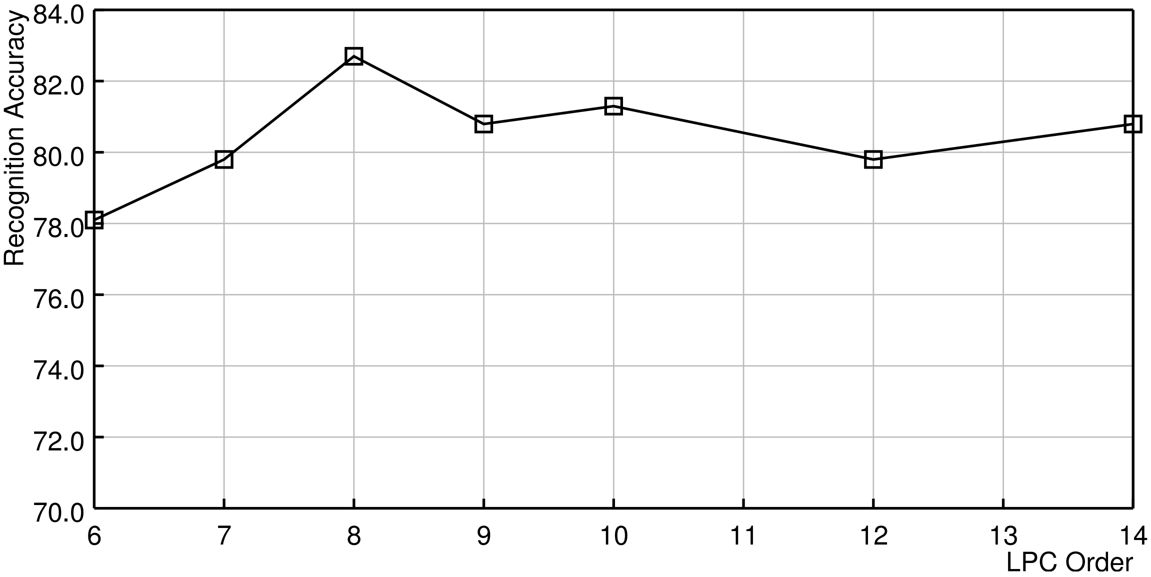


Figure 6-1:Recognition Accuracy versus LPC order for the downsampled version of the AN4 database.

Similarly the results of the same experiment on the NTIMIT database are shown in the same figure. In this case we do not observe any significant variation in accuracy for a wide range of LPC orders.

We conclude that the proper order of LPC model for all of these databases is 8 or higher, and that differences in recognition accuracy for higher orders in Figures 6.1 and 6.2 can be attributed to chance.

6.4.2 Analysis Window Size

Another of the parameters in the LPC analysis is the size *N* of the window. In general *N* must be on the order of several pitch periods to ensure reliable results (Makhoul [1975]). Since the sampling rate of the speech files is reduced by half, the size of the window must be reduced by one half to 160 samples in order to maintain the same 20-ms duration for the analysis window.

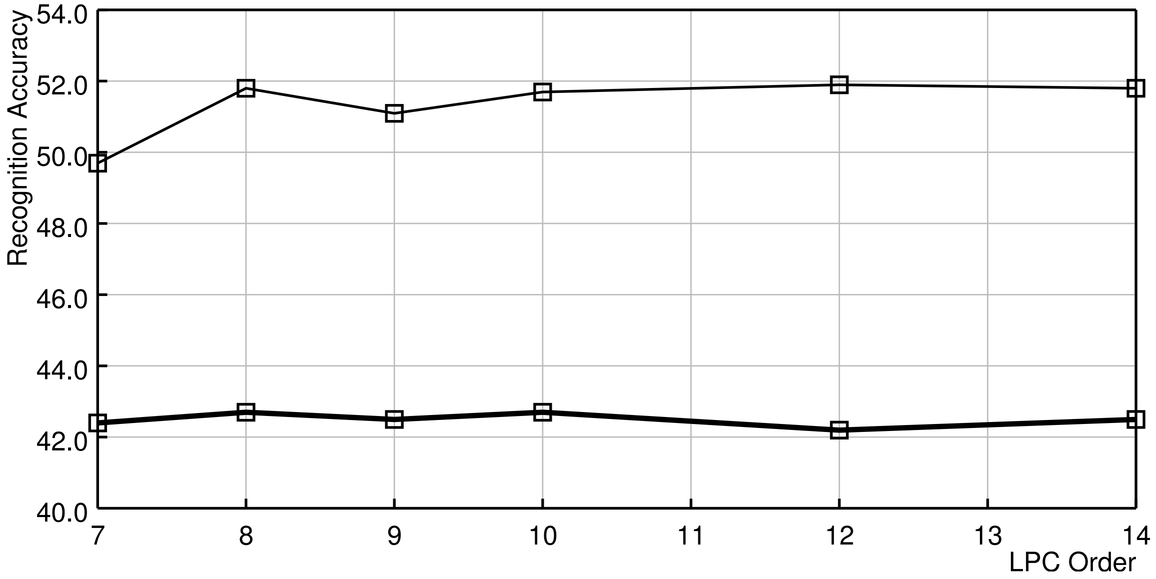


Figure 6-2: Recognition Accuracy versus LPC order for the downsampled version of the TIMIT database (upper curve) and the NTIMIT database (lower curve).

In Figure 6-3 we observe the effect of window size on the recognition accuracy of a downsampled version of the AN4 database.

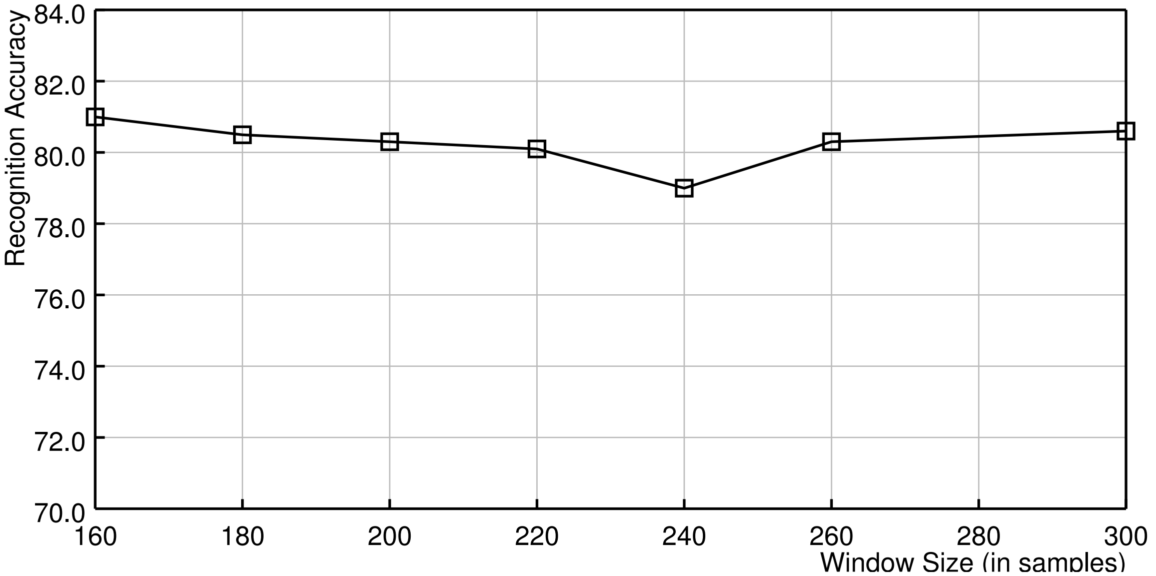


Figure 6-3: Effect of the window size (in number of samples) on the recognition accuracy in a downsampled version of the AN4 database.

From these experiments it seems that a window of 160 samples is best, maintaining the 20-ms window size, no matter what sampling rate we use. In the rest of the experiments reported in this chapter whenever the TIMIT or NTIMIT database have been downsampled, the analysis window contains 160 samples, with a time shift of 80 samples.

Tests of statistical significance suggest that the difference in performance due to the different window size is not due to chance for the AN4 database. Nevertheless, the results of similar tests using the NTIMIT database reveal no statistically-significant effect of window size.

6.4.3 Pascal Window Size

The autocorrelation function is multiplied by a Pascal window in the SPHINX front end. This window has the form

$$v[k] = \frac{\binom{\tau - k + 1}{k}}{\binom{\tau + k - 1}{k}}$$

where τ is set to 1500.

The purpose of this window is to suppress harmonics in the autocorrelation function that appear at multiple frequencies of the pitch period. For speech sampled at 16-kHz, these harmonics can appear at 32-sample lag intervals. Since the speech is downsampled to an 8-kHz sampling rate, these harmonics would appear at 16-sample lag intervals. In order to reduce this effect, a smaller value of the τ parameter would be needed.

In Figure 6-4 different Pascal windows are represented for different τ parameters. The narrowest window corresponds to a τ value of 200, and the widest one corresponds to a τ value of 3000.

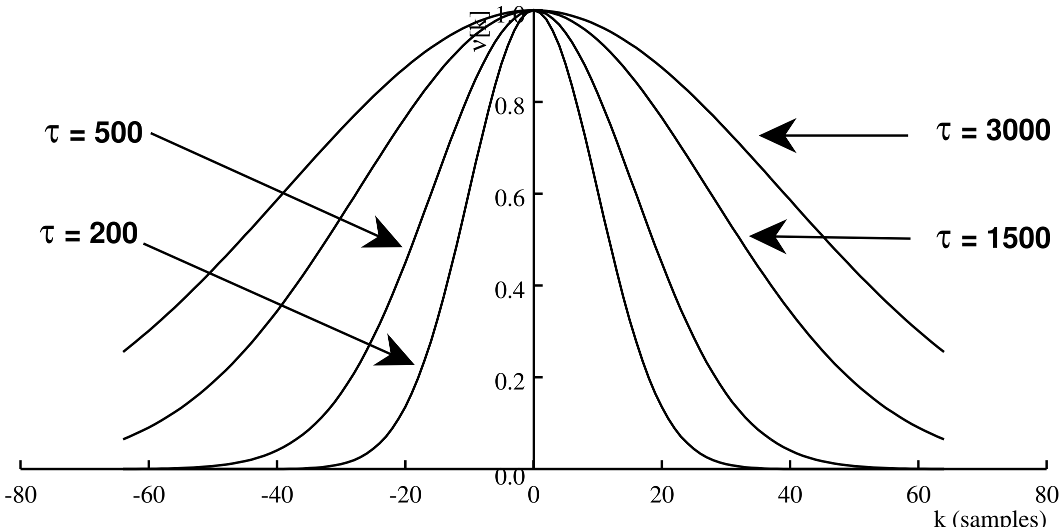


Figure 6-4: Pascal windows for a τ parameter of values 200, 500, 1500 and 3000.

A set of experiments was done on the AN4 database, using downsampled clean speech, to explore the possible effects of the Pascal window. In Figure 6-5 we show the results of experimenting with several Pascal window sizes. We obtained the best recognition result with a τ value of 500.

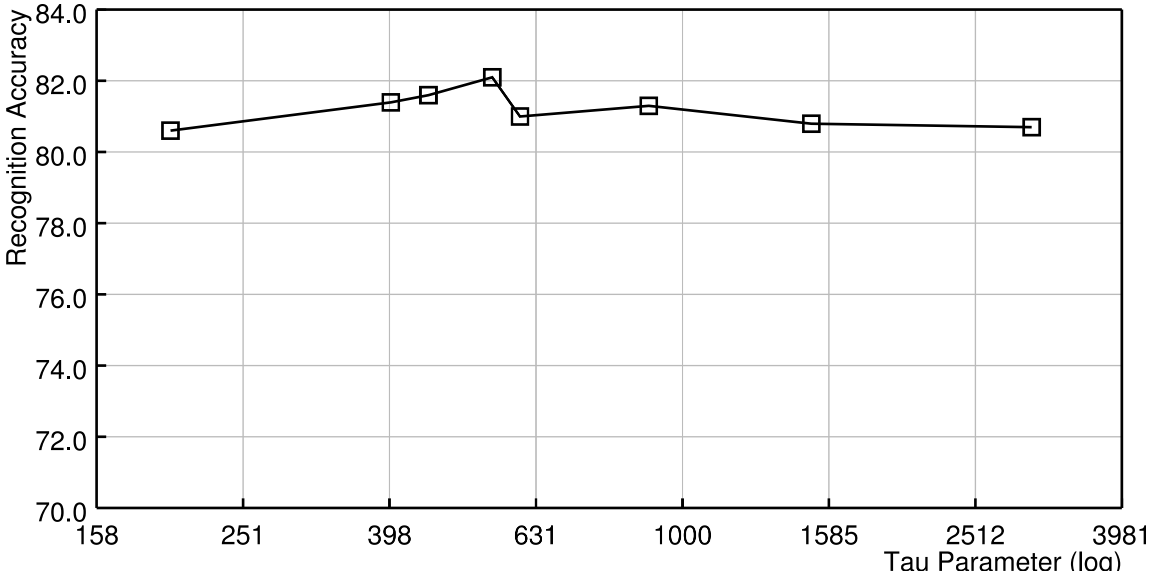


Figure 6-5: Recognition Accuracy versus τ value. The experiments were done on the AN4 database, using speech that was downsampled to 4000-Hz to simulate a telephone channel.

From our previous experiments with the AN4 database, we can assume that the difference in performance for different values of τ is insignificant. In similar experiments on the NTIMIT database no clear dependency on τ was observed.

The results are summarized in Table 5. The experiment used an 8th-order LPC analysis with downsampled speech.

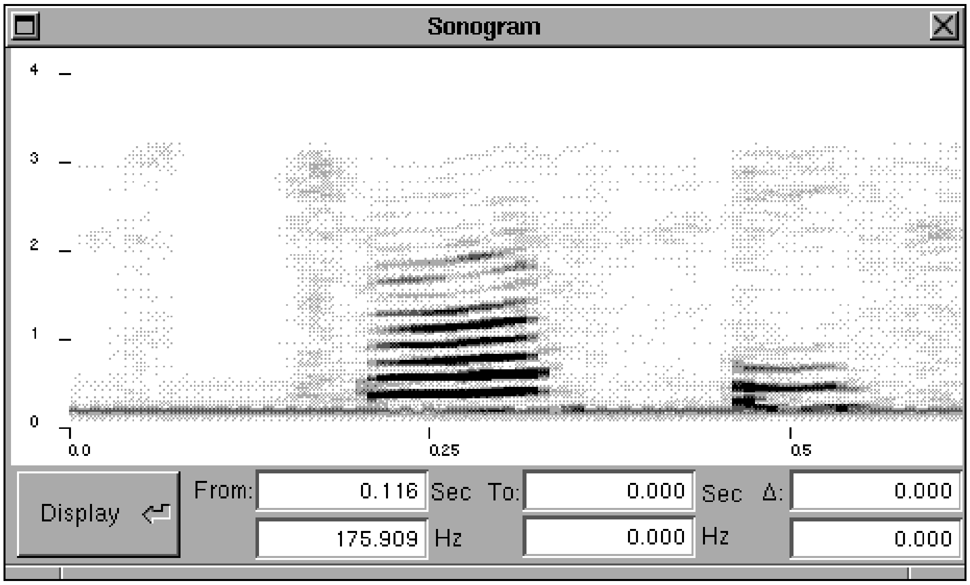
TRAIN & TEST	τ	RECOGNITION ACCURACY
NTIMIT	1500	42.7%
NTIMIT	500	42.8%

Table 5: Comparison of different Pascal τ parameters in the NTIMIT database

The motivation for using the Pascal window comes from the field of speech coding, where harmonics of the pitch period if not properly filtered out can be damaging in the speech quality. However its value in the filed of speech recognition remains unclear, as our experiments suggest that no clear improvement is obtained by using it.

6.4.4 Low-Frequency Tones

In our search for possible explanations for the 10% reduction in recognition accuracy using the NTIMIT database compared to the TIMIT database, a detailed audition of the testing subset of NTIMIT was performed, labelling the 160 sentences with 3 features: noisy, containing a low-frequency tone, and containing some kind of distortion. In this informal study we observed many utterances with some low-energy tones in the low-frequency band below 300-Hz, such as sentence *si567-b-fcmy0.adc* in the NTIMIT database. In Figure 6-6 we show a spectrogram of this particular sentence. The horizontal line at about 200 Hz indicates the presence of a tone at a signal level of about 20 dB below the normal signal. We believe that this kind of low-frequency tone can be due to AC-power interference in the local exchange.During the recognition process, these tones are detected as a normal word rather than as silence in intervals where there is no speech.



From: 0.116 Sec To: 0.000 Sec Δ: 0.000
175.909 Hz 0.000 Hz 0.000

Figure 6-6: Spectrogam of part of the utterance *si567-b-fdmy0.adc*: Notice the line at about 200-Hz representing a tone.

One possible way to deal with this problem is to build tuned Hidden Markov Models for the tones, in a similar way as was done by Ward [1989] and Wilpon [1990] building special models for the detection of “noise words” (*i.e.* non lexical utterances such as “um” or “ah”, or coughs, etc.). This technique of noise-word modeling, however, has been used only in small size vocabulary tasks, and with the help of language models (grammars). We believe that in the TIMIT/NTIMIT task where no grammar is used the technique would not be so effective.Another problem the noise-word method poses is that some kind of relabelling of the whole database (more than 2000 sentences) with presence-of-tone indicators would be necessary.

Since the tones appear consistently always at frequencies below 250 Hz, and the speech signal is not transmitted through the telephone net below 300 Hz, a simpler way to avoid this problem is to use a high-pass filter with cutoff frequency at 250 Hz.

In Table 6, we report recognition results obtained by passing the speech through a high-pass filter. We can observe that in spite of the tone reduction, the recognition accuracy decreases rather than increases.

HIGH PASS FILTER	RECOG. ACCURACY
NO	42.7%
YES	41.9%

Table 6: Using lowpass filtering in the NTIMIT database

These experiments were run with an LPC analysis of order 8, using downsampled speech.

In the next section a possible explanation for this reduction in recognition accuracy is given.

6.4.5 Channel Response and LPC Modelling

As we have shown in previous chapters, the speech in the NTIMIT database is filtered by the channel response of the telephone network that corresponds to a bandpass filter in the 300 to 3400-Hz band. We have also shown that best recognition results are obtained using an eighth-order AR process in the SPHINX front end (*i.e.* an all-pole model is used to characterize the speech spectra). While LPC is generally a very effective technique for speech modelling, we are attempting to model the filtered spectrum, so the channel is going to be modelled as well as the speech. In the following figures we show this effect.

The channel response of a typical telephone channel was already described in a previous section. In general it is well characterized as a strong attenuation at frequencies below 300 Hz, and also at frequencies above 3400 Hz. In Figure 6-7 the effect of the telephone channel can be observed. The FFT of a 20-ms window of telephone speech is shown. The effect of the channel can be observed in the attenuation of the spectrum at high and low frequencies. Naturally the LPC model is going to follow this attenuation in high and low frequencies. In doing so the LPC model is not characterizing so well the spectrum in the 300-3400-Hz band, where the speech energy is located.

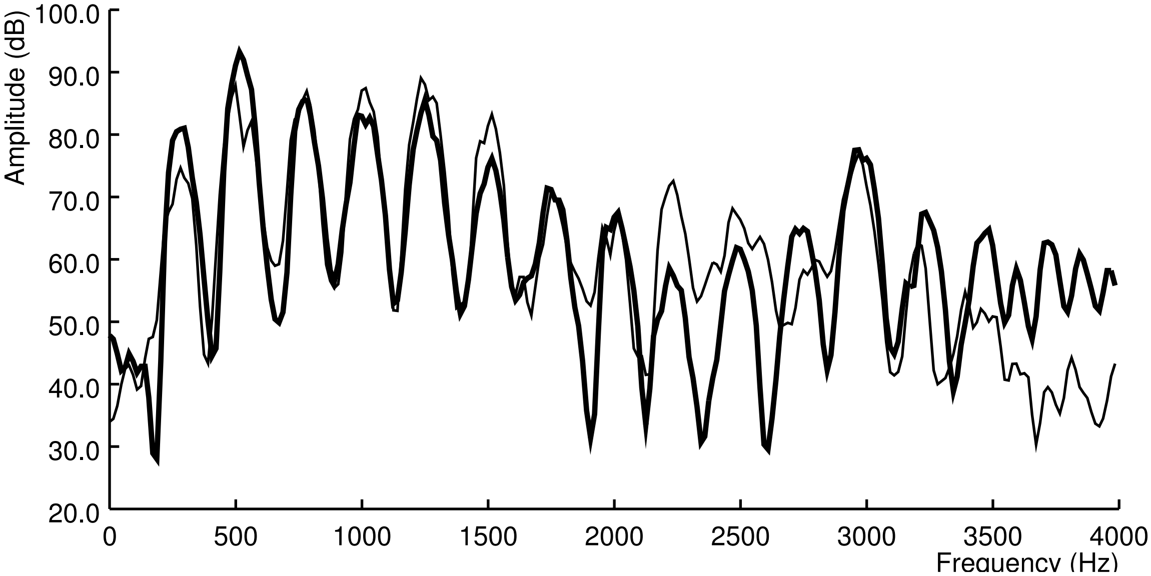


Figure 6-7: Comparison of NTIMIT speech (light line), versus TIMIT speech (bold line). The power is not normalized

6.4.6 Selective LPC Method

We believe that the introduction of a high-pass filter, proposed in the previous section to reduce the effect of the low-frequency tones, has a negative effect on the LPC modelling. Specifically, it increases the problem of the channel effect, so that the LPC model allocates more modelling effort to the combined high-pass filter and telephone-channel filter than to the speech spectrum.

Since the effects of the telephone channel are more apparent in the region of 0 to 300 Hz, and in the region of 3.4 to 4.0 kHz where the spectrum is highly attenuated, it would be interesting to model the spectrum only in the 300-Hz to 3400-Hz band. In this way we obtain two benefits:

- No LPC modelling effort is put into modelling the channel
- The tones below 300 Hz are filtered out

A simple technique for achieving this purpose is the selective LPC method, described in Makhoul [1970]. This technique allows LPC modelling to be applied only to certain selected frequency bands.

In Figure 6-8 the spectrum of a segment of noisy speech is shown using two LPC models. The solid curve, corresponds to an eighth-order LPC model in the 0 to 4000-Hz band, and the dashed curve represents an eighth-order LPC model in the 300-3300-Hz band. It can be observed that the solid curve follows the attenuation of the spectrum at high and low frequencies.

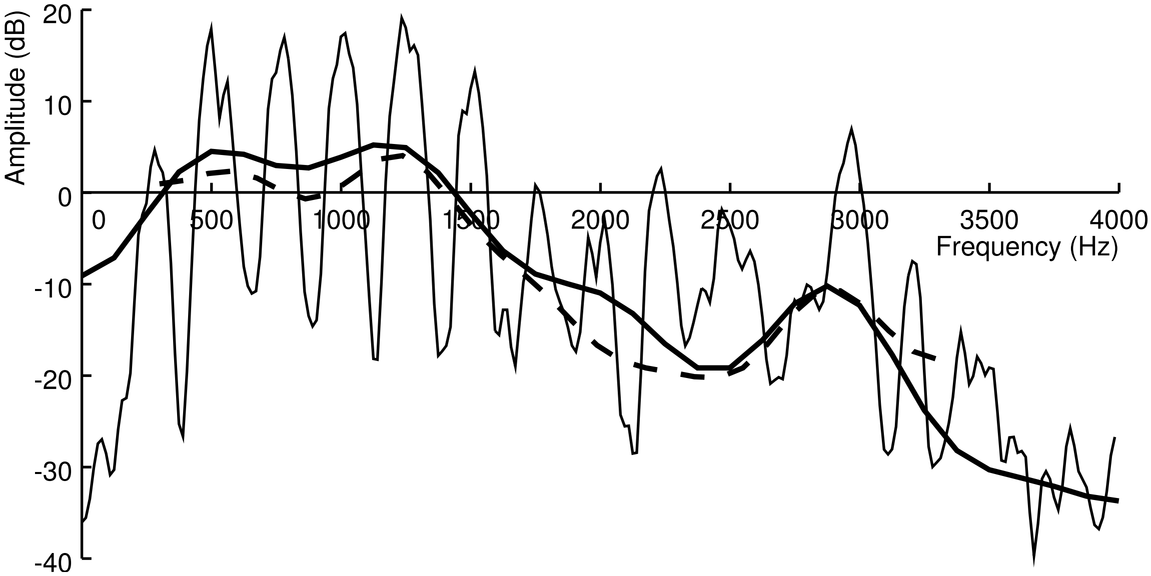


Figure 6-8: FFT Spectrum of a noisy speech frame, comparison of normal LPC (solid line) versus Selective LPC in the 300-3300-Hz band (dashed line).

We examined the use of selective LPC in telephone speech to deal with the problem of low frequency tones and proper modelling of the speech spectrum. In the next table the results of running the SPHINX system with this modified front end are shown.

TRAIN	TEST	WARPING	LPC	FREQ. RANGE	RECOG. ACCURACY
NTIMIT	NTIMIT	0.6	8	0-4000 Hz	42.7%
NTIMIT	NTIMIT	0.3	8	300-3000 Hz	42.7%
TIMIT	NTIMIT	0.6	8	0-4000 Hz	27.4%
TIMIT	NTIMIT	0.3	8	300-3300 Hz	31.3%

Table 7: selective LPC analysis (300-3400 Hz) vs. normal LPC analysis (0-4000 Hz)

As we can see, the use of the selective LPC technique provides no net improvement, except for cases in which training and testing conditions are different.

6.4.7 Modified Warping

The use of a warped frequency scale as opposed to a linear one has proven advantageous in speech recognition systems. Most researchers agree that frequencies above 4000 Hz contribute much less to speech intelligibility and recognition accuracy than frequencies below 4000 Hz. Further evidence for the use of a nonlinear frequency axis can be extracted from the behavior of the human auditory system. Zwicker [1961] introduced the Bark scale[5] as an approximation to the discrimination power of different frequencies in the human auditory system.The net effect of this warping is to compress the high frequency spectrum, reducing the contribution of this part of the spectrum to the recognition engine.

In the standard SPHINX front end designed for speech sampled at 16 kHz, a bilinear transform is applied to the cepstral coefficients to provide frequency warping along the frequency axis, approximating the nonlinear frequency resolution of the Bark scale. A warping coefficient of 0.6 seemed to provide the best approximation to the Bark scale.

It was believed that the warping coefficient should be reduced for speech sampled at 8 kHz so that important frequencies below 4 kHz would not be over-compressed. In order to verify this idea a set of experiments was performed, using a downsampled version of the AN4 database with different values of the warping coefficient. The results of these experiments are shown in Figure 6-9. For downsampled speech, best results are obtained using a warping coefficient of about 0.4. Additional experiments confirm that this value of 0.3 or 0.4 for the frequency-warping parameter is best for the NTIMIT database as well when the selective LPC approach is used. These experiments were run with an 8th order LPC analysis.

6.5 Environmental Compensation Algorithms

The two main sources of variation in the telephone channel are additive noise and a changing channel frequency response from telephone line to telephone line. We investigated the extent to which two standard compensation algorithms, CDCN and RASTA, are able to cope with these types of environmental degradation in telephone channels. CDCN compensates simultaneously for the effects of linear filtering and additive noise, while the much simpler RASTA method compensates primarily for differences in the frequency response of the channel.

[5] The Bark scale is approximately linear for frequencies below 1000-Hz and logarithmic above that level

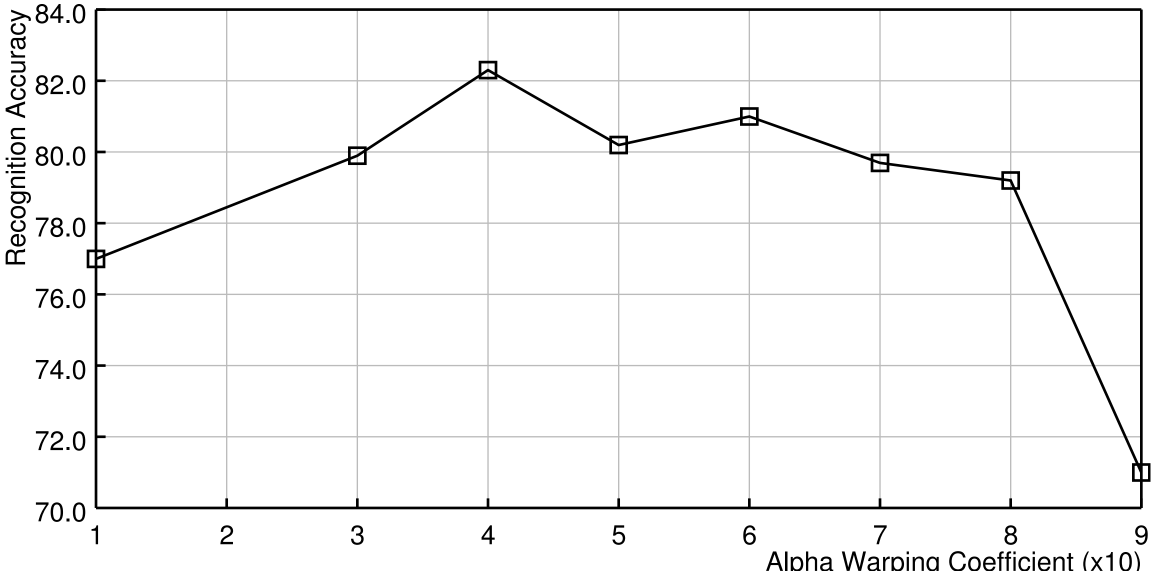


Figure 6-9: Warping Coefficient versus Recognition Accuracy in the AN4 database (d0wnsampled).

6.5.1 CDCN

As noted above, the CDCN procedure is attractive because it can simultaneously compensate for the effects of additive noise and linear filtering. We performed a series of experiments to determine the extent to which CDCN could provide benefit for recognition of speech over a telephone channel. Two conditions were examined: training and testing with the NTIMIT database while applying CDCN, and training on TIMIT and testing on NTIMIT, also applying CDCN.

In a second series of experiments, the system was trained with downsampled clean speech, selective LPC of order eight, a frequency-warping coefficient of 0.3, and tested with noisy speech. We compared the performance obtained with and without the use of CDCN.

TRAIN	TEST	CDCN	RECOGNITION ACCURACY
TIMIT	TIMIT	NO	48.9%
TIMIT	TIMIT	YES	47.7%
NTIMIT	NTIMIT	NO	42.7%
NTIMIT	NTIMIT	YES	42.1%
TIMIT	NTIMIT	NO	31.3%
TIMIT	NTIMIT	YES	39.0%

Table 8: Crossed Results

When the signal-to-noise ratio is very low (below 5 dB), compensation algorithms have a very important effect, improving recognition accuracy over the case of no compensation. When the signal-to-noise ratio is moderate (above 15 dB) the best way to get good results is to train the recognition system on noisy speech.

In this case, we can get up to 42.7% of recognition accuracy. If the system is trained on clean speech and tested on noisy speech there is a degradation of about 11.4%. CDCN is able to reduce this degradation to only 3.7%.

When we compare the performance of CDCN in this NTIMIT task with the performance obtained in the AN4 task (Acero [1990]), CDCN seems to produce slightly worse results[6], perhaps because the algorithm is not designed to deal with the non-stationary noise that could appear in the noisy test set and which may not have been observed during training in the crossed conditions. Another possible explanation could be in the nonlinearities present in the network.

6.5.2 RASTA

We also performed a series of experiments using the RASTA technique. The RASTA technique can be applied to the reduction of additive noise, or to the reduction of linear filtering effects of the channel. In the following experiments the RASTA technique was applied in the cepstral domain, dealing then with linear channel distortion.

The RASTA technique was applied to the TIMIT and NTIMIT databases using the selective 8th-order LPC technique, downsampled speech, and a bilinear warping coefficient of 0.3. The results are given in Table 9. The RASTA filtering was not applied to the *c[0]* coefficient (the average energy in the frame), since this coefficient contains information about the low-frequency content of speech which would be removed by RASTA filtering. In fact in previous experiments we have observed that the use of the RASTA algorithm on the *c[0]* cepstral component reduces the recognition accuracy significantly.

TRAIN	TEST	RASTA	RECOGNITION ACCURACY
TIMIT	TIMIT	NO	48.7%
TIMIT	TIMIT	YES	44.9%
NTIMIT	NTIMIT	NO	42.7%
NTIMIT	NTIMIT	YES	39.2%
TIMIT	NTIMIT	NO	31.3%
TIMIT	NTIMIT	YES	33.2%

Table 9: Results of using RASTA in the TIMIT/NTIMIT task

In the case of training on the closetalk speech and testing using the telephone channel, RASTA is able to improve recognition accuracy from 31.3% to 33.2%, less than 2%. Perhaps the reason why the improve-

[6] In previous experiments CDCN was applied to a noisy version of the AN4 database. Therefore when comparing the benefit of using CDCN in the crossed conditions (train on clean AN4, test on the noisy AN4), the use of CDCN improved recognition accuracy significantly over not using it.

ment is no greater lies in the fact that RASTA is operating to remove the linear channel distortion. In the case of telephone speech additive noise could be the dominant problem.

6.6 New Spectral Representations: Mel Frequency Scaled Cepstrum

Another well-known front end is the Mel Frequency Scaled Cepstrum (MFCC), defined as the DFT of the logarithm of the power spectral density function interpolated over a warped frequency scale (the Bark scale). The main difference between MFCC and the current SPHINX front end is that the power spectral density function is estimated directly using the DFT of the windowed speech, rather than describing the speech by an AR process.

MFCC appears to have two advantages over the conventional LPC approach. The first advantage is that the covariance matrix of MFCC coefficients appears have smaller values outside the main diagonal (and most speech recognition systems assume that the off-diagonal elements are zero). Second, the estimation of cepstral coefficients using MFCC seems more robust to the presence of noise compared to the LPC approach.

Table 10 provides a summary of experimental results comparing the use of the Mel Frequency Scaled Cepstrum as a front end with the LPC Frequency Cepstrum (LPCC).

TRAIN	TEST	FRONT END	LPC ORDER	BANDWIDTH	RECOG. ACCURACY
TIMIT	TIMIT	MFCC	N.A.	8000 Hz	53.1%
TIMIT	TIMIT	LPCC	14	8000 Hz	52.7%
NTIMIT	NTIMIT	MFCC	N.A.	4000 Hz	43.9%
NTIMIT	NTIMIT	LPCC	8	300-3300 Hz	42.7%
TIMIT	NTIMIT	MFCC	N.A.	4000 Hz	29.8%
TIMIT	NTIMIT	LPCC	8	300-3300 Hz	31.3%

Table 10: Comparison of LPCC vs. MFCC

We selected the front end parameters (LPC order, alpha warping coefficient, number of channels in the MFCC front end,...etc) of these experiments so that for a given method (LPCC or MFCC) and bandwidth (4000 Hz or 8000 Hz) we obtained the best performance.

In general MFCC provides an improvement of less than one percent under clean conditions, and slightly more than one percent on noisy conditions. In the case of crossed conditions there is no improvement, however there is still room for improvements of the MFCC modelling, since at this stage it has not been fully adjusted to the reduced bandwidth conditions.

6.7 Nonlinearities

An additional problem with telephone speech is the possibility of nonlinearities introduced by carbon-button microphones in the handset and by the network itself. We performed an initial series of experiments in an attempt to develop a simple scheme to compensate for the effects of these nonlinearities. In these experiments we assume that the nonlinearities in the net are memoryless*, i.e.* they can be modelled as an operation that does not take into account previous time values, such is the case of quantizers and amplifiers working in a nonlinear region. A further assumption we make is that the nonlinearities are soft, *i.e*. they don't have sharp transitions.

A set of experiments was designed to see if under these assumptions the nonlinearities could be reversed. Since we have access to the input of the telephone channel (the TIMIT speech) and its output (the NTIMIT speech), assuming statistical independence of one sample from another, we calculated a histogram of the signal that approximates the probability density function (PDF) for the input and output signals. Once these two histograms are estimated, a nonlinear warping function can be found to match them minimizing the difference between the two.

In Figure 6-10 we show histogram plots for a sentence from the TIMIT and NTIMIT databases. In Figure 6-11 we show the nonlinear mapping that minimizes the distance between the two histograms.

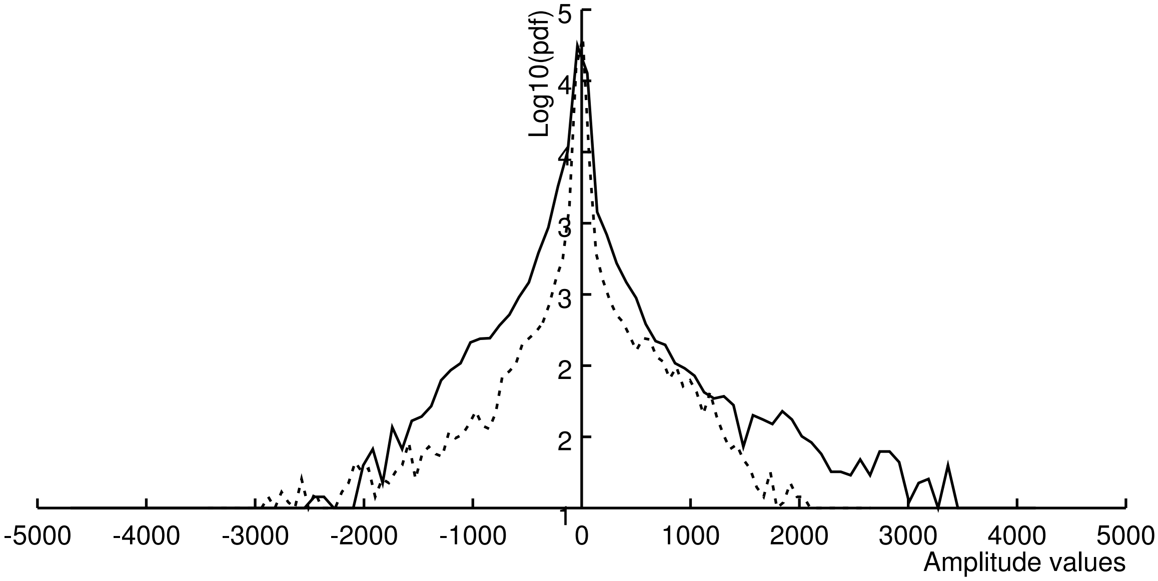


Figure 6-10: Amplitude histograms for one sentence. The solid curve represents clean speech, and the dotted one represents noisy telephone speech.

Using the clean speech as a reference, for every noisy sentence we were able to find the mapping between histograms that minimized the distortion. These mappings were used to reverse the nonlinearities in the noisy testing set. We then measured the recognition accuracy obtained using this testing set with compensation for the nonlinearities. The system was trained using the TIMIT database and tested using the NTIMIT database.

The results, shown in the next table, indicate that this technique produces a degradation in recognition accuracy rather than an improvement in performance. We believe that the problem with this scheme is that two other effects found in the telephone network, additive noise and linear filtering, are not considered.

TRAIN	TEST	LPC	NONLINEARITY	RECOG. ACCURACY
TIMIT	NTIMIT	8	NOT REVERSED	31.3%
TIMIT	NTIMIT	8	REVERSED	29.9%

Table 11: Comparison of reversed vs. not reversed nonlinearity in the crossed conditions

In general if *h(m)* is the telephone channel, *n(m)* the additive noise and *T[]* is a nonlinear function that can be considered memoryless, the output *y(m)* will have the form:

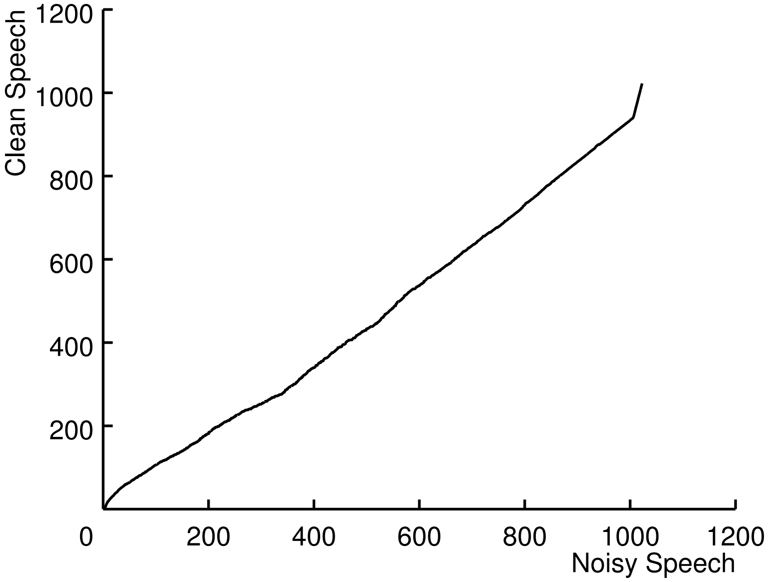


Figure 6-11: A dynamic warping curve minimizing the distance between a clean and noisy speech amplitude histogram.

$$y(m) = h(m) * T[x(m)] + n(m)$$

$$h(m) = \sum_{k=-\infty}^{\infty} h(k)\,T[x(m-k)] + n(m)$$

The combination of these three effects results in a highly nonlinear and difficult problem. In fact we would like to be able to estimate simultaneously these three parameters, *T[], h(m)* and *n(m)* in a similar way that the CDCN algorithm estimates simultaneously the linear filtering of the channel and the additive noise.

6.8 Summary

In this chapter we have explored the effect of the telephone network on speech recognition systems, and we have observed a 10% degradation in performance when comparing telephone speech to clean speech. A number of reasons for this degradation have been explored, from bandwidth reduction to inappropriate front end parameters (LPC order, analysis window size, frequency-warping coefficient, and Pascal window width). The effect of low-frequency tones in the LPC analysis has also been explored, and a technique to deal with this problem, selective LPC, has been proposed. A general fine tuning of the SPHINX front end has produced an improvement of about 3% when compared with the standard SPHINX front end.

Two environmental compensation techniques have been considered, RASTA and CDCN. CDCN has proved to be much more effective in the case of training with clean speech and testing using speech over the telephone network, although its performance over the telephone network is worse than in other noisy environments.

Finally, the use of spectral representations using Mel-frequency cepstral coefficients has been considered, yielding similar results to what was obtained using the more traditional LPCC front end.

In conclusion we have shown that the general degradation of 10% introduced by the telephone network does not appear to be produced by downsampling or by inappropriate parameters used in the LPC model. In fact when a complete new front end based on FFT analysis is introduced this 10% degradation remains. Furthermore, our experiments using the CDCN algorithm suggest that the degradation in recognition accuracy cannot be accounted for by simple additive noise and linear filtering in the telephone channel; our results suggest that other sources of degradation are present as well.

Chapter 7
Conclusion and Future Work

In Section 7.1 we summarize the major findings of this report, and in Section 7.2 we offer suggestions for future work.

7.1 Summary of Findings

In this report we have performed several experiments to explore the effects that the telephone network has on the recognition accuracy of a speech recognition system. In general we have observed a degradation of about 10% in recognition accuracy. A degradation of 20% was observed when the system is trained on clean speech and tested on telephone-quality speech.

We studied several hypothesis for the reduction in performance in the NTIMIT task, including bandwidth reduction introduced by the telephone channel and inappropriate choice of front end parameters. Of the parameters considered (LPC order, window size, pascal window width and bilinear alpha coefficient) only the bilinear alpha coefficient was found to affect recognition accuracy. Bandwidth reduction was also affected the recognition accuracy. Fine adjustment of the LPC parameters did not improved performance, except for some conditions in which the system was trained on clean speech and tested using telephone speech.

In the following three tables a summary of major results is presented.

TRAIN	TEST	ANALYSIS	FREQ. RANGE	COMPENSATION	RECOG. ACCURACY
TIMIT	TIMIT	LPCC	0-8000 Hz	NONE	52.7%
TIMIT	TIMIT	LPCC	300-3300 Hz	NONE	48.9%
TIMIT	TIMIT	MFCC	0-8000 Hz	NONE	53.1%
TIMIT	TIMIT	LPCC	300-3300 Hz	RASTA	44.9%
TIMIT	TIMIT	LPCC	300-3300 Hz	CDCN	47.7%

Table 12: Summary of results for the TIMIT task

TRAIN	TEST	ANALYSIS	FREQ. RANGE	COMPENSATION	RECOG. ACCURACY
NTIMIT	NTIMIT	LPCC	0-8000 Hz	NONE	41.3%
NTIMIT	NTIMIT	LPCC	300-3300 Hz	NONE	42.7%
NTIMIT	NTIMIT	MFCC	0-4000 Hz	NONE	43.9%
NTIMIT	NTIMIT	LPCC	300-3300 Hz	RASTA	39.2%
NTIMIT	NTIMIT	LPCC	300-3300 Hz	CDCN	42.0%

Table 13: Summary of results for the NTIMIT task

TRAIN	TEST	ANALYSIS	FREQ. RANGE	COMPENSATION	RECOG. ACCURACY
TIMIT	NTIMIT	LPCC	0-8000 Hz	NONE	23.9%
TIMIT	NTIMIT	LPCC	300-3300 Hz	NONE	31.3%
TIMIT	NTIMIT	MFCC	0-4000 Hz	NONE	29.8%
TIMIT	NTIMIT	LPCC	300-3300 Hz	RASTA	33.2%
TIMIT	NTIMIT	LPCC	300-3300 Hz	CDCN	39.0%

Table 14: Summary of results for the crossed conditions task

We were disappointed to find that environmental-compensation algorithms such as RASTA and CDCN which have provided improvements in other domains in which different conditions were used for training and testing did not by themselves provide much help in dealing with the special problems of telephone speech. Specifically, we found that when the system was trained and tested using the NTIMIT database neither CDCN nor RASTA provided better performance than the performance obtained with the baseline system. Similarly, the use of CDCN did not completely compensate for loss of recognition accuracy incurred by training with TIMIT speech and testing using the NTIMIT database. (This is indicated by the fact that TIMIT/NTIMIT performance using CDCN is worse than NTIMIT/NTIMIT performance with no processing.) This is the first instance in which CDCN was unable to provide complete compensation for different training and testing conditions. Hence it appears that there are important sources of degradation in telephone speech beyond the simple effects of quasi-stationary additive noise and linear filtering.

We also considered a compensation algorithm for dealing with nonlinear effects and studied alternative spectral representations.

We believe that the standard front end parameters are reasonable, although some simple adjustments are necessary to account for the reduction in sampling rate. Alternate spectral representations such as MFCC seems to provide slightly better performance.

7.2 Future Work

The results of this study point to three directions of research that are likely to be needed to further improve recognition accuracy of telephone speech.

The first and simplest technique to be adopted is the use of noise-word models, which were not examined in this study. These models have the ability to provide some resilience to the effects of non stationary noises which can be very troublesome in a database as NTIMIT. So far noise-word models have been applied to very simple tasks, with limited vocabularies and with the aid of grammars with great success, but it remains to be seen if these methods can achieve a significant improvement in more complex tasks and for speech over telephone lines.

We also believe that further research in telephone speech recognition would be helped by a more analytical approach to modeling of the telephone channel. A number of physical models have been developed for research in telephone-channel simulation. This knowledge should be exploited to develop compensation procedures that are more specific to the vagaries of long distance telephone lines.

Finally we believe that further improvements in compensation techniques such as CDCN can be obtained if the algorithms are applied at the level of phonetic models in the speech recognition system. So far all of our work has taken place at the waveform level before speech is input to the system. Application of this compensation at the HMM-state level remains an open field of study.

References

Acero A. (1990) Acoustical and Environmental Robustness in Automatic Speech Recognition. Ph. D. Dissertation, ECE Department, Carnegie Mellon University, September 1990.

Carey M. B., Chen H. T., Descloux A, Ingle J. F., and Park K. I. (1984) 1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network, *AT&T Bell Labs Tech. J.* , 63, November 1984, 2059-2119

Chigier B. (1991) Phonetic Classification on Wide-Band and Telephone Quality Speech, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1991.

Ephraim Y., Malah D. and Huang B.D. (1989), Speech Enhancement Based Upon Hidden Markov Modelling. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Glasgow, UK, May 1989.

Gales M. J. F. and Young S. (1992), An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1992.

Gillick L. and Cox S.J. (1989), Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* Glasgow, UK, May 1989.

Gray R., Buzo A., Gray A. and Matusyama Y. (1976), Distance Measures for Speech Processing, *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, October 1976.

Hermansky H., Morgan N., Bayya A., and Kohn P. (1991) Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP), *Proc. EUROSPEECH 91*, pp 1367-1370, Genova, 1991.

Hirsh G., Meyer P. and Ruehl H. (1991), Improved Speech Recognition Using High-pass Filtering of Subband Envelopes, *Proc. EUROSPEECH 91*, pp 413-416, Genova, 1991.

Jankowski C., A. Kalyanswamy, Basson S., and Spitz J.(1990).NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database,*International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 1990

Lee K. F. (1989) Automatic Speech Recognition: The Development of the SPHINX Recognition System, Kluwer Academic Publishers, 1989.

Lee K.F. and Hon H. W. (1988) Speaker-Independent Phone Recognition Using Hidden Markov Models, CMU-CS-88-121, Carnegie Mellon University, Computer Science Research Report.

Lim J.S. (1983), Speech Enhancement, Prentice Hall, Englewood Cliffs, NJ, 1983.

Lim J.S. (1979), Enhancement and Bandwidth Compression of Noisy Speech, *Proce. of the IEEE* , vol 67, Dec 1979.

Makhoul J. (1975), Linear Prediction: A Tutorial Review, *Proce. of the IEEE* .Vol 63, 1975, 561-580.

Makhoul J., and Wolf J. (1970), Spectral Linear Prediction: Properties and Applications, *IEEE Trans. on Acoustics, Speech, and Signal Proc.*Vol ASSP-23, June 1975.

Rabiner R. L. and Juang B.H. (1986), An Introduction to Hidden Markov Models, *IEEE ASSP Magazine* 3, January 1986.

Rabiner R.L. and Schafer R.W. (1978), Digital Processing of Speech Signals, Prentice-Hall

Seneff S. (1988), A Joint Synchrony/Mean-Rate Model for Auditory Speech Processing, *Journal of Phonetics* 16:55-76, 1988.

Stockham, T.G., Cannon T. M. and Ingebretsen R.B. (1975) Blind Deconvolution Through Digital Signal Processing, *Proce. of the IEEE* 63 (4), April 1975, 678-692.

Ward W. (1989), Modelling Non-Verbal Sounds for Speech Recognition, 1*992 DARPA Workshop on Speech Recognition*, pages 310-318, October 1989.

Wilpon J., Rabiner L. R., Lee C. and Goldman E. R. (1990), Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models, *IEEE Trans. on Acoustics, Speech, and Signal Proc.* November 1990.

Zue V. *et al*, (1990), The Summit Speech Recognition System: Phonological Modelling and Lexical Access, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Alburquerque, NM, pages 49-52, April, 1990.

Zwicker E. (1961) Subdivision of the Audible Frequency Range into Frequenzgruppen,*J. Acoust. Soc. Am.* 33, 248,February 1961.