AN INVESTIGATION OF THE USE OF AN AUDITORY MODEL IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

by

Claude Lefèbvre

A thesis submitted to the Department of Electrical Engineering at the University of Ottawa in partial fulfillment of the requirements for the degree of Master of Applied Science

ISBN 0-315-33347-2

The University of Ottawa requires the signatures of all persons using or photo-copying this thesis. Please sign below. and give your address ans date.

ACKNOWLEDGEMENTS

Sincere thanks go to my advisor Dr. Nicholas Georganas, who first directed me to the field of speech analysis. I am grateful to him for the time he spent correcting the thesis and pushing me to improve my communication skills.

Special thanks are due to my co-supervisor Dr. Melvyn Hunt, for the time he spent guiding me in the research and in writing my thesis. His support and patience has permitted me to improve my knowledge in the field of speech research.

I thank the National Research Council of Canada, in particular the head of the Aeroacoustics Section Mr. Bob Westley, who permitted me to use Speech Processing facilities in the Laboratory. I also thank Carl Swail, Michael Baranowski and Claes Harvenberg for providing speech spectrograms, plots and comments.

I would also like to acknowledge the financial support from the NSERC Operating Research Grant A-8450 of Dr. N. D. Georganas and the financial support provided by the Department of National Defense through DCIEM for partially funding the computing facilities I have used in the Aeroacoustics Section of the National Research Council.

PREFACE

In recent years, intensive research has been carried out on modeling the auditory system. It is partly motivated by the need to increase the quality of acoustic analysis in speech recognition systems and by the hope of gaining a better understanding of the auditory system. It is believed that a computational model of the auditory system will be more robust than conventional front-end analyses and will reproduce salient features in the speech signal.

The purpose of this work is to implement a computational auditory model and evaluate it as the front-end analysis of a template-matching speech recognition system. The auditory model used is primarily based on the computational model presented by S. Seneff at the 1984 IEEE ICASSP meeting [53]. A feature of this model is its ability to extract important perceptual cues in speech signals.

Since the field of speech recognition is very broad, we first present various topics pertinent to the field. This is intended to provide the reader with a review of the topics that are important for modeling and testing the front-end of a speech recognition system.

The first chapter gives an introduction to how speech sounds are produced and what the perceptual cues are that permit us to differentiate speech sounds. Since we measure the recognition performance resulting from the use of the auditory model as the front-end of a template-based recognition system, a description

of such recognition system is also given. In the last part of this first chapter, we present some of the reasons why the performance obtained with current speech recognition systems is far inferior to human speech perception. In this discussion, we also compare speech representations obtained from the two most common front-end analyses, namely linear predictive and filter-bank analysis.

Chapter 2 gives a description of the human auditory system. Some important details of the known mechanisms of the ear and its properties are given. Then we present Seneff's auditory model, which is based on some of the ear's known mechanisms and properties. In contrast to other computational auditory models recently presented [20][37][52], this model employs in the final stage process a synchrony measure which is based on the observation that each auditory nerve fiber responds synchronously with the corresponding stimulus.

Chapter 3 describes how the auditory model has been modified to be used in a template-matching recognition task. Small-scale recognition experiments have shown that representations obtained from Seneff's computational auditory model do not give good recognition performance. The reason appears to be that the model produces too much frequency masking. Because we consider that the human ear would derive a close to optimal representation of speech sounds, we set the parameters of the new computational auditory model to replicate human frequency masking.

In Chapter 4 the recognition performance of the auditory model representation is compared with that of a commonly used filter bank representation in a cross-speaker digit-recognition task. The conventional dynamic time-warping (*DTW*) algorithm (which is described in Chapter 1) is used in the digit recognition

test for both representations. Various speech conditions are considered, namely, clean speech, speech in the presence of noise and linearly distorted speech. Finally, chapter 5 presents some conclusions about the modified version of Seneff's auditory model as a front-end analyzer of a recognition system and some suggestions are given for future research.

TABLE OF CONTENTS

CHAPTER 3 - DETAILS OF THE IMPLEMENTATION OF THE AUDITORY MODEL

CHAPTER 4 - SPECTROGRAPHIC COMPARISONS AND SPEECH RECOGNITION EXPERIMENTS

CHAPTER 5 - CONCLUSION

APPENDIX

CHAPTER 1

SPEECH ANALYSIS FOR RECOGNITION PURPOSES

1.1 INTRODUCTION

In the past decade the science of speech recognition has advanced to the state where it is now possible to communicate with a computer by speaking to it in a disciplined manner using a vocabulary of moderate size. Although we have made a lot of improvements in this field, the performance of speech recognition systems is still much inferior to human speech perception. This is partly due to the fact that present-day speech recognition systems do not extract appropriately the relevant perceptual cues contained in speech sounds and partly because they are very sensitive to variations between speakers and to noisy environments.

The purpose of this chapter is first to review what are the most important perceptual cues used to differentiate speech sounds. Then we describe how a conventional word recognition system extracts these cues and how it compares them. Finally we compare the two most frequently used front-ends in a speech recognition system and we make some comments on their performance in extracting perceptual cues from noisy speech and from speech produced by different speakers.

1.2 PRODUCTION OF SPEECH SOUNDS

Speech sounds are produced by the flow of air coming from the lungs and passing through the vocal tract. This can be represented by a speech production model consisting of a source exciting a vocal tract filter. In particular, there are two different modes of excitation, and this permits us to divide speech sounds into two distinct classes: *voiced* and *unvoiced* sounds.

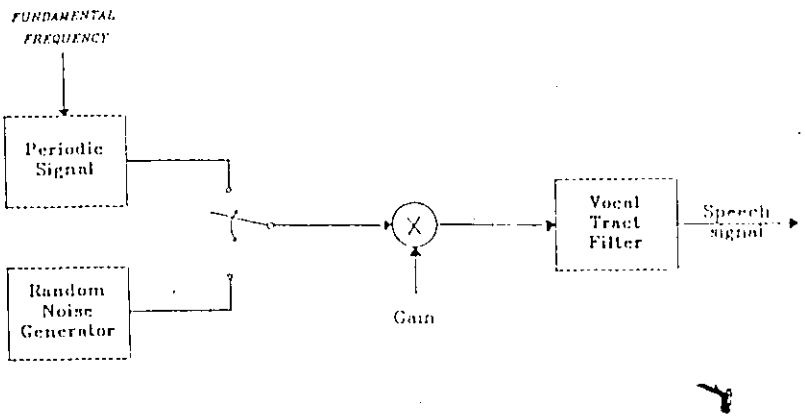

Figure 1.1 : *Model of speech production.*

Voiced sounds are produced when the excitation consists of quasi-periodic pulses of air generated by the opening and closure of the vocal cords in the larynx. When the vocal cords are not vibrating, but there is turbulence created by the air passing through constrictions of the vocal tract, the speech sounds produced are called unvoiced or *voiceless*. In the speech production model, unvoiced excitation is represented by random noise exciting the vocal tract filter.

Figure 1.2 shows the power spectrum of a time-differenced voiced sound. The spectrum can be seen to be made up of a sequence of equally spaced spikes. The first spike corresponds to the frequency of closure of the vocal cords, known as the *fundamental frequency*. The succeeding spikes correspond to *harmonics* of the fundamental and occur at all integer multiples of the fundamental. The spacing between them is thus equal to the fundamental.

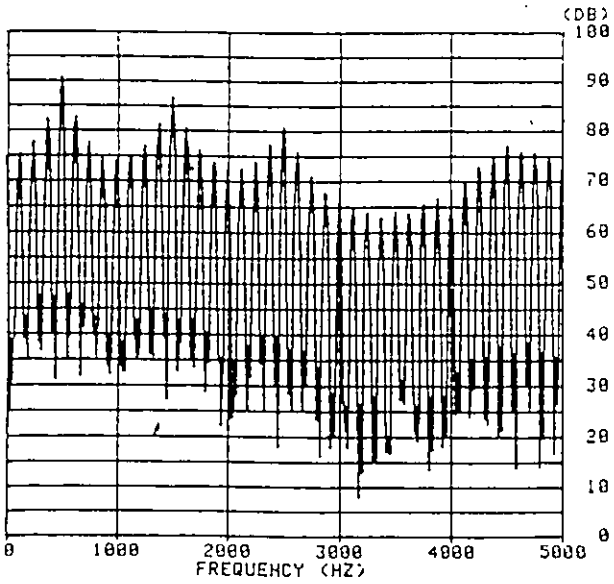

Figure 1.2 : The power spectrum of a time−differenced vowel sound. (after Hunt [27]).

We also notice in the figure that the intensity of the harmonics varies across the spectrum. It depends partly of the excitation, partly on the effect of radiation from the lips and partly on the impulse response of the vocal tract. The combined effect of the first two factors is approximately to impart a -6 *dB/octave* slope to the spectrum. In figure 1.2, this slope has been removed by the time-differencing, and the remaining variation in the intensity of the harmonics is due

to the vocal tract. Peaks in the envelope of the spectrum corresponds to *resonances* of the vocal tract.

The resonant structure of the vocal tract depends on the position of the tongue, jaws and lips. For different resonant structures different kinds of voiced and unvoiced sounds are produced. Some examples of voiced sounds are pure vowels as in: *see*, *two*, diphthongs as in: *toy*, *cow*, *hay*, and some consonants e.g. : *b*,*d*,*g*,*v*,*z*, etc. Examples of unvoiced sounds are : *f*,*sh*,*p*,*t*,*k*.

Some other mechanisms of the vocal tract occurring when voiced and unvoiced sounds are produced give rise to sounds that have *plosive* or *nasal* characteristics. Plosive sounds result from making a complete closure of the mouth, building up the pressure behind the closure, and abruptly releasing it as in: *p*,*t*,*k*,*b*,*d*,*g*. Nasal sounds (*n*,*m*,*ng*) occur when the nasal tract is connected and the airflow going through the mouth is stopped off at some point. A branched tube of this kind has *antiresonances* and the speech waveform produced by exciting it has prominent holes in the envelope of its power spectrum.

1.3 INFORMATION CONTAINED IN SPECTRAL REPRESENTATIONS OF SPEECH SIGNALS

Words and sentences are produced by varying in time the excitation going into different vocal tract configurations. To analyze the time-varying characteristics that produce speech signals, researchers and phoneticians have used a spectrographic representation. It consists of representing short-time power spectra of

speech sounds in a two-dimensional pattern called a *spectrogram* such as in figure 1.3, where the vertical axis corresponds to frequency and the horizontal axis corresponds to time. A similar time and frequency representation of sounds can be observed in the human ear.

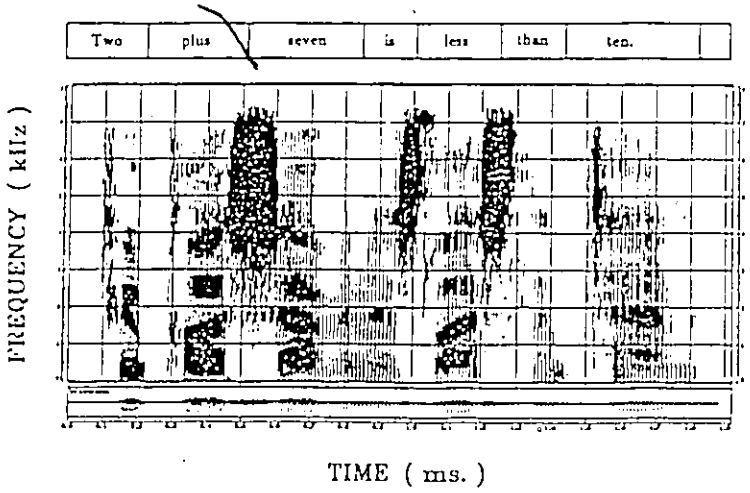

Figure 1.3 : *Speech spectrogram of the sentence : Two plus seven is less than ten (after Zue* [58])

The most obvious feature of the spectrogram shown in figure 1.3 is the vertical stripes. The separation between them corresponds to one complete cycle of opening and closure of the vocal cords. Inside some of the vertical stripes, we notice energy bands concentrated around four or five frequencies. They correspond to resonances of the acoustically excited vocal tract and are called *formants*. We would expect to see formants parallel to the time axis for pure vowels, moving slowly in frequency for diphthongs and more rapidly for *glides* (r, y), and *liquids* (w, l), etc.

The second class of sounds, known as unvoiced, can also be seen in the spectrogram of figure 1.3. Segments composed of them consist of energy grouped in the higher frequency part of the spectrogram (~ 2-7 kHz) rather than in the lower part. As with voiced sound segments, unvoiced sounds have formants but their peaks are less distinct because they have broader bandwidths.

Another characteristic visible in the spectrogram namely a period of silence followed by a sudden rise in energy, permits us to identify plosive sounds. This can be seen just before the *t* in *two* and *ten* and the *p* in *plus*.

Although many spectral features can characterize speech sounds, we must take into consideration that only some of them have important perceptual meanings. The most useful of them are the formant frequencies and changes in frequencies [34] since we observe these characteristics in almost every speech sound and they are quite specific to each sound. On the other hand, it has been demonstrated that formant amplitudes and bandwidths have little effect on the phonetic content [24][34]. A feature corresponding to energy rising rapidly or not in some frequency bands is also a relevant perceptual cue, since it permits us to differentiate, for example, voiced plosive sounds from voiced fricative sounds. As we might expect, whether the energy is concentrated in the lower part of the spectrum or not is also important, since it permits us to differentiate voiced from unvoiced sounds.

The important conclusion that we may draw about the most relevant features contained in a speech waveform is that they are present in different frequency bands and that they can therefore be extracted from a speech signal by particular kinds of spectral estimation. As we shall see at the end of this

chapter, the performance of a speech recognition system depends on how well they are represented in the spectral estimation and how they are affected by noise and by differences between speakers.

1.4 FUNCTIONS WITHIN A WORD RECOGNITION SYSTEM

At present, small word-based speech recognition systems have two or three main applications. They can be used in a speaker-dependent recognition mode recognizing 50 to 100 words for controlling various tasks. They can also be used in a speaker-independent mode with a vocabulary of approximately ten or fifteen words, such as the ten digits. For these applications the sophistication of the recognition algorithms used depends to some extent on whether if they are intended for connected and isolated word recognition tasks.

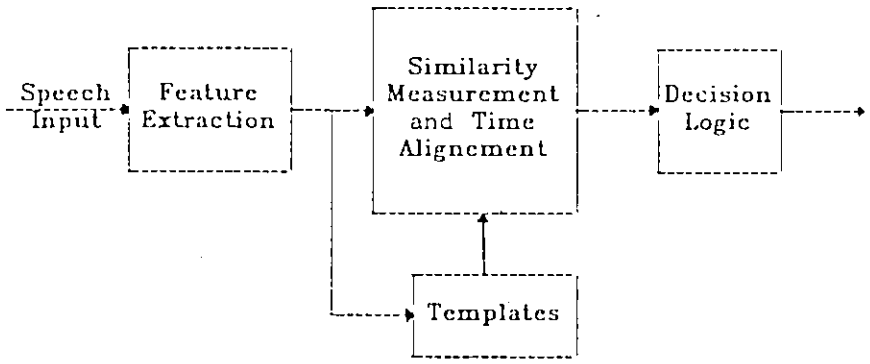

Figure 1.4 : Block diagram representing functions accomplished in word recognition systems

A word recognition system can be represented by the block diagram of functions illustrated in figure 1.4. To use the system, one or more speakers must first train the system by providing examples of the words in the vocabulary to be recognized. Reference templates are then created for the words. These templates are later used to match the incoming spoken word.

The process of recognizing a word can be seen as an extraction of its features - somewhat as speech sounds are represented in a spectrogram - and then finding the best match between the representation of word features and those of the words in the reference set. For the second task, very efficient *dynamic − programming* algorithms [7] or *Hidden − Markov modeling* [44] have been used both for connected or isolated word recognition.

Since in this work we are comparing the performance obtained with two feature-extraction processes and we are interested in relative and not absolute performance levels, we use a simple template-matching recognition system. The major functions in such a system are the feature extraction process and a time alignment algorithm of the features. Both of these functions are described in the following paragraphs.

1.4.1 Feature extraction process

The task of feature extraction in a speech recognition system is to generate short-time spectral representations. Speech recognizers most commonly use a filter-bank analysis to extract the energy contained in the different frequency bands where most speech information is contained (0-5 kHz). Other feature measurement methods have been proposed. One of them, which is closely related to filter-bank analysis, consists of analyzing the zero-crossing rate (periodicity) in each frequency band. Some other sophisticated methods use representations such as the short-time discrete Fourier Transform (DFT) spectrum or linear-predictive coding (*LPC*). Each method has certain advantages, but due to their ease of implementation and their recognition performance, LPC and filter-bank analyses are the most frequently used methods. In the following paragraphs, an overview of the recognition functions used in a filter-bank analysis is presented. The reason for choosing a filter-bank analysis is that it is closest to the computational auditory model that is presented later in this work. It is also generally considered to be at least as effective as any of the commonly used alternatives.

Figure 1.5 shows a typical system using filter-bank analysis [11][43]. In this system the speech signal is passed through a number of parallel bandpass (BP) filters. The number (K) of filters can vary from four to a hundred and the filter spacing is generally linear below 1000 Hz and logarithmic above 1000 Hz, the so-called *technical mel scale*.

The output of each bandpass filter is passed through a rectifier (NL) and a low-pass filter (LP) to give a signal that is related to the energy of the speech signal in that band. The output is generally expressed as log energy (LOG). This

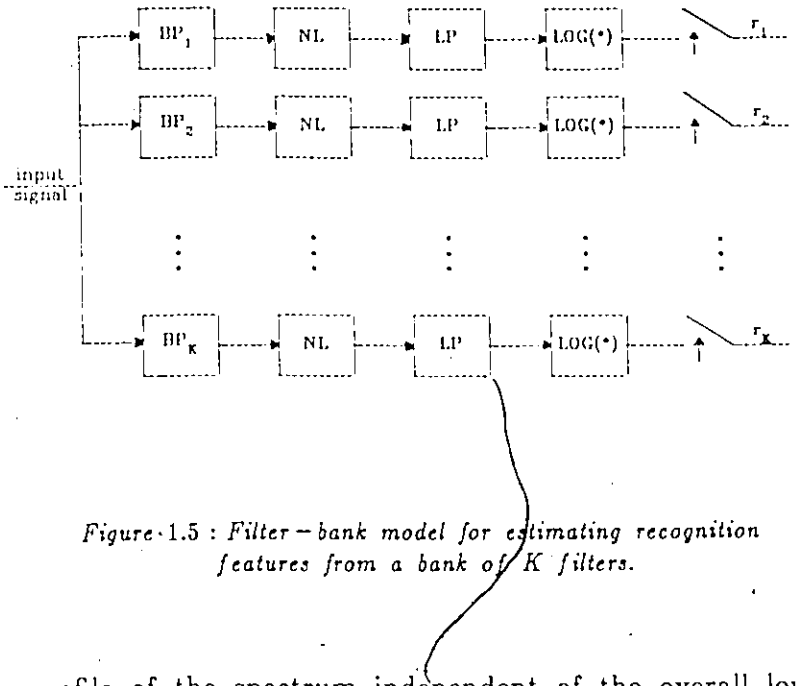

Figure 1.5 : Filter – bank model for estimating recognition features from a bank of K filters.

makes the profile of the spectrum independent of the overall loudness of the input and corresponds better to a perceptual loudness. Finally the output of each filter is sampled. The sampling rate is chosen to be between 5 to 20 ms. [12]. The m'th spectral estimation frame is represented by :

$$R_j(m) = \left\{ r_1(m), r_2(m), \ldots\ldots, r_K(m) \right\} \tag{1.1}$$

where the total spectral pattern goes in time from $m = 1$ *to* M, and in frequency from channel 1 to channel K, and r_k is the log-energy in the k'th channel of the m'th frame of word j.

1.4.2 Measuring similarity between spectral patterns.

The next task in a speech recognition system after the feature extraction process consists of comparing the extracted spectral test pattern with the previously recorded spectral reference patterns. This task is based on a measure of similarity between spectral patterns of the same word. That is, since the sequence of speech sounds of the same word is the same, we would expect spectral patterns representing this word to be quite similar, although the time-duration might be slightly different. Spectral reference patterns representing other words would not be expected to correspond to spectral patterns of this word since some of their spectral frames will be quite different.

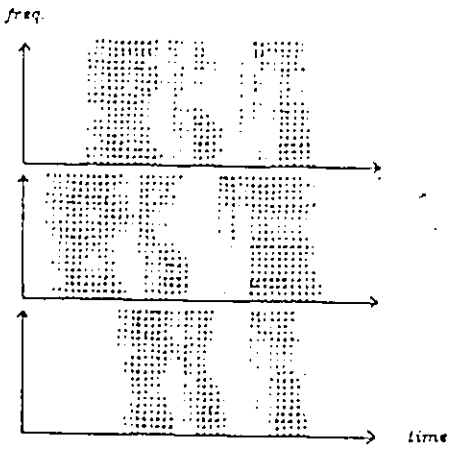

Figure 1.8 : *Three spectrograms of the word helicopter.* (*after Moore* [38])

As suggested by figure 1.6 illustrating three spectral representations of the word *helicopter*, the major problem encountered when measuring similarity between spectral patterns of the same word is how to deal with the duration of

the speech sounds being compared, since the same word is very rarely pronounced at the same rate by different speakers or even by a single speaker. It is obvious that a function measuring the overall similarity between two of these spectral patterns by a trivial sum of the difference between frame one to frame one, frame two to frame two, and so on, would not work.

One possible solution would consist of determining the beginning and end points of the words and linearly stretching or shrinking the spectral frames of one word such that both spectral patterns would have the same length. Then the sum of the frame-to-frame comparisons would give the overall dissimilarity. But this method does not optimize the time-alignment between spectral patterns, since the rate at which a word is pronounced also varies within the word. Rather, what we need is a non-linear time-alignment, as illustrated in figure 1.7, in order that spectral frames in the middle of the words will also be time-aligned.

Now that we have defined one possible way to carry out the comparison of spectral patterns, we need an algorithm to do it. We need a function to calculate the similarity or dissimilarity between frames of the two spectral patterns. For example, in a filter-bank analysis the dissimilarity between the n'th frame of the spectral test pattern ($\mathbf{T}(n)$) and the m'th frame of the j'th spectral reference pattern ($\mathbf{R}_j(m)$) can be calculated with the Euclidean distance measure:

$$d(\mathbf{T}(n),\mathbf{R}_j(m)) = \sum_{k=1}^{K} (t_k(n)-r_k(m))^2 \qquad (1.2)$$

We also need an automatic process to time-align non-linearly spectral patterns and to calculate the overall dissimilarity between them. The basis of this process is to find the warping function, (which is $w(i)$ in figure 1.7), among the

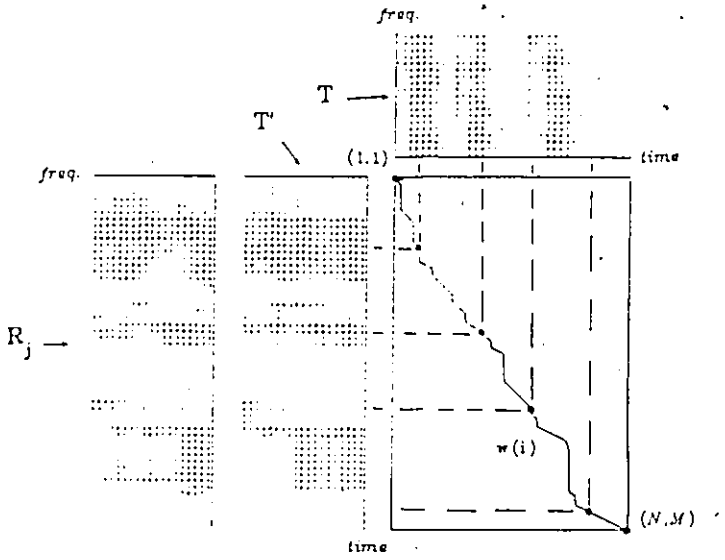

Figure 1.7 : Non−linear time−alignment of two spectrograms.
T and R_j are the spectrograms being aligned.
T' is the non−linear time−normalized version of T.

infinite possibilities $w_c(i)$, $c=1,2,...$ that optimally adjusts the timing difference between spectral frames of T onto the ones of R_j. This can be done by calculating the minimum sum of the frame-to-frame dissimilarity along the points of one of the warping functions :

$$D(T,R_j) = \min_{w_c} [\sum_i d(w_c(i))\, g(i)] \qquad (1.3)$$

where $d(w_c(i))$ is the Euclidean distance at the point i, $g(i)$ is a weighting function, and $D(T,R_j)$ is the overall dissimilarity.

Equation 1.3 can be automatically solved by the use of a dynamic programming *time − warping* technique [5][49]. In such a technique, equation 1.3 is represented by a recursive summation for which a set of local decisions representing the warping function are automatically calculated to minimize the overall dissimilarity.

1.4.3 Dynamic time-warping calculation

The local decisions in the dynamic time-warping calculation can be defined by representing the warping function in a two-dimensional lattice as shown in figure 1.8. The horizontal and vertical axes correspond to the time axis of the two spectral patterns to be compared and each point is associated with a Euclidean distance. Referring to figure 1.7, the goal is to find a warping path through the lattice beginning at (1,1) and ending at (N,M) which is continuous and monotonic and which minimizes the sum of the Euclidean distances over the points it traverses.

Two different basic step formulations (symmetric and asymmetric) [43][50] have been used to represent the possible moves between pairs of points in the warping path in the two-dimensional lattice. These are represented in figure 1.9. By allowing the basic steps shown in figure 1.9 both monotonicity and continuity constraints are respected - i.e. the relative ordering of the points in the warping path is always positive or null and the motion of the warping path is continuous.

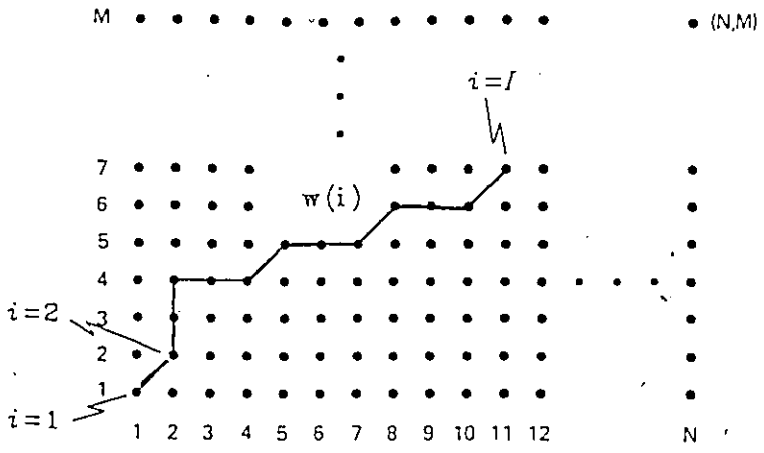

Figure 1.8 : Two − dimensional lattice used to describe the variants of the time − warping algorithm. Each point is associated with a frame − to − frame distance.

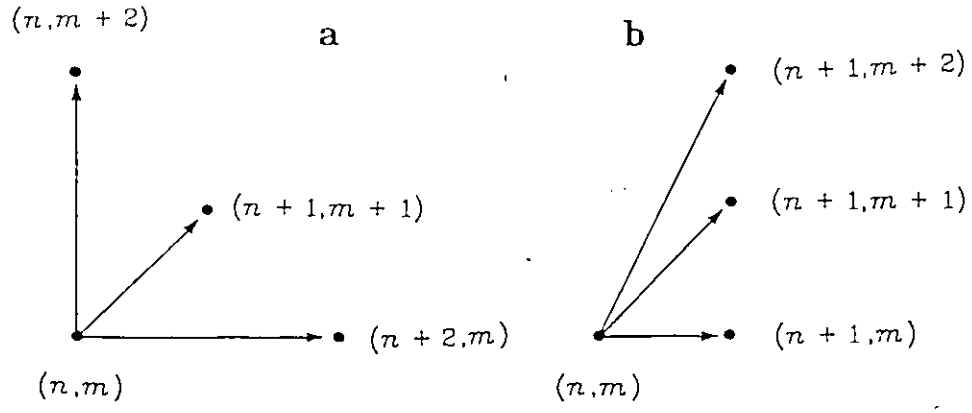

Figure 1.9 : Set of possible transitions in the two − dimensional lattice. a) Symmetric formulation (Checker − board method). b) Asymmetric formulation.

Recognition experiments do not show a clear advantage for one of the two formulations over the other [49]. In this work we have decided to choose the *checker−board* symmetric formulation method [50]. As we see in figure 1.9, this formulation defines that a point (n,m) may be followed by points $(n+1,m+1)$, $(n+2,m)$ or $(n,m+2)$. It may appear that we are loosing information by ignoring points $(n,m+1)$ or $(n+1,m)$. In fact, since spectral frames are generally derived from overlapping windows, successive spectral frames are sufficiently redundant that omitting single frames rarely leads to a loss of the useful information.

1.4.4 Constraints imposed on the elasticity of the time-alignment.

The basic step formulation presented above specifies only the ordering of parameter frames but does not constrain the compression or dilation of the time scales of the two spectral patterns being compared. Given that for speech sounds, timing can also carry information, it is desirable to restrain the time-elasticity of the warping algorithm with some constraints. One of these constraints, referred to as a hard slope constraint, places an absolute limit on the slope of the path. For example, the warping path may be prevented from having more than one or two horizontal or vertical steps in a row. Another kind of constraint, which is referred to as a soft slope constraint, consists of adding or multiplying a penalty cost to every vertical or horizontal step in the path. It can easily be seen that these constraints favor the warping path to be diagonal, and consequently they will tend to favor the comparison of spectral representations with similar segmental durations, and such patterns tend to be examples of the same word.

Having defined the monotonicity, continuity and slope constraints, equation 1.3, calculating the overall dissimilarity between template and reference pattern, can be represented by the following recursive equation:

$$D(n,m) = \min \left\{ \begin{array}{l} \delta_1 d(\mathbf{T}(n),\mathbf{R}_j(m)) + D(n,m-2) \\ \delta_2 d(\mathbf{T}(n),\mathbf{R}_j(m)) + D(n-1,m-1) \\ \delta_3 d(\mathbf{T}(n),\mathbf{R}_j(m)) + D(n-2,m) \end{array} \right\} \qquad (1.4)$$

where δ_1, δ_2, δ_3 are the soft slope penalty costs and the calculation goes from $m = 1$, $n = 1$ to $m = M$, $n = N$.

The constraints can be set to maximize the recognition results. They will probably be dependent on the front-end analysis used. Since in this work we evaluate the performance of a front-end recognition analysis compared with the results obtained with another analysis, the constraints are set so as not to favor one of the two analyses. That is, there is no hard slope constraint and the soft slope penalty costs are set to 1.

The result of equation (1.4) is time-normalized to compensate for the number of points introduced by the warping function. Since the checker-board method ensures that the number of local distances to be $(N+M)/2$ or $(N+M+1)/2$ depending on whether the sum $N+M$ is even or odd, the result is divided by. $(N+M)/2$ or $(N+M+1)/2$. Thus, the final distance between test and reference pattern would be :

$$D(\mathbf{T},\mathbf{R}_j) = \left\{ \begin{array}{ll} \dfrac{D(N,M)}{(N+M)/2} & \text{if} \quad N+M \ \ EVEN \\ \dfrac{D(N,M)}{(N+M+1)/2} & \text{if} \quad N+M \ \ ODD \end{array} \right. \qquad (1.5)$$

1.4.5 Representation of each spectral pattern frame by cepstrum coefficients

The comparison of spectral patterns can be computationally expensive considering the number of frame-to-frame dissimilarity measures that need to be calculated and the number of spectral parameters (K) used in each of them. We might consider reducing the number of bandpass filters, but this would affect the recognition performance. On the other hand, the spectral parameters in each frame can be converted into *cepstrum coefficients*. The cepstrum analysis consists of the spectral analysis of the short-time spectrum which can be calculated by a Cosine Transformation:

$$C_l(n) = \sum_{k=1}^{K} t_k(n) \cos\left[l \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \qquad l = 1,2,...,L \qquad (1.6)$$

where L is the number of cepstrum coefficients and $t_k(n)$, k=1,2,...,K represents the log-energy output of the k'th filter.

Due to the orthonormality of the cosine basis functions, and because the frame-to-frame dissimilarity measure is calculated from a Euclidean distance metric, there is no difference in measuring frame-to-frame dissimilarity from K spectral parameters or K cepstrum coefficients. Moreover, the recognition performance decreases only slightly when a number as small as 6 cepstrum coefficients is used [12], because while energies in adjacent channels are highly correlated, cepstrum coefficients are largely uncorrelated.

A property of the cepstrum analysis also provides a simple method of improving the similarity between spectral frames of the same sound segment being pronounced at different intensities. It consists of excluding the 0'th cepstrum coefficient in the Euclidean distance calculation since it corresponds to the mean energy level of the short-time spectrum, while the other terms do not in principle depend on the overall intensity.

1.5 DIFFERENCES BETWEEN VARIOUS FRONT-END ANALYSES

1.5.1 Front-end analysis based on Linear Predictive Coding

One of the most commonly used methods over speech analysis for the last twenty years has been linear predictive coding (LPC). This method has become a predominant technique for digital communication systems because it provides very low-bit-rate transmission and storage of speech without reducing the intelligibility excessively. It also performs well for speech recognition applications, and with the growing availability of VLSI LPC circuits, speech recognizers based on this analysis method are becoming common.

The idea behind LP analysis is that the vocal tract can be modeled as an all-pole filter corresponding to the linear prediction equation:

$$s[nt] = a_1 s[(n-1)t] + a_2 s[(n-2)t] + \ldots\ldots + a_p s[(n-p)t] + e_n \qquad (1.7)$$

or in the Z-domain by:

$$\frac{e(z)}{s(z)} = 1 + \sum_{q=1}^{p} a_q z^{-q} \qquad (1.8)$$

where the s[nt]'s are the current speech samples, the a_q's are the linear predictive coefficients, p is the order of the predictor, and e_n is the *prediction residual*.

Figure 1.10 illustrates the process inside an LP analysis system. It begins by windowing the input signal (m samples) to obtain short-time segments. For each segment the squared error average is calculated :

$$E_n = \sum_m e_n^2 \qquad (1.9)$$

The LP coefficients a_q are found by minimizing the squared error average. That is by setting $\partial E_n / \partial a_q = 0$, q=1,2,...,p. It can be shown that this minimization can be achieved by solving a matrix equation involving the autocorrelation properties of the windowed waveform. Depending on the details of how the autocorrelation properties are estimated, an *autocorrelation* or *covariance* method is said to be used [45].

The important point we want to make with this introduction to LP analysis is that a suitable number of LP coefficients (~10) permits us to obtain a good approximation to the formant frequencies and bandwidths of each speech sound segment analyzed. This can be seen in the relation $s(z)/e(z)$, which is similar to a digital filter equation where the poles (z_1, z_2, ...) correspond to the resonances of a speech waveform. That is, we have the equation:

$$\frac{s(z)}{e(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots} = \frac{1}{(1 - z_1/z)(1 - z_2/z) \cdots} \qquad (1.10)$$

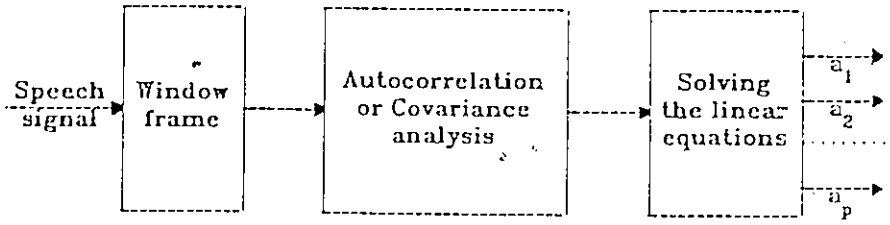

Figure 1.10 : *LP analysis*

Figure 1.11 shows a voiced speech waveform segment (upper right) and the corresponding spectrum together with spectral envelope defined by the filter equation $s(z)/e(z)$. We can see that the spectral envelope estimated from the LP analysis is a very good approximation to the short-time spectrum. That is, the peaks corresponding to the harmonics of the excitation (vibration of the vocal cords) are not present in the envelope and only a smooth energy contour surrounding the formant peaks is modeled. Such a short-time analysis therefore allows us to derive appropriate representations suitable to be used for recognition.

Another important characteristic of LP analysis is that it is a fit that is biased towards matching the high-energy parts of the spectrum. This results from the fact that no zeros are present in the digital filter equation 1.11. This can be seen as beneficial in the sense that quantizing noise which may dominate the signal in low-energy regions of the spectrum will be "masked" or made inaudible by the simultaneous presence of the speech signal. Such an effect can be compared to the frequency masking encountered in human auditory perception,

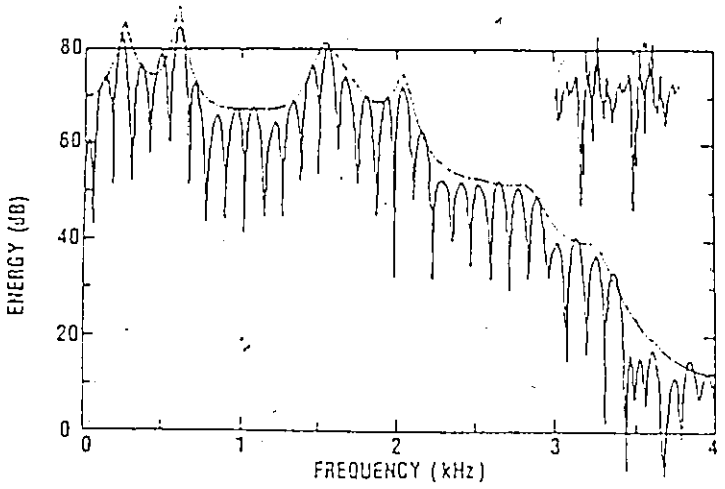

Figure 1.11 : Brief segment of a voiced speech signal (upper right), its short−time spectrum showing the harmonics of the fundamental frequency and the spectrum envelope obtained by linear prediction (after Schroeder [51]).

where high-energy frequency components of a signal mask low-energy frequency components. On the other hand, this kind of analysis causes information about anti-resonances such as occur in nasal consonants or nasalized vowels to be deemphasized in the LP speech representation.

But the good recognition performance obtained with LP analysis demonstrates that it is more advantageous to have some low energy masking effect even if some information is lost.

1.5.2 Comparison between LPC and filter-bank representations

The description of LP analysis in the previous section might suggest that a representation obtained from it would perform better than a representation obtained from a filter-bank analysis in a digit recognition task. This appears not to be the case, since some studies comparing both methods have shown that they were essentially equivalent when used in a dynamic programming time-warping speech recognition system [57], and some others [12] have even indicated that the filter-bank method gives better results. The reason is thought to be that a filter-bank analysis with filters spaced linearly at low frequencies and logarithmically at high frequencies captures the phonetically important characteristics of speech while LP analysis does not.

Another advantage of filter-bank-based recognizers over LPC-based recognizers is their robustness to additive noise and to other forms of channel distortion. It is known that LPC analysis systems tend to become error prone with high background noise levels ($SNR \leq 6\ dB$) and other transmission distortions whereas filter-bank systems seem to be more robust to noise [11].

1.6 DISTORTIONS AFFECTING PERCEPTUAL CUES IN SPEECH REPRESENTATIONS

1.6.1 Effect of speaker differences

Major sources of difference between the voices of different speakers are differences in the vocal tract physiology, in the laryngial physiology and in usage of the larynx and the vocal tract. Differences stemming from first two sources can be better appreciated by comparing the production of speech by males and females. For example, a man's larynx is much larger than a woman's and consequently the average value of adult male fundamental frequency (around 100 Hz) is about half that of an adult female (around 200 Hz) [27]. Also, because a woman's vocal tract is typically about 15% shorter than a man's vocal tract, the vocal tract resonances in the speech are some 15% higher in frequency [27].

If we refer to figure 1.2, it may be seen how the differences in vocal tract length and larynx sizes between men and women would affect the short-time power spectrum of similar speech sounds. Since the average fundamental frequency for women is higher than for men, the separation between harmonics in the power spectrum of a speech sound produced by a woman will be generally greater. Also, because the vocal tract resonances are 15 % higher, the peaks of the spectral envelopes will be shifted higher in frequency

The performance obtained from a front-end analyzer in a speaker-independent speech recognition application depends partly on the extent to which the representations are affected by the sources of differences. Since the difference

between men's and women's voices leads to very different spectral representations, most of the conventional recognition systems deal with this problem by using different templates for the two sexes.

Conventional front-ends are somewhat insensitive to the fundamental frequency of the voiced excitation signal. For example, figure 1.11 shows that the spectral envelope obtained from LP analysis is almost not unaffected by the separation of the harmonics produced by the excitation. In the case of a filterbank, the effect of the excitation is reduced by having the bandwidths of the bandpass filters large enough that at least one harmonic of the excitation is included in each channel.

Although conventional front-end analyses may decrease some differences between speakers, the recognition performance is still affected by other factors. The best technique that has been used to deal with these factors consists of having multiple averaged reference templates covering the voice differences of the expected users [12]. Typically, six spectral patterns are used for each word when the speaker population forms a homogeneous dialect group. Of course, there is a limit to how far this technique can be applied, since when the number of speaker accents is increased, the pronunciations of different words begin to overlap.

There are also other techniques that have been used to adapt to new speakers or to changes in the speaker's voice over the period when he is using the recognition system [36]. Speech recognition systems using these techniques update units of speech sounds (syllables,words) corresponding to the units in the utterance that it to be recognized. Such a method has been found to work very well in a speech recognition system, although there is a danger that the

recognition system may misrecognize an utterance resulting in a deterioration of the units in the reference patterns.

1.6.2 Effects of noise on the speech representations

A major factor which causes similar speech sounds to be different in the reference and test representations of a speech recognition system is noise. Some studies comparing filter-bank and LPC analyses in a noisy speech recognition task show that the performance begins to degrade for signal-to-noise ratios as high as 24 dB [11].

Noise can affect speech representations in different ways. For example, we will expect the noise to obscure features in a short-time spectrum of a speech sound. In addition, at moderate and high noise levels, a person usually modifies his voice when he speaks. That is, to make it understandable, he will raise his voice. This will affect the spectral representation of his speech because when a person raises his voice, the spectral envelope of his excitation is usually tilted upwards in frequency [23].

Various techniques have been presented to resolve the noise problem, including, recording the reference templates in the environment in which the machine is to be used, taking noise into account in the spectrum comparison process [23], and using noise-canceling microphones. Improved results have been obtained, but they are still far from what human listeners can do.

CHAPTER 2

MODELING THE AUDITORY SYSTEM

2.1 INTRODUCTION

It seems surprising considering the advances in knowledge in the field of auditory physiology and psychoacoustics that very few auditory models have been investigated as front-end analyses for speech recognition systems. Yet, there are many good reasons for carrying out such an investigation. Up to now, the recognition performance obtained with current systems is worse than human performance on the same task. The difference is even greater when the operating conditions become more difficult. One might therefore expect that the extraction of speech feature by an auditory model would contain the relevant cues and would be more robust.

2.2 THE AUDITORY SYSTEM

Before presenting the computational auditory model that is investigated in this work as the front-end analysis for a speech recognition system, we review in the following paragraphs some properties of the auditory system and some of the operations carried out there. We focus mainly on the description of properties that seem to be important in the extraction of perceptual cues characterizing speech sounds.

The auditory system is functionally subdivided into the peripheral and the central auditory systems. Figures 2.1 and 2.4 show the various processing stages corresponding to these two systems. The peripheral auditory system can be divided into three processing stages, namely, the outer ear, the middle ear and the cochlea. The function of these stages is to derive features from the incoming air sound pressure signal and to retransmit them as a series of nerve impulses to higher auditory centers. The central auditory system may be divided into four or five main processing stages. We can reasonably assume that each of these processing stages must have different functions in analyzing place, temporal and amplitude features in nerve fiber responses for sound localization, sound identification, speaker identification and speech recognition.

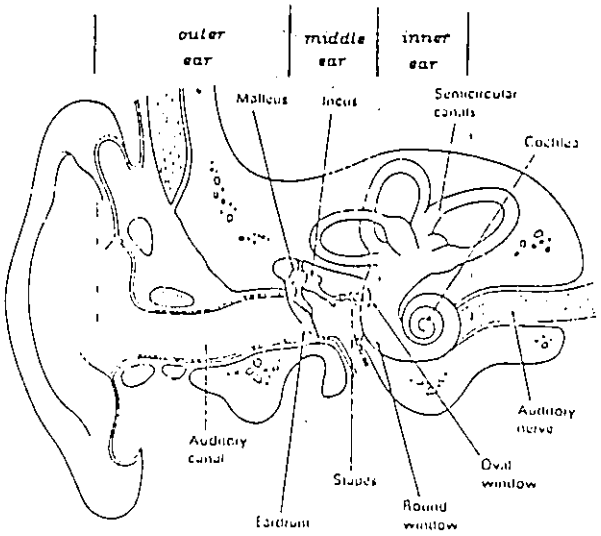

Figure 2.1 : *Peripheral auditory system.*

2.2.1 Peripheral auditory system

2.2.1.1 Outer and middle ear

The outer ear consists essentially of a cavity open at one end to the external sound field and closed at the other end by the relatively stiff tympanic membrane (the eardrum). Its geometry provides directionality cues, which help in sound localization. The middle-ear consists of the three ossicles : *malleus*, *incus* and *stapes*, which are coupled together and provide a mechanism for matching the impedance at the eardrum with that of the cochlear fluid.

Although the function of the external and middle ear is principally to transfer sound signals carried by an air pressure waveform into a fluid pressure waveform, the transformations accomplished by these organs also affect the

perception of speech sounds. The combined transformation of these organs can be seen as a preemphasis filter. That is, it provides a resonance which enhances components of an incoming sound in the range of frequencies of 1 to 7 kHz [2]. Such a transformation is similar to a first-order differentiation, which can be simulated by a finite impulse response of the form :

$$H(z) = 1 - az^{-1} \qquad 0.5 \leq a \leq 1 \tag{2.1}$$

2.2.1.2 The cochlea

The cochlea, represented in figure 2.1, consists of a coiled tube filled with fluid, which is divided into two chambers along its length by the cochlear partition. The cross-sectional view of the cochlea shown in figure 2.2a reveals that the cochlear partition includes the *basilar membrane* (BM), the *organ of Corti* and *inner hair cells*. The signals are transduced into nerve impulses in these cochlear areas.

The transformation occurring in the cochlea can be described in the following way. The sound signal transmitted via the ossicles of the middle ear to the oval window of the cochlea creates a fluid waveform in the chambers. Because the mass and compliance (stiffness) of the BM vary monotonically along its length, the waveform produces a specific motion of the BM. As illustrated in figure 2.2b, for a single stimulus, starting from the oval window (base), the amplitude of the BM motion builds up to a maximum, then decreases rapidly. The position of the maximum depends on the frequency. In this way, the

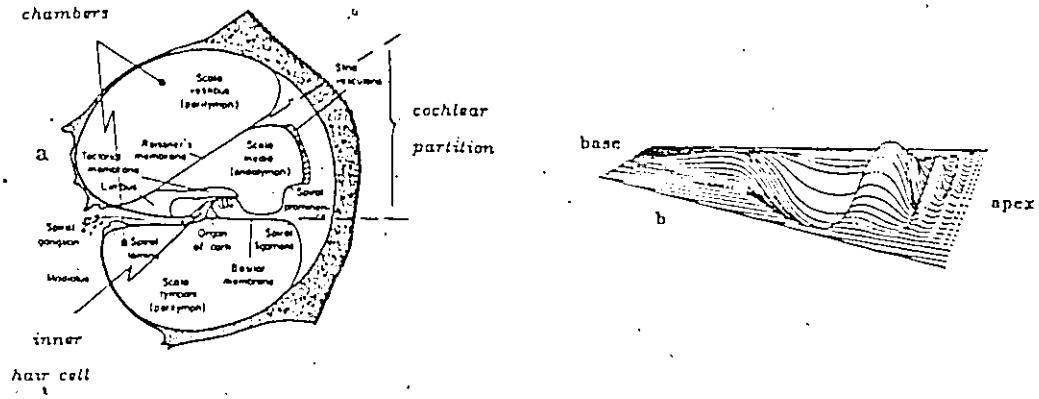

Figure 2.2 : a) cross−section of the cochlea. (after Pickles [41])
b) basilar membrane displacement. (after Warren [55])

envelope of the BM displacement represents a spectral analysis of the incoming signal, with frequency converted into place.

From the above description we can regard each BM area as a tuning filter, with high frequencies tuned at the base of the cochlea and low frequencies tuned closer to the apex. Physiological experiments have shown that typical BM tuning curves have an asymmetric bandpass characteristic with a slope of approximately a 6-12 dB/octave at the low frequency-end, a very steep slope ($>$100 dB/octave) and fast rolloff (curvature of the response close to the center frequency) at the high-frequency end [1]. Figure 2.3 shows the filter response corresponding to BM tuning.

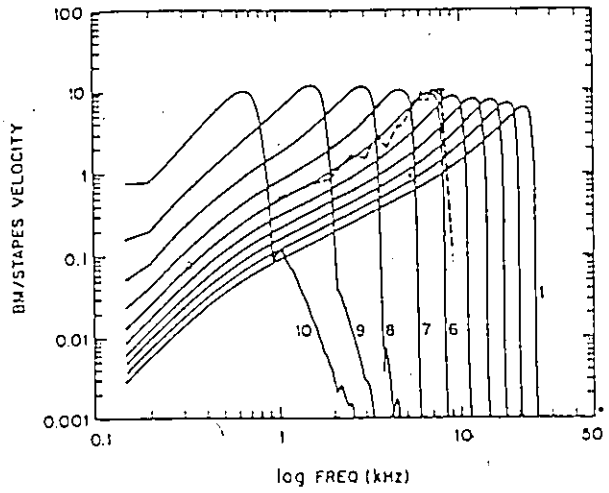

Figure 2.3 : Tuning response at various points along the BM. (after Allen [1])

In the organ of Corti the BM motion is converted into nerve impulses by means of receptor cells known as hair cells. The exact process is still not very well understood but researchers have suggested various non-linear processes which would enhance components of the signal at different sound levels. Among them, are suggestions that active mechanisms are present in the cochlea to enhance the gain applied to very low level signals [13]. It has also been suggested that compressive adaptation mechanisms are present to produce the lateral inhibition measured in the response of the nerve fibers [37].

The only well understood non-linear transformation occurring in the organ of Corti is the mechanical-to-electrical transduction taking place at the inner hair cell sites. The relationship between the inner hair cell input to the neural output is that of a half-wave rectifier. The purpose is similar to that of an amplitude

demodulator such as a diode in an AM radio, the positive envelope is used to represent the information contained in the signal.

2.2.2 Central auditory system

After the sound signal has been extracted and transformed by the peripheral auditory system, it is sent to the central auditory system. The principal processing centers of the central auditory system are the four brain stem nuclei: *cochlear nucleus*, *superior olivary complex*, *inferior colliculus* and the *medial geniculate body* and the *auditory cortex* (see Figure 2.4). In contrast to the peripheral auditory system, we have few notions about the methods by which the central auditory system is able to produce the observed phenomena of hearing - sound localization and analysis of complex sounds such as the understanding of the speech signal and the identification of the speaker.

Several experiments have been carried out to measure the different responses in neurons of the higher auditory centers. We know that different representations of an incoming sound signal are present in the neurons of the cochlear nucleus [41]. We may speculate that the function of latter auditory centers would be to extract cues in some of these representations to differentiate speech sounds. The function of the olivary complex is to carry out correlations between signals coming from the two ears. Thus, it decodes from the two signals the direction of the source of sounds in space. We know that the inferior colliculus is also involved in auditory reflexes, such as the startle reflex in response to loud sounds. Finally, the auditory cortex is responsible for the analysis of complex sounds. This part of

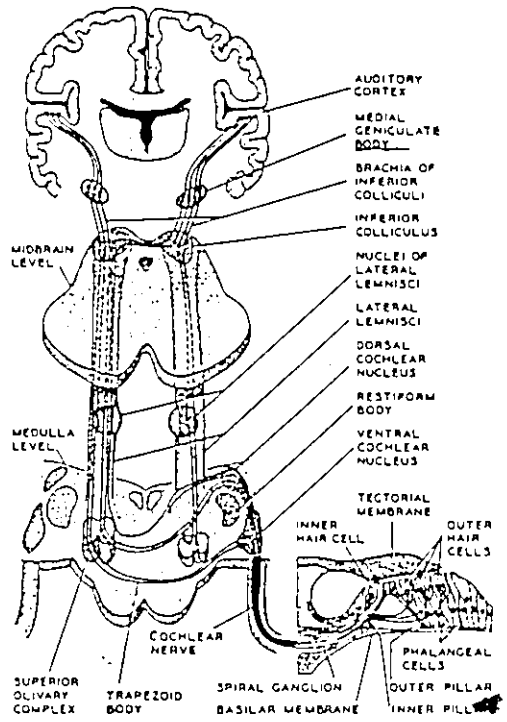

Figure 2.4 Schematic diagram of ascending auditory pathways.
(after Flanagan [18])

the brain would be similar to a processor with a certain amount memory and which takes decisions about the information received.

2.2.3 Properties of the auditory system

Because the details of some mechanisms of the auditory system are not very well understood, we are to some extent obliged to model the properties of hearing as found in psychoacoustic experiments and physiological measurements rather than modeling the mechanisms directly. These properties have been determined

by testing the perception of some listeners to various sound waveforms. The most important of them are phase-insensitivity, frequency masking, critical bands, and the fact that the ear's perception of loudness is proportional to the log-magnitude of the incoming sound. In order to see their importance for the perception of speech, we review each of them and we give some examples of their effects in the following paragraphes.

Considering the first property listed, since we are relatively insensitive to phase information contained in a signal, it follows that it would make little sense to build a speech recognizer that tried to recognize specific waveforms, since a slight change in the relative phases of its spectral components such as would be caused by room reverberation would not be detectable by human ears, yet it would cause the device to conclude that the sound was different.

The second property, which is one of the most interesting properties of human perception, is the masking effect. It is known that the perception of one sound can be obscured by the presence of another, or more specifically the presence of one sound raises the threshold of hearing of the other sound. A classic masking experiment, so-called two-tone suppression, measures the masking level obtained by a tone with fixed amplitude and frequency on a second tone [41]. The figure 2.5 shows the effect of a tone at 1200 Hz and 80 dB SPL (the Sound Pressure Level is the level of a pressure waveform in dB relative to the amplitude reference of 0.0002 dyne/cm^2 [55]) on a second tone whose frequency and amplitude are varied to measure the masking level. As we can see, the second tone is not masked if its frequency is lower than about half the frequency of the fixed tone. However, when the second tone is within 100 Hz of 1200 Hz, it needs at

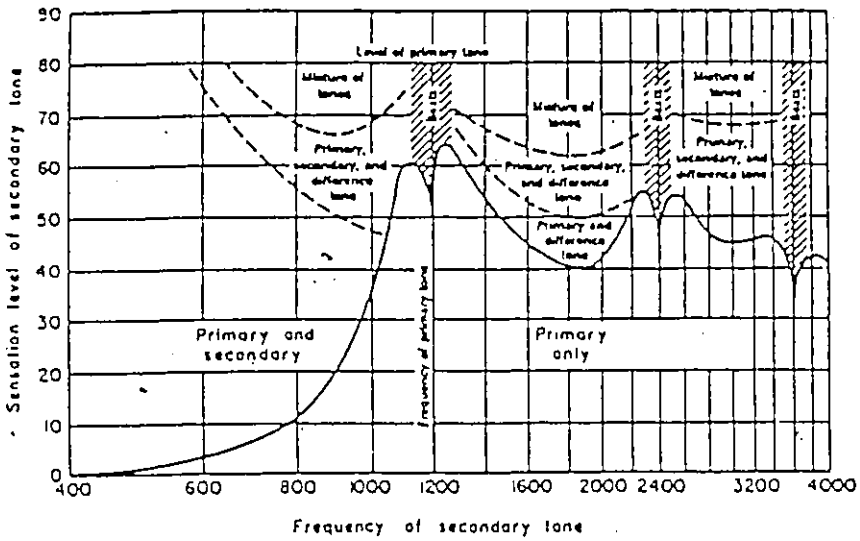

Figure 2.5 : Sensations produced by a two − tone stimulus. The primary tone at 1200, 80 dB above auditory threshold is fixed, while the secondary tone varies in frequency (horizontal axis) and intensity (vertical axis). Below the solid curve, the secondary tone is masked (not heard); above it, beats and various combinations of the tones are heard. (after Fletcher [19])

least 50 dB to be heard. This masking frequency remains for higher frequencies as well: the second tone requires at least 40 dB to be perceptible.

It can be seen from the last example that frequency masking should improve the perception of voiced sounds in noise. That is, frequency components of some noises near the frequency of a narrow-band signal will be masked if the noise intensity is lower by a certain amount than the intensity of the narrow band signal.

The masking effect is also closely related to another property of hearing namely the critical band effect. As an example of this effect, in the 1950's

Zwicker demonstrated that a signal consisting of a narrow band of noise can be masked by two tones, one above it in frequency and the other below. As the tones were moved together in frequency, the intensity threshold at which the noise was just detectable was constant at first and then rose steeply when the tones came within a certain frequency bandwidth called the critical band [41]. In particular, it has been measured that below 500 Hz the critical bandwidths are approximately constant in width, and above 500 Hz, the critical bandwidth increases linearly with frequency. The critical bandwidth (CB_c) can be calculated for a given frequency (f_c) in *kHz* from the equation derived by Zwicker [59] :

$$CB_c \ / \ Hz \ = \ 25 \ + \ 75 \ [\ 1 \ + \ 1.4 \ (\ f_c \ / \ kHz \ )^2 \ ]^{0.69} \qquad (2.2)$$

Another useful equation that has been used to represent the critical bands in terms of a linear scale is the critical band rate given in *barks*. For a frequency, f, in *kHz*, critical band rate, z_c, in barks and arctan in radians, the following analytical expression is proposed :

$$\frac{z_c}{Bark} \ = \ 13 \ arctan \ (\ 0.76 \ \frac{f}{kHz} \ ) \ + \ 3.5 \ arctan \ (\frac{f}{7.5 \ kHz})^2 \qquad (2.3)$$

The bark range is usually used to specify the spacing of bandpass filter responses in an auditory processor. It can also be used to explain the behavior of the responses in auditory nerve fibers to various sounds.

To see how masking and critical bands may help in analyzing speech, we can consider how a person can differentiate speech sounds by looking at a speech spectrogram [9]. He will differentiate sound segments mainly by paying attention mainly to the regions where the energy is concentrated or where the contrast is

more pronounced and by paying relatively little attention to low energy regions. In doing this, irrelevant information is removed and consequently only the relevant speech cues contribute to the comparison.

It could be argued that information is also contained in the lower energy portion of the spectrum. To defend the idea of masking for this situation, we can consider the case of LPC analysis as presented in chapter 1, which demonstrates that low energy components do not contribute significantly to the representation of sounds. Also, it can be demonstrated that masking occurs mostly in the neighborhood of regions of high energy, so that lower energy areas will not be masked if they are not surrounded by higher energy areas.

One exception where masking can occur in larger areas is when we observe the attenuation of the 4'th and 5'th formants. As we observe in figure 2.5, the sensation level of hearing for high frequency tones is increased when another tone of lower frequency is presented in the same time. Also, to support the fact that higher formant frequencies are not important, experiments on human perception with synthetic voiced sounds indicate that just the first two or three formants characterize these sounds almost completely [34].

2.2.4 Neurophysiological properties of hearing

We might hope to gain a more detailed understanding of hearing properties found in psychoacoustic experiments by studying their physiological properties. In particular, studies of the activity in the auditory nerve fibers of a cat in response to complex sounds have made it possible to speculate on how speech

sounds could be analyzed in a human brain [32][47][48].

Signals appearing in a single auditory nerve fiber consist of a train of spikes. Even in the absence of sound stimulation, the nerve fibers have a non-zero discharge rate called the spontaneous rate [41]. In response to sound, the discharge rate of spikes is initially constant until the stimulus intensity is large enough to cause the spontaneous rate to be exceeded. It then increases approximately linearly with tone level, up to a point when the discharge rate is saturated. For low frequencies (below 5 kHz) the train of spikes occurs during the positive half-cycle of the stimulus waveform.

One physiological measurement that has been used to study the train of spikes in single nerve fibers is the calculation of the average rate response by a post-stimulus time histogram (PSTH) [48]. This histogram is made by presenting a stimulus many times, and recording the times of occurrence of each spike, relative to the time of presentation. For example, when a tone burst was used as the stimulus, histograms showed an initial peak of activity, followed by a reduction in activity which was initially rapid, then slower. The application of this measurement has therefore been found useful in determining adaptation behavior to different sound signals.

Another measurement called the synchronized rate response [32] was introduced later, principally to study the fact that the train of spikes observed in the nerve fibers discharges in a half-cycle of the stimulating waveform. The method consists of creating a period histogram by calculating the number of discharges in a time interval as a function of the stimulus phase. Fourier transforms of period histograms of periodic stimuli have shown that the train of spikes in many

auditory nerve fibers were phase-locked to periodicities of the stimuli.

Interesting studies have been done by Sachs and Young [47] to determine whether the rate response alone was sufficient for vowel identification or whether a synchronized measure would help in recovering formant peaks. They measured averaged and synchronized rate responses to synthetic vowel stimuli presented at several stimulus levels in a large population of auditory nerve fibers. While the vowels were easily perceptible at stimulus levels corresponding to a normal level of hearing, they found that formant information was almost completely lost in averaged rate responses because the rates had saturated. They found, however, that formant-peaks could be easily identified in Fourier transforms of period histograms. From this study, they concluded that some form of synchrony analysis is performed at higher auditory stage of the auditory system to recover formant peaks.

2.3 SENEFF'S AUDITORY MODEL

A computational model of the auditory system might consist of a filter-bank analysis followed by different non-linear functions such as non-linear compressions and rectifications. Among the auditory models that have been presented in the past few years, some of them use compressive non-linearities to increase contrast between spectrum areas of high energy and those of low energy by a form of lateral inhibition. In some of the models, this has been implemented by inserting a non-linear function between two different bandpass filters [10][40]. In

another, it has been implemented by using a compressive network [37].

One particular approach that has been recently used by S. Seneff [53] is to design a non-linear function that performs a synchronicity measurement of bandpass filtered signals similarly to what we believe is done in the auditory system. This approach is attractive since the lateral inhibition produced by the model is mainly used to enhance formant peaks in spectral representation of speech sounds.

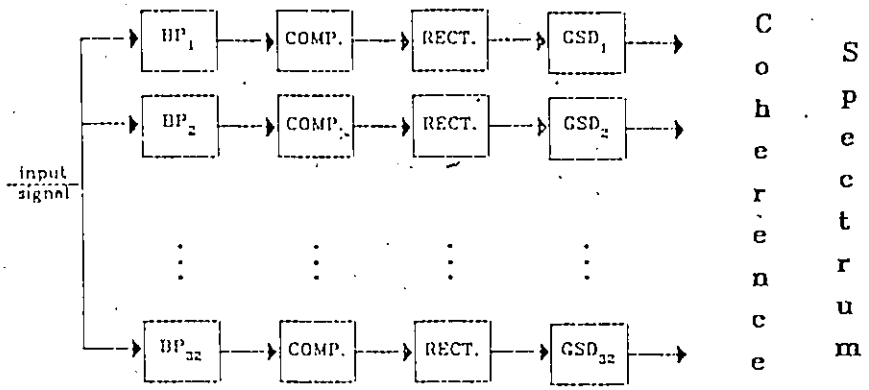

Figure 2.8 : Seneff's auditory model.

As schematized in figure 2.6, Seneff's auditory model consists of a bank of digital filters (BP), followed by a non-linear compressive stage (COMP), half-wave rectifiers (RECT) and by a set of generalized synchrony detectors (GSD). The first three stages of this computational auditory model simulate some known cochlear mechanisms. The last stage is the non-linear transformation that measures the periodicity of bandpass filtered stimuli. Since Seneff's model is based on

the detection of periodicities in a signal, we will call the output a *short−time coherence spectrum*.

The bandpass filtering stage in Seneff's auditory model simulates the transformation accomplished along the basilar membrane of the cochlea. It consists of 32 bandpass filters spaced by half a bark. The frequency responses are similar to those shown in figure 2.3, although their rolloff on the high frequency-end is slower. The relative amplitude response of each bandpass filter is increasing with frequency to simulate preemphasis accomplished in the outer and middle ear.

The second processing stage is a non-linear compression consisting of two automatic gain controls AGC's of the form :

$$y[t] = \frac{x[t]}{k + <|x|>_\tau} \tag{2.4}$$

where $<|x|>_\tau$ is an estimate of the average magnitude of the input signal, x, obtained by leaky integration. The constant k will cause the output of the compression stage to be linearly related to the input signal at low amplitudes and log related at high amplitudes. Such a relationship is in agreement with the adaptation response measured in auditory nerve fibers. Values of k that would permit us to reproduce the adaptation response in nerve fibers range between 400 and 1000 for the slowest AGC and range from 1 to 20 for the fastest AGC [20][54]. An example of the response of the non-linear compression stage for a square wave is given in figure 2.7.

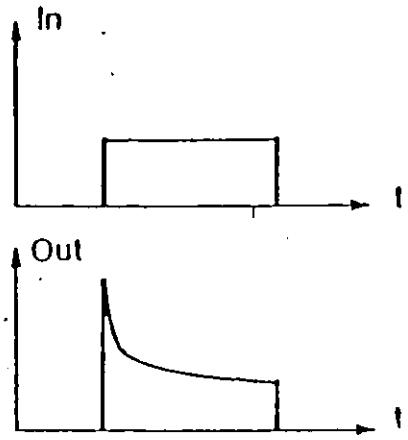

Figure 2.7 : Non−linear compression response for a square pulse. (after Seneff [53])

The equation representing the leaky integration in the two AGC's can be represented by the infinite impulse response of the form :

$$H(z) = \frac{1 - \lambda}{1 - \lambda z^{-1}} \qquad (2.5)$$

The constant λ in the leaky integration is a function of the sampling rate and of the time constants τ. This time constant is set to 3 ms. in one AGC and 40 ms. in the other. These values correspond to the averaged adaptation response measured in auditory nerve fiber responses [20].

The third stage in Seneff's auditory model is a half-wave rectifier to reflect the property measured in hair cell transduction that only positive portions of the

waveform contribute to the final response. This transformation can be represented by :

$$y[t] = \begin{cases} x[t] & \text{if } x[t] \geq 0 \\ 0 & \text{if } x[t] < 0 \end{cases} \tag{2.6}$$

where x[t] is the input and y[t] is the output.

2.3.1 Generalized synchrony detector (GSD)

The generalized synchrony detector is the transformation that provides a synchrony measure of the bandpass filtered stimuli. It can also be regarded as a non-linear function that reproduces some non-linear effects that have been measured in the responses of auditory nerve fibers.

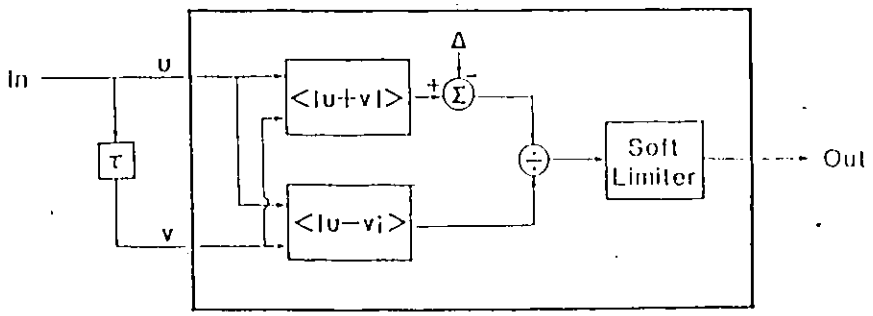

Figure 2.8 : Generalized Synchrony Detector. (after Seneff [53])

A block diagram of the GSD is given in Figure 2.8. The two input waveforms u and v consist of a stimulus detected in one channel and the same stimulus delayed by a time τ, which is the inverse of the channel center frequency. A sum and a difference waveform are constructed from u and v and the full-wave rectified outputs of both of these are passed through identical low-pass filters to obtain an envelope response. A small threshold, Δ, is subtracted from the sum envelope to prevent a strong response to weak but synchronous signals. This threshold may be considered to be the spontaneous rate response in nerve fibers when no sound is present. The output is then obtained from the division of the sum envelope by the difference envelope. In order to prevent a potentially infinite output when the denominator is near zero, a soft limiter is included.

2.3.2 Combined effects of different stages in Seneff's auditory model.

The most important advantage obtained with the Seneff's auditory model as a speech analyzer is the enhancement of formant peaks in spectral representations of voiced sounds. The division of the sum envelope by the difference envelope in the GSD can be interpreted as an excitatory/inhibitory process. Therefore, when the filtered stimulus being analyzed by one of the GSD's is periodic with period corresponding to the time constant of the GSD, which is the case if a formant is present at this particular frequency, the difference envelope is low and the result of the division is therefore high. On the other hand, when the input of the GSD is varying in amplitude, such as in tone onset and offset, the difference envelope is high so that the result of the division is low.

Moreover, it can be seen that when a tone (suppressing signal) at a frequency nearby is added to the initial stimulus, it will decrease the response of the GSD output. This is due to the fact that the second signal increases the difference envelope in the GSD, which has the effect of decreasing the response at the output. This particular effect can be associated with the *two−tone suppression* or *frequency masking* effect that has been observed in psychoacoustic experiments.

2.3.2.1 Effect of the bandpass filtering on the GSD analysis

The degree of suppression in Seneff's auditory model depends critically on the amount of filtering by the bandpass filter of the suppressing signal. It turns out that for this model higher frequencies would mask lower frequencies more effectively than the converse if a symmetric bandpass filter were used. Filtering the incoming signal through an asymmetric bandpass response similar to the Basilar Membrane tuning therefore compensates for the asymmetry in the masking.

Preemphasizing the incoming signal before the bandpass filtering will reduce the suppression of high-frequency components by low-frequency components. That is, higher energy formants, which for voiced sounds are located in the lower frequency part of the spectrum, inhibit the synchronous analysis of higher frequency formants less if the incoming signal is preemphasized.

2.3.2.2 Effect of the non-linear compression on the GSD analysis

The principal benefit of the non-linear compression stage is to improve synchrony detection when the amplitude of filtered stimuli is varying. In such a case the difference envelope of a periodic signal in the GSD's is kept small because the fastest of the two AGC's can adapt to one period of the filtered stimulus.

In addition to the frequency masking effect present in Seneff's auditory model, there is also an effect related to *forward masking* [41]. This particular effect can be observed by testing human auditory perception of noise sounds immediately followed by a low intensity tone. It has been found that when both the noise and the tone are present in the same frequency band then the perception of the tone is masked in the first 20 to 40 ms. For the auditory model, it can be seen that in such a case the slower of the two AGC's will take 40 ms or so to adapt to the lower intensity of the periodic input, and the input to the GSD will be inhibited during this period.

2.3.2.3 Effect of the half-wave rectification on the GSD analysis

The choice of rectification process may affect the synchrony analysis in some ways. For example a full-wave rectifier would cause the GSD to be sensitive to periodicity at half the center frequency. In a half-wave rectified signal, the GSD's respond only to filtered stimuli of periodicity close to the center frequency of the corresponding channel.

2.3.3 Resistance of the auditory model to distorted speech.

The spectral representations of voiced sounds obtained from Seneff's auditory model are consistent with some psychoacoustic experiments done by Klatt [34]. He found that spectral tilt, high-pass, low-pass and notch filtering as well as formant amplitude changes affect phonetic judgement only slightly. In Seneff's auditory model, a periodic filtered stimulus at the frequency of the corresponding channel is amplitude demodulated by the non-linear compression stage and the output of the GSD's is independent of the filtered stimulus amplitude. The coherence spectrum obtained from this model is therefore more resistant to various linear filtering operations such as occur when different microphones are used to record the speech. It also provides spectral representations that are less sensitive to differences in the envelopes of excitation spectra occurring between speakers and within speakers under, for example, varying degrees of vocal effort.

The lateral inhibition or frequency masking introduced in the GSD causes the valleys between formant peaks in the short-time coherence spectrum to be smoothed. Also, higher frequency formants will be attenuated since they have less energy than lower frequency formants. We may conclude that the auditory model will be more robust in background noise than conventional front-end analyses because it discards speech features that can be easily distorted by low noise levels.

CHAPTER 3

DETAILS OF THE IMPLEMENTATION OF THE AUDITORY MODEL

3.1 INTRODUCTION

As we noted in the previous chapter, Seneff's auditory model exhibits a form of lateral inhibition that strongly increases the contrast in the area of a formant peaks. Although this formant emphasis appears to be very useful for formant tracking, it turns out that this auditory model does not produce speech representations that are well suited to a pattern-matching recognition task. To find out what had to be modified in the computational auditory model in order to use it as a front-end of a recognition system, we tested it with single tones, two-tone signals and speech sounds. All of these tests suggested that we needed to change the lateral inhibition produced in the model.

3.2 WEAKNESSES OF SENEFF'S AUDITORY MODEL.

Seneff's auditory model extracts spectral representations of speech sounds by analyzing the periodicity in each channel corresponding to the center frequency. Tests with single tones have shown that the output of each channel is very sensitive to the frequency of the filtered stimuli. That is, when the input to a GSD is perfectly periodic at the center frequency (f_c) of one of the channels, the result of the division is infinite (or large since the arc tangent operation of the soft-limiter keeps the result finite). When the input is periodic at a frequency between two center frequencies, the result of the division is much lower.

Spectrogram-like representations of the output of the auditory model for speech signals showed that formant peaks could be affected in two different ways. One is due to the separation of harmonics of the fundamental frequency. For example, when the two strongest harmonics in the neighborhood of the first formant are close together, the first formant in the short-time coherence spectrum is represented by single peak and when the two harmonics are somewhat further apart the formant gives rise to two peaks because of the strong lateral inhibition.

Formants in the higher frequency part of the spectrum may also be affected by lateral inhibition. That is, as two formants approach to each other in frequency, they should gradually merge into one peak. Due to the lateral inhibition, this gradual merging does not occur and formants jump suddenly from two separate peaks to one merged peak. Since formant frequencies in the same speech sound may differ by 5 or 10%, the lateral inhibition in this case may increase the differences between the representation of equivalent speech sounds.

As we show in part 3.3.5 of this chapter, lateral inhibition exibited by Seneff's auditory model does not agree quantitatively with lateral inhibition measured in mammalian auditory nerves. Since human speech production presumably takes lateral inhibition into account, we might hope that by emulating quantitatively the lateral inhibition measured in mammalian ears we would improve the spectral representation for speech recognition purposes.

Another weakness of Seneff's model, which we mentioned at the end of chapter 2, is the insensitivity of the GSD analysis to tone onsets and to rapid frequency changes. This weakness would cause some perceptually important cues related to the onset of voiced sounds and of plosive sounds to be poorly represented in the spectral representation. To extract all relevant perceptual cues in speech sounds we would therefore need some other kind of analysis in parallel with the GSD analysis.

3.3 MODIFICATIONS TO THE AUDITORY MODEL

The computational auditory model has been implemented in a manner corresponding to the block diagram shown in figure 2.6. A pre-emphasis filter has been added at the input of the model. We have also applied a raised-cosine window of 25.6 ms length at the output of each channel, and we take the output at a 6.4 ms frame rate in order to provide an output that can be used in a template-based recognition task. To obtain speech representations that would permit us to emulate human frequency masking and to maximize the correlation

in a frame-to-frame comparison, modifications have been made to bandpass filter characteristics and to the GSD analysis formulated by Seneff.

3.3.1 Modifications to the filter-bank structure.

Like Seneff's auditory model, the new model has 32 channels spaced by half a bark. However, the channels span the frequency region from 100 to 3400 Hz rather than 100 to 2700 Hz. The four extra channels in the range 2700-3400 Hz are intended to increase information about voiceless sounds.

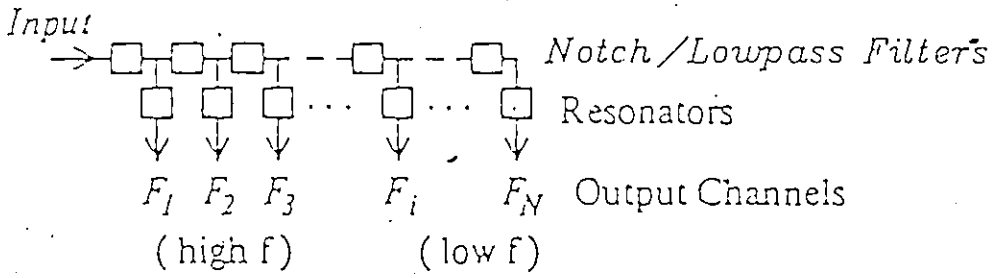

Figure 3.1 : Filter−bank structure based on Lyon's model.

The filter structure for each channel consists of a 12'th-order butterworth lowpass filter and second-order notch and resonator filters (see Appendix 1). The filters have been cascaded in the manner described by Lyon [37]. Some other researchers have also used a similar structure in their auditory models [14][15]. Since Lyon has 96 channels instead of 32, we have added a lowpass butterworth

filter in series with the notch filter to obtain the same lowpass response as multiple notches in series would provide. As shown in figure 3.1, the new filter structure consists of cascaded notch/lowpass filters in parallel with resonator filters. For such a structure, the filters compound their effects to increase the high frequency fall-off in the responses of the lower frequency filters. The resulting bandpass filter response obtained for each channel is somewhat more similar to that of the basilar membrane (see figure 2.3, p. 41) than the filter responses proposed by Seneff in her paper [53].

3.3.2 Insertion of an interpolating function in the GSD analysis

The synchrony detection in the GSD is calculated for lower frequency channels by using an integral number of samples corresponding to the time delay τ_c. For higher frequency channels, the approximation to the delay period provided by an integral number of samples becomes too inaccurate, and we have to resort to an interpolation between samples.

In her work, Seneff did not say how samples of high frequency signals have been interpolated, though we believe that she used a linear interpolation. As shown in figure 3.2, such a method would lead to considerable inaccuracies. On the other hand, since the bandwidth of higher frequency channels is small compared to the center frequency, the waveform at the output from the filter can be represented as a sine wave at the center frequency with slowly varying amplitude and phase modulation. We can therefore approximate the signal locally as a sine wave at the center frequency. To interpolate between a pair of samples, the

samples are used to estimate the local phase and amplitude of the sine wave and this function is then used for the interpolation.

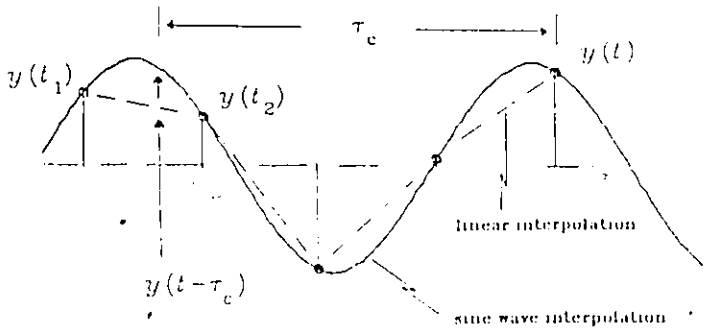

Figure 3.2 : Samples of the signal to be interpolated.

We can approximate the samples $y(t)$ of a signal as shown in figure 3.2 by :

$$y(t) = A \cos\left(2\pi f_c \frac{t}{SF} + \phi\right) \tag{3.1}$$

where SF is the sampling frequency, A is the amplitude of the sine wave, f_c, its frequency and ϕ is the phase at $t = 0$.

For two adjacent samples at time t_1 and t_2, we have the relation $t_2 = t_1 + 1/SF$, and substituting $2\pi f_c t_1/SF + \phi$ by ψ in equation 3.1, we can represent their amplitudes by the equations :

$$y(t_1) = A \cos\psi \tag{3.2}$$

$$y(t_2) = A \cos\left(\psi + 2\pi \frac{f_c}{SF}\right) \tag{3.3}$$

The value ψ can be calculated by dividing equation 3.3 by 3.2 and substituting the relation inside the brackets of equation 3.3 by the trigonometric relation $cos(\alpha + \beta) = \cos\alpha \cos\beta - \sin\alpha \sin\beta$. We therefore have :

$$\frac{y(t_2)}{y(t_1)} = \frac{\cos\psi\cos 2\pi \frac{f_c}{SF} - \sin\psi\sin 2\pi \frac{f_c}{SF}}{\cos\psi} \tag{3.4}$$

and solving 3.4 for ψ, we have :

$$\psi = \tan^{-1}\left(\frac{\cos 2\pi \frac{f_c}{SF} - \frac{y(t_2)}{y(t_1)}}{\sin 2\pi \frac{f_c}{SF}}\right) \tag{3.5}$$

The phase ϕ and amplitude A of the sine wave are given by :

$$\phi = \psi - 2\pi f_c \frac{y(t_1)}{SF} \tag{3.6}$$

$$A = \frac{y(t_1)}{\cos\psi} \tag{3.7}$$

Then the interpolated value needed to measure the synchronization for a delay τ_c can be calculated from equation 3.1.

3.3.3 Addition of adjacent-channel cross-correlation

As we noted earlier, the auditory model responds more strongly to a tone at a frequency corresponding to the center of one of the channels than to a tone of equal intensity whose frequency lies between two channels. We are reluctant to solve this problem by adding more channels because we do not want to increase the already high computational cost. Replacing the output from each channel by a sample-by-sample product of the output of pairs of adjacent channels mitigates the problem since a tone lying between two channel center frequencies produces weak outputs in both channels, while a tone centered on one channel produces a strong output in that channel and a very week output in the other. The correlated signal of pairs of channels will therefore provide a more even response across the spectrum.

3.3.4 Modification to the denominator in the GSD.

Although cross-correlating adjacent channels mitigates the problem of the sensitivity of the GSD to frequency of single tone, short-time coherence spectra are still affected by the splitting effect described earlier in this chapter. To reduce this effect, we have added a constant to the denominator of the GSD. The constant also prevents division by zero, so the inverse tangent operation is no longer necessary. The new expression for the GSD is :

$$\frac{<|\, u \,+\, v \,|>}{<|\, u \,-\, v \,|> + C_c} \tag{3.8}$$

where C_c is the constant for the c'th channel.

Small-scale digit-recognition experiments have shown that the recognition performance could be optimized for a certain value of C_c. Since it is not feasible to determine the optimal value of C_c separately for each channel and the splitting problem is worse at low frequencies than at high or intermediate frequencies, it seems appropriate to make C_c larger for low-frequency channels than for intermediate or high-frequency channels. Because the "Q" values of the channels (defined as the ratio of center frequency to critical bandwidth) are low for low frequencies and rise to a constant value above 500 Hz, it therefore seems reasonable to make the values of C_c inversely proportional to the Q of the channel.

3.3.5 Measurement of two-tone suppression contours.

Details of the lateral inhibition exhibited in the auditory nerve fibers have been investigated by Javel who measured *two-tone synchrony suppression* contours in the auditory nerve fibers of cats. His experiment consisted of measuring the response in nerve fibers to a fixed tone when a suppressor tone of variable frequency and fixed amplitude is also present.

In his experiments, Javel has shown that the suppression magnitude depends over a wide intensity range of the fixed tone, only upon the frequency and intensity of the suppressor. He expressed the dependence of suppression to frequency of suppressor tone in terms of two-tone suppression contours. Two of these contours measured in auditory nerve fibers of cats for channels that would correspond to center frequencies of a) 1.0 kHz and b) 1.5 kHz and for which the fixed tone is at 1.0 and 1.5 kHz respectively are shown in figure 3.3. As we see,

suppression is approximately 6 dB/Bark when the frequency of the suppressor tone is below the center frequency and 25 dB/Bark above it.

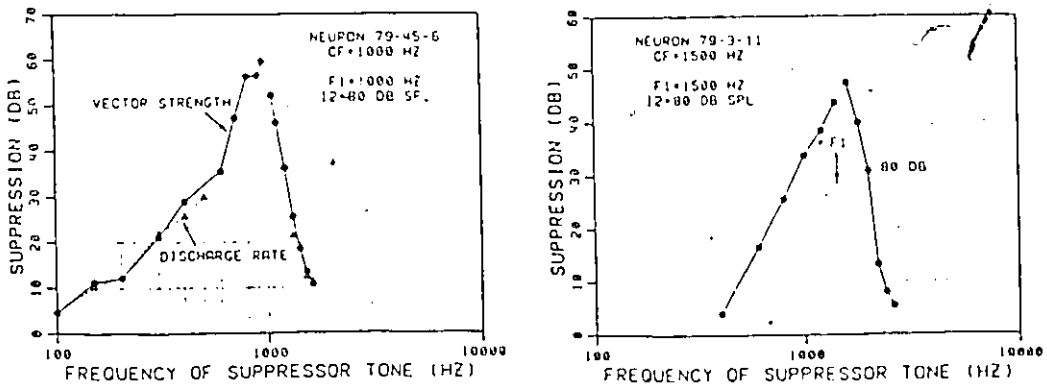

Figure 3.3 : *Two−tone suppression contours measured in physiological experiments at CF* = 1.0 ,1.5 *kHz for a fixed intensity* (80 *dB SPL*) *variable−frequency suppressor tone and a fixed frequency tone at* 1.0, 1.5 *kHz respectively.* (*after Javel* [31])

In addition to two-tone suppression contour measurements, Javel has measured the behavior of suppression as a function of the intensity of the suppressor tone, and he found that increasing the intensity of the suppressor tone produced monotonically increasing amounts of suppression. We show in figure 3.4 a measure in one nerve fiber of the degree of suppression as a function of the intensity of the suppressor tone. Javel found that when the intensity of a suppressor tone is greater than a certain level, the suppression of the fixed tone in dB is linearly related to the intensity in dB of the suppressor tone.

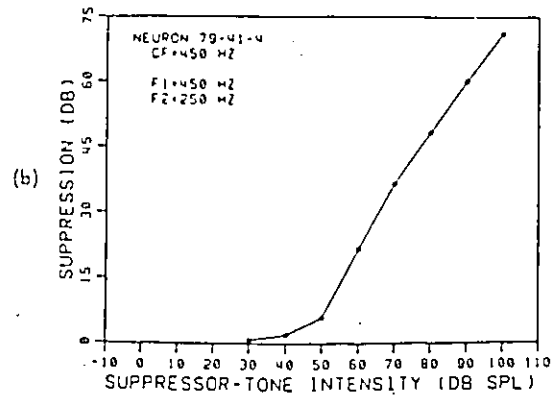

Figure 3.4 : Suppression magnitude (in dB) at CF = 450 Hz as a function of suppressor−tone intensity. The frequencies of the fixed and suppressor tones are 450 Hz and 250 Hz respectively. (after Javel [31])

It seems reasonable to suppose that if a computational auditory model can reproduce the suppression behavior measured by Javel, it will reproduce the frequency masking properties of mammalian ears when the input is a periodic signal. As illustrated in figure 3.5, we have measured the two-tone suppression produced in the appropriate channel of Seneff's auditory model under roughly the same conditions. The suppression contour we obtained is quite different from that measured by Javel. We notice that it is a much stronger above the center frequency than below it, while Javel's contours show the converse. Suppression is also present far away from the center frequency, while in Javel's suppression contour, it drops steadily as the frequency of the suppressor tone is lowered and it is not present at all for suppressor tones of high frequency.

We have also measured suppression properties of the new model with modified bandpass filter responses, with adjacent-channel cross-correlation and with the addition of a constant in the denominator of the GSD. As shown in

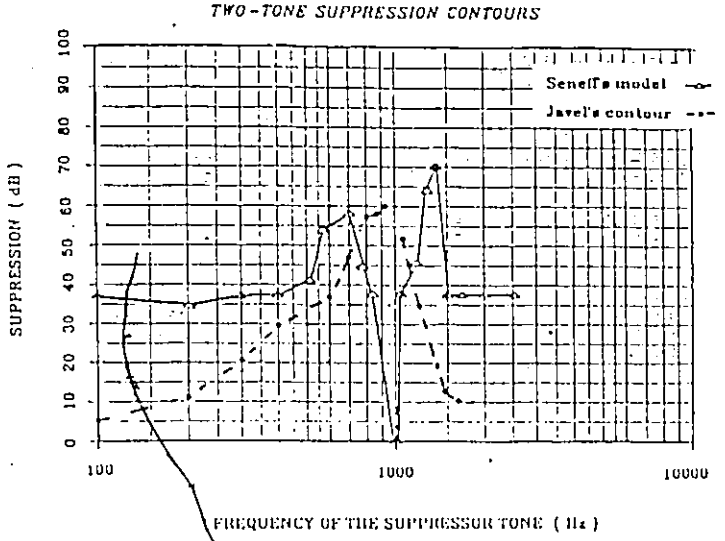

Figure 3.5 : Comparison of two−tone suppression contours measured at the output of the channel corresponding to f_c = 1000 Hz between Seneff's model and Javel's measurements.

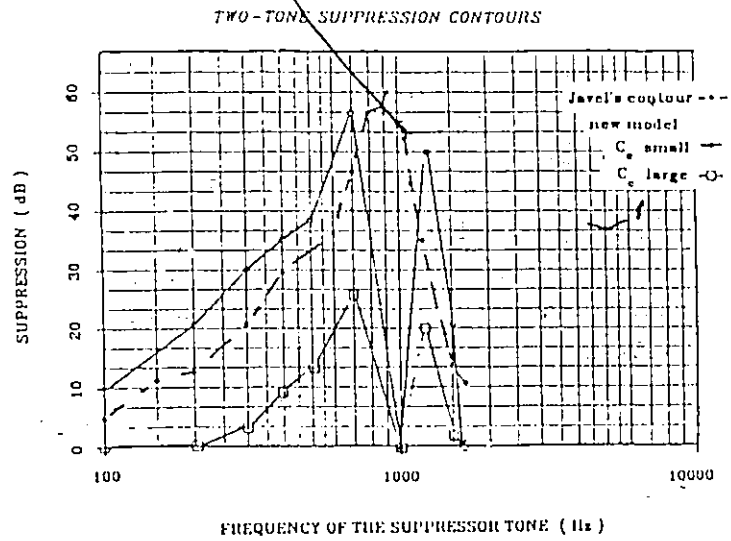

Figure 3.6 : Two−tone suppression contours measured at the output of the channel corresponding to f_c = 1000 Hz for different values of C_c.

figure 3.6, we have measured the two-tone suppression contours for different values of C_c at channel $f_c = 1.0\ kHz$.

The use of different values for the constants in the denominator of the GSD seems only to affect the suppression when the frequency of the suppressor tone is close to the center frequency. This permits us to believe that we can adjust the suppression behavior to reduce the formant peak splitting effect described earlier without affecting the behavior of suppression far from the center frequency.

The behavior of suppression as a function of suppressor tone intensity has also been tested. As shown in figure 3.7, different two-tone suppression contours are obtained from different intensities of the suppressor tone. Figure 3.8 indicates that the growth of suppression as a function of suppressor tone intensity in the model is similar to the results obtained by Javel shown in figure 3.5.

We cannot say at this point that we are able to reproduce quantitatively two-tone suppression behavior encountered in mammalian ears since the comparison has been carried out for only a limited set of results and we do not know if we have the same tone intensity levels in our experiment as in Javel's experiment. But we can conclude that the modifications have permitted us to reproduce the behavior of suppression in a better qualitative way than with the original version of the auditory model.

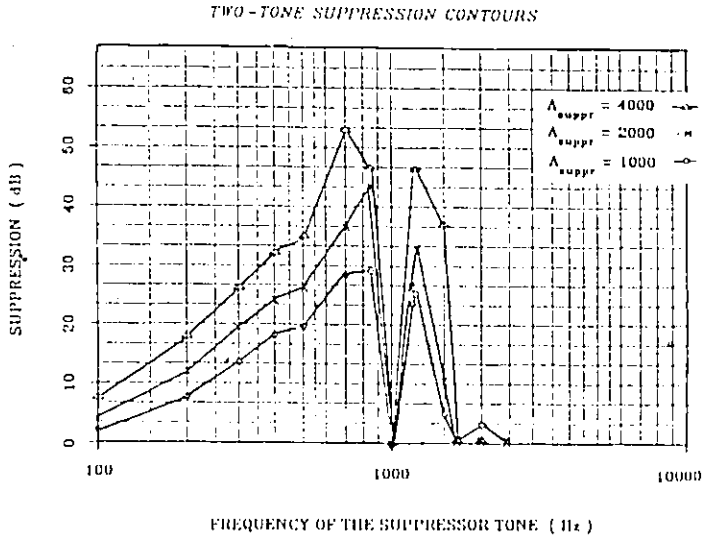

Figure 3.7 : Two−tone suppression contours measured at the output of the channel corresponding to f_c = 1000 Hz for different intensity levels of suppressor tones.

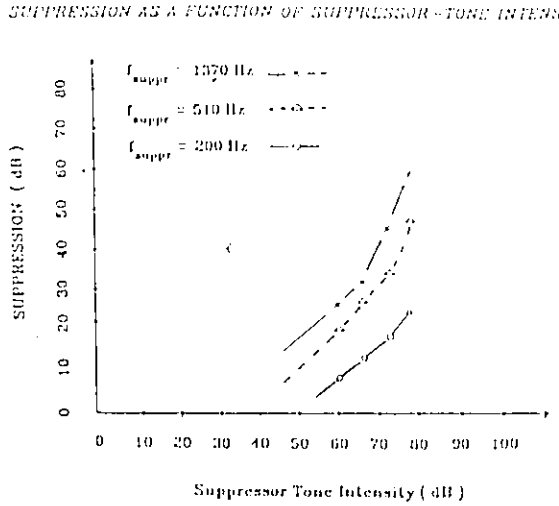

Figure 3.8 : Suppression behavior in one channel (f_c = 1.0 kHz) of the computational auditory model as a function of the suppressor tone intensity. The suppression produced by three different frequency suppressor tones are presented.

3.4 SUMMARY

We have modified Seneff's computational auditory model in order to use it as the front-end of a template-matching recognition system. The major problem encountered was that Seneff's model produced excessively strong lateral inhibition, which resulted in short-time spectral representations with sharp peaks. Since we are interested in reproducing lateral inhibition measured in mammalian ear because we believe it will optimize the extraction of perceptual cues, we have also tested the Seneff's computational auditory model with two-tone signals. The resulting two-tone suppression contours were different from those measured by Javel in his two-tone synchrony suppression experiments.

Three modifications have been made to the original version of Seneff's auditory model. The first modification consisted of changing the bandpass filter response in order to reproduce more precisely the BM tuning response. We have also added a sample-by-sample product of the output of pairs of adjacent channels to reduce the sensitivity of the GSD's to the frequency of single tones. The third modification consisted of adding a constant in the denominator of the GSD to reduce the formant peak splitting effect. Together, these modifications have permitted us to obtain smoothed short-time representations that we expect would contain perceptual information about speech sounds as it is presented to the higher auditory centers by the peripheral auditory system.

CHAPTER 4

SPECTROGRAPHIC COMPARISONS AND SPEECH RECOGNITION EXPERIMENTS

4.1 INTRODUCTION

In this chapter we present spectrograms of speech sounds obtained from the modified version of Seneff's computational auditory model and we compare them to spectrograms obtained from a Discrete Fourier Transform (DFT) analysis and a conventional filter-bank analysis for various speech conditions. This comparison provides us with some illustrations of the ability of the computational auditory model to emphasize formant peaks in spectral representations of speech sounds, its resistance to speech degradations such as additive noise, and its insensitivity to spectral tilt. As we said in chapter 1, such tilts, when caused by excitation differences, are a major source of speaker differences, and the model may therefore offer some advantage for speaker-independent speech recognition.

In order to verify that the computational auditory model does offer some advantages for recognition of degraded speech and speech from different speakers, we have compared its performance with that of the filter-bank representation in a cross-speaker digit recognition task. Since the filter-bank representation has

been compared with other commonly used representations and shown to be as good or better than them [12], the comparison between the two analyses should give to us an indication of the attractiveness of the auditory model as a front-end.

4.2 AUDITORY MODEL VERSUS FILTER-BANK REPRESENTATIONS

4.2.1 Extraction of spectral representations by a filter-bank analysis

The filter-bank representation of speech is obtained from an analysis which simulates that shown in figure 1.4 of chapter 1. That is, the speech signal sampled at 8 kHz is windowed into 204-points (25.6 ms.) every 6.4 ms. and an FFT is calculated. Each short-time FFT amplitude spectrum is multiplied by twenty overlapping triangular bandpass filter responses equally spaced on the technical mel scale of frequency. The energy in each channel is then converted to a log scale to obtain 20 log channel energies. In this work we will call this representation *LCE* for *log channel energies*, while we call the ones computed by the auditory model *GSD* for *generalized synchrony detector*.

4.2.2 Comparison of speech spectrograms extracted from different analyses

Figures 4.1 through 4.6 show speech spectrograms extracted from the three different analyses that were listed above for different speech degradations. Figure 4.1 a and b shows spectrograms of the word *zero* derived from a DFT analysis and from the computational auditory model. In examining the two spectrograms, we notice that the formant peaks are clearer in the spectrogram of the auditory model than in the other. We also notice that there are no pitch striations across time and also that valleys between formant peaks are somewhat smoother. Another important characteristic is that the third and fourth formants are inhibited in the GSD spectrogram while they are well represented in the DFT spectrogram.

Figure 4.2 shows the same spectrograms as in figure 4.1 except that white noise (SNR ~ 6 dB) has been added to the speech. The first and second formants show up in the GSD spectrogram, while they are almost indistinguishable in the DFT spectrogram. Moreover, the GSD spectrograms for the word *zero* in noise and without noise are quite similar, suggesting that the auditory model should offer good performance when clean speech templates are used to recognize speech in noise.

We may suppose that the auditory model will also be insensitive to noise having other characteristics than white noise, although it is to be expected that it will be sensitive to some periodic signals other than speech sounds. To verify this hypothesis, we have processed in-flight recordings in a *T*33 fighter/trainer and in a *Bell* 205 helicopter. GSD spectrograms of digit *zero* in such environment are

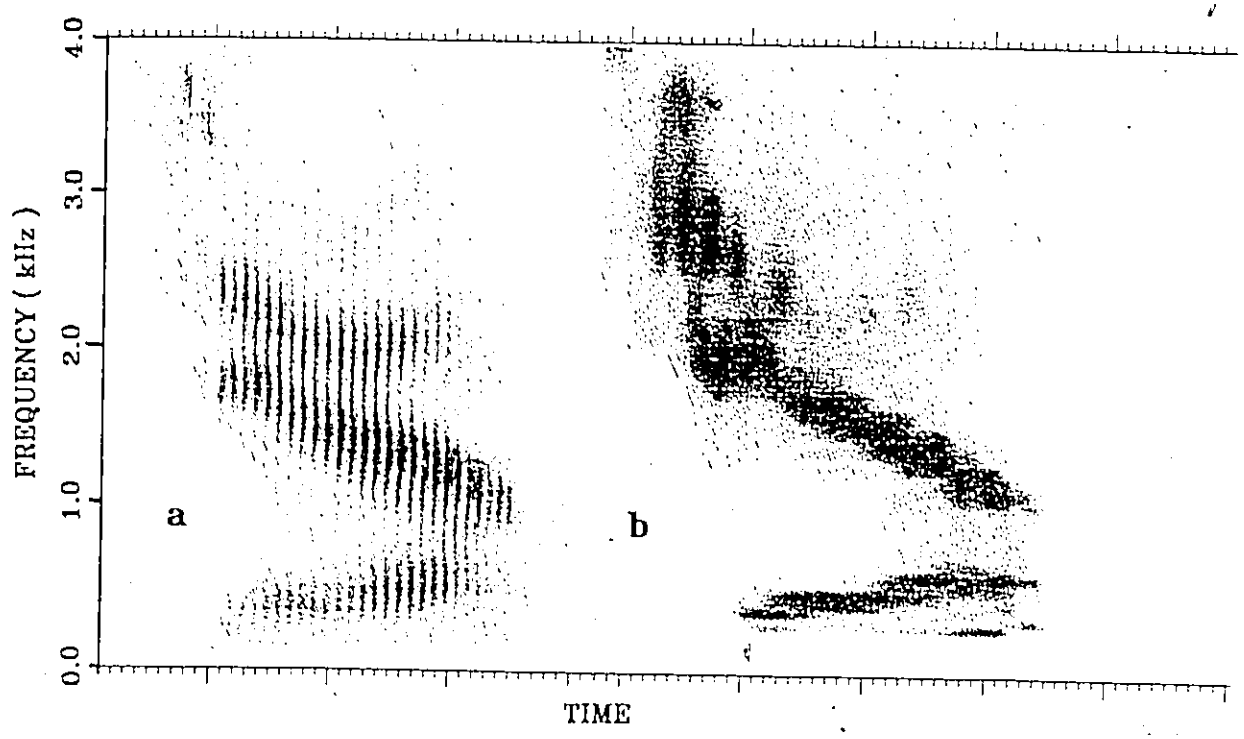

Figure 4.1 : Discerete Fourier transform spectrogram a) and GSD spectrogram b) for the word zero.

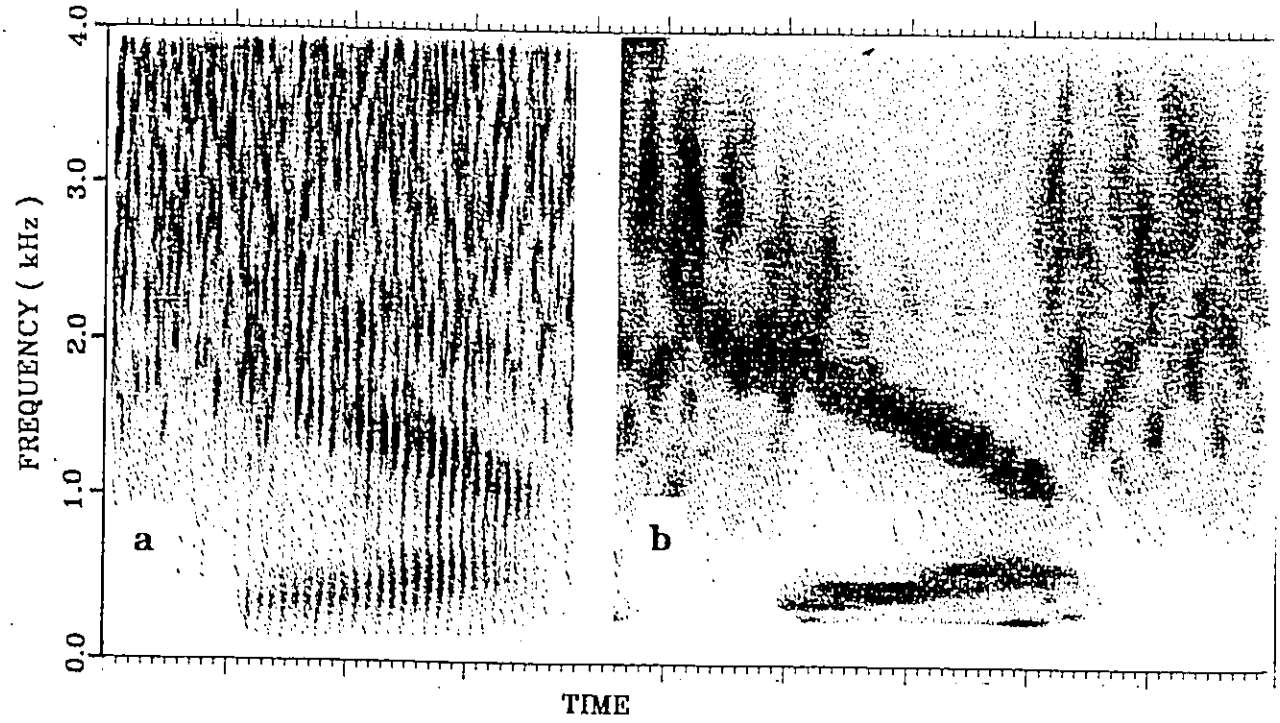

Figure 4.2 : Discerete Fourier transform spectrogram a) and GSD spectrogram b) for the word zero with added noise. The signal−to−noise ratio is 6 dB.

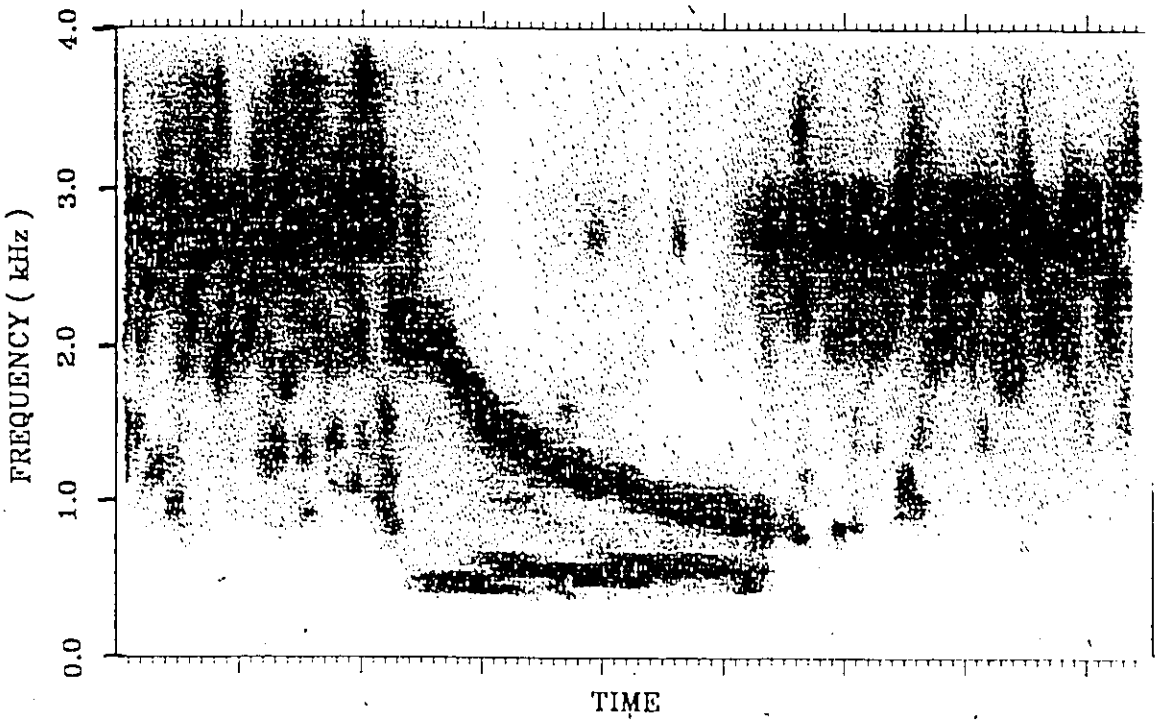

Figure 4.3 : GSD spectrogram of the digit zero recorded in an helicopter. The noise around 3.0 kHz is due to the gearbox whine. (The narrow horizontal line in center of the noise is an artifact of the printing process).

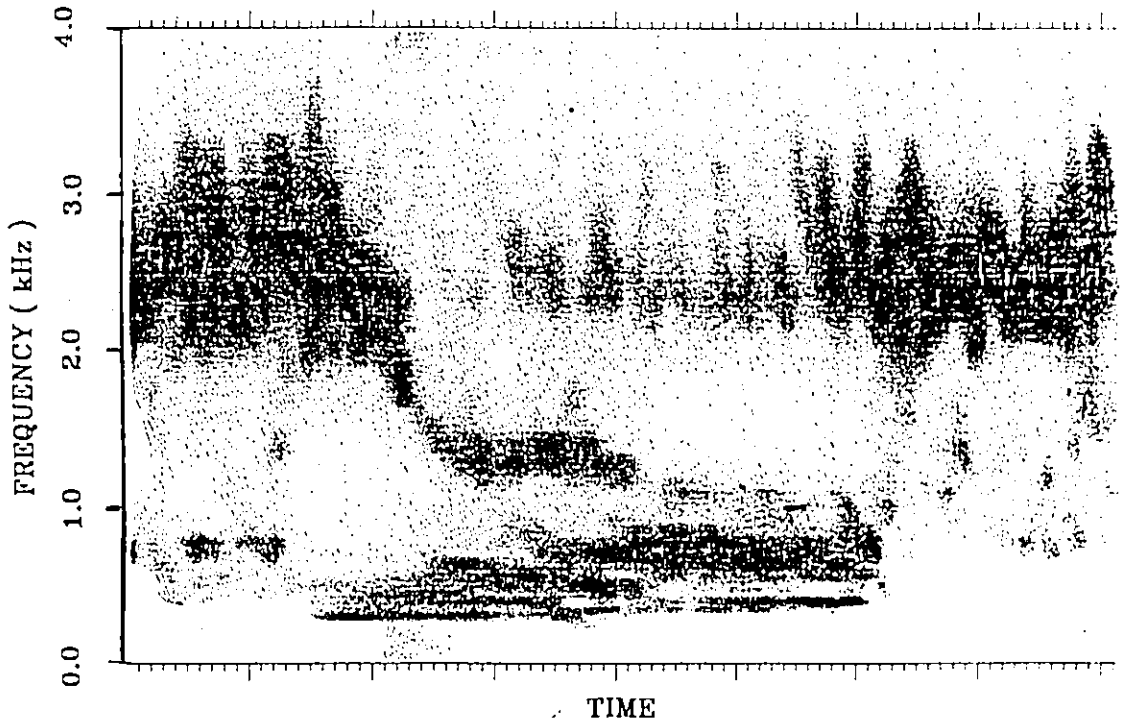

Figure 4.4 : GSD spectrogram of the digit zero recorded in a T33 airplane cockpit.

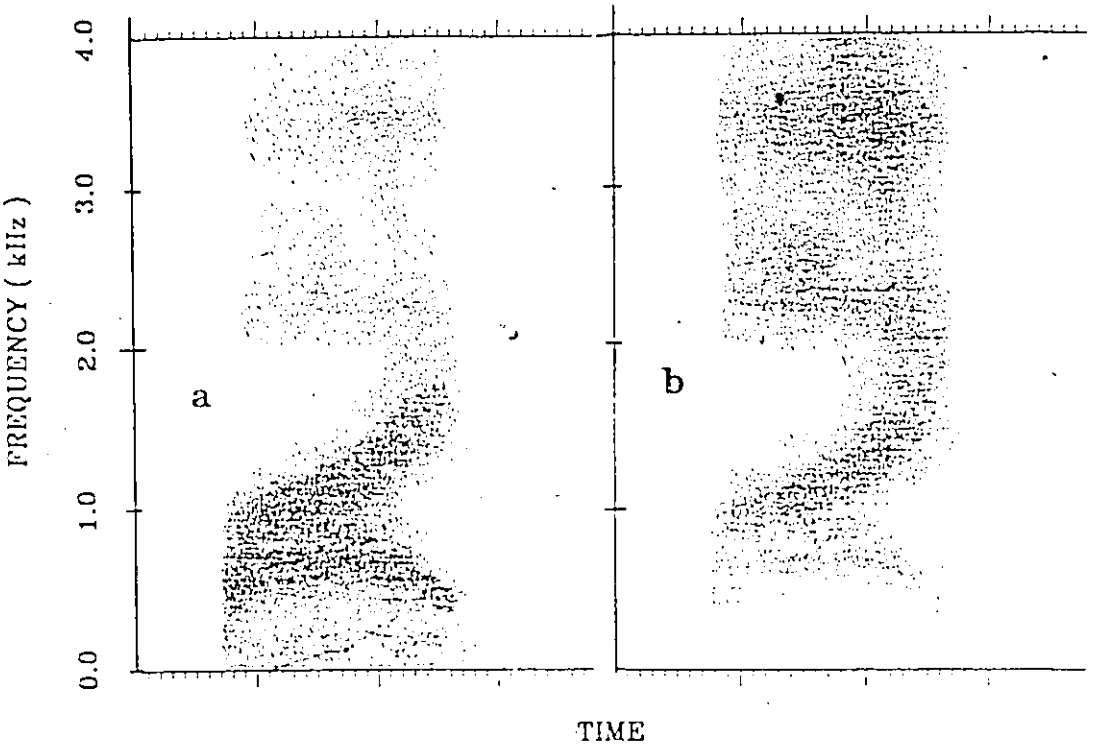

Figure 4.5 : LCE spectrograms of the undistorted a) and of the spectrally tilted (+6 dB/octave) b) word five.

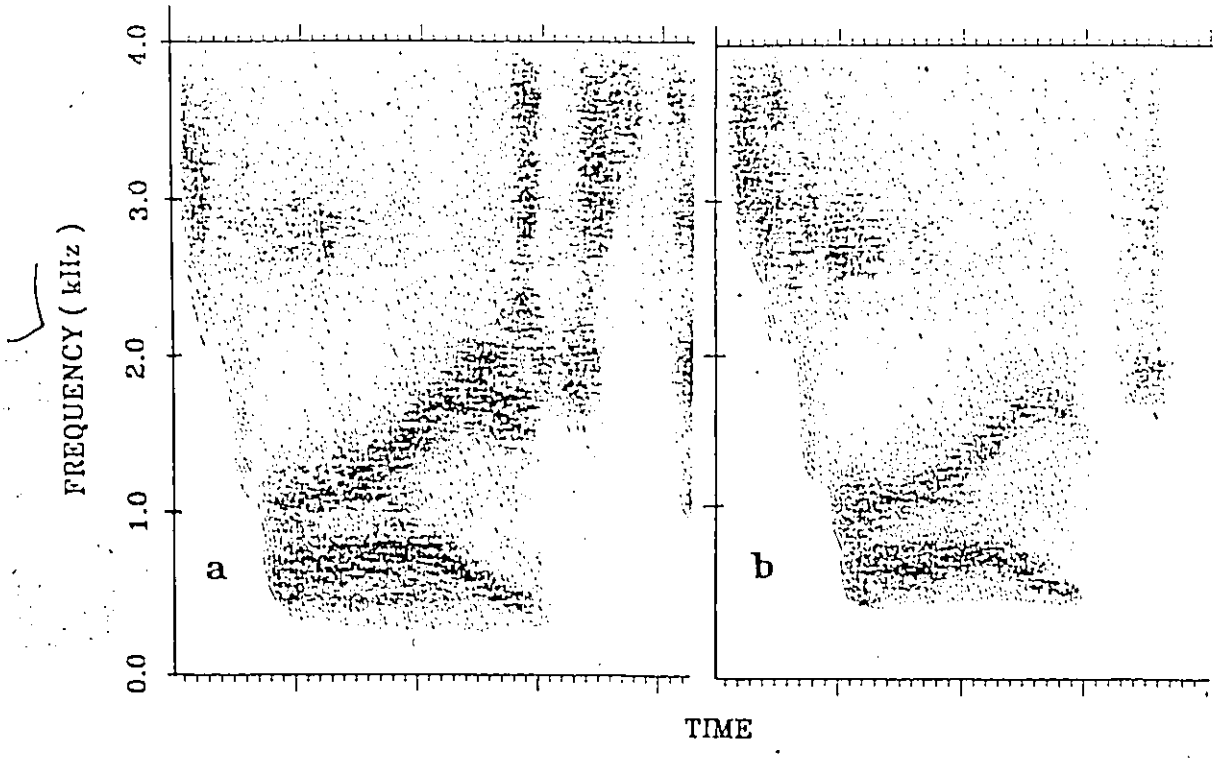

Figure 4.6 : GSD spectrograms of the undistorted a) and of the spectrally tilted (+6 dB/octave) b) word five.

presented in figure 4.3 and 4.4. Examination of figure 4.3 shows that the periodic interference at 3.0 kHz due to the gearbox whine is suppressed during speech. This may be explained by the fact that the first two formants inhibit the higher frequency parts of the spectrum.

On the other hand, if we examine the GSD spectrogram of the word *zero* recorded in T33, we see that the noise suppression is less effective. It appears that suppression produced by the auditory model is only effective in suppressing low intensity noise levels.

Figures 4.5 and 4.6 show spectrograms comparing an example of the word *five* in one case unmodified and in the other spectrally tilted at 6 dB/octave from both LCE and GSD analysis. The region of the first two formants in the GSD spectrogram shows little effect of tilting, while we clearly see the effect of the tilt in the LCE spectrograms. Thus, as we expected, the GSD analysis is to some extent insensitive to spectral tilts.

4.3 DIGIT RECOGNITION EXPERIMENTS

4.3.1 Reference and test data

The reference data used in the recognition experiment consists of two sets of ten digits with one example of each digit in each set. One was recorded in an anechoic chamber by a male speaker of British English. The other is synthetic speech produced by the text-to-speech system *MITalk* [3]. The test speech data

consists of digits spoken in connected groups of three by North American native English and French speakers, with 150 digits per speaker. The word boundaries in the reference and test data have been determined by hand.

Only 18 output channels of the filter-bank analysis are used in applying the recognition algorithm, while only 29 output channels computed from the auditory model are used. The bottom two channels in both methods are dropped because they contain information about the fundamental and this not useful for speech recognition. Because we cross-correlate adjacent channels, we dropped the top channel in the auditory model. Seven cepstrum coefficients are used in the frame-to-frame dissimilarity measure of the dynamic time-warping algorithm.

4.3.2 Recognition performance in clean speech

Table 4.1 shows the digit recognition error rate obtained for six male speakers and the two reference sets.

Digit Recognition Error Rate (%)				
	LCE analysis		CSD analysis	
Speaker	British English	Synthetic Voice	British English	Synthetic Voice
AA	15	57	16	48
CM	32	54	38	53
DS	10	21	10	35
GH	7	15	13	14
JT	12	42	25	45
ML	18	45	21	42
Average Error	17	41	21	40

Table 4.1 : Digit recognition error rate between six speakers and two templates consisting of digits pronounced by a British English speaker and of digits produced synthetically.

4.3.3 Recognition performance in noisy speech

The recognition performance for LCE and GSD representations of noisy speech has been tested with different speakers for different signal-to-noise ratios, by adding white noise to the test digits. The SNR is calculated by dividing the power of the speech signal by the power of the noise. Different SNR's are obtained by scaling the noise waveform. The reference digits are the British speaker set without degradation.

Two different tests have been carried out. In one we measured the digit error rates for five different speakers with SNR at 15 dB. The results are shown in Table 4.2. In the other, we measured the average error rate for two speakers (AA and DS) as a function of SNR. Curves representing the average error rate versus SNR for both the LCE and GSD analyses are shown in figure 4.7.

Digit Recognition Error Rate (%)		
Speaker	Speech added with noise	
	LCE analysis	GSD analysis
AA	41	21
DS	34	17
JT	40	48
ML	49	32
RO	48	39
Average Error	43	31

Table 4.2 : Digit recognition error rate in noise for the LCE and GSD analysis with five speakers. The SNR is 15 dB.

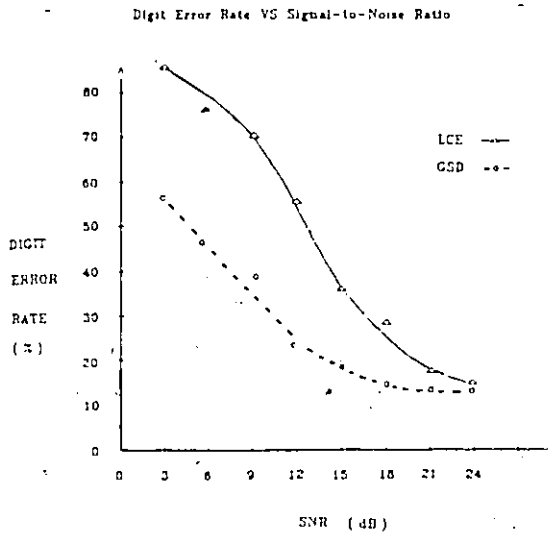

Figure 4.7 : Average digit recognition error rate of noisy speech for the LCE and GSD analysis. Noise at various power levels has been added to digits pronounced by two male speakers. The reference digits are the undistorted British English set.

4.3.4 Recognition performance of spectrally tilted speech

To compare the recognition performance of both the LCE and GSD analyses for linearly distorted speech, the signal representing the reference digits pronounced by the British speaker has been spectrally tilted by 6 dB/octave. It may not seem realistic to apply the linear distortion to the reference data rather than the test data, but we believe that the effect on recognition performance will be equivalent since what is important is the relationship between the test and reference data. We chose to apply the linear distortion to the reference digit because it involves much less processing. The digit error rates obtained for six test speakers are shown in table 4.3.

Digit Recognition Error Rate (%)		
Speaker	Spectrally Tilted Speech	
	LCE analysis	GSD analysis
AA	73	35
CM	79	52
DS	53	19
GH	21	15
JT	63	41
ML	71	43
Average error	60	34

Table 4.3 : Digit recognition error rate for spectrally tilted speech. Six test speakers are used. Template digits pronounced by a British English speaker are spectrally tilted.

4.4 DISCUSSION OF THE RESULTS

We have obtained an average digit error rate for six speakers of 17% for the LCE analysis and of 21% for the GSD analysis when using the British English reference templates. We may attribute the fact that the auditory model analysis does not perform as well as the conventional filter-bank analysis to the fact that the model deemphasizes onsets, offsets and rapid formant transitions in speech sounds. The error rates obtained when the synthetic voice is used as the template

do not show any advantage of using either one or the other method. The average error rate of about 40% obtained for both analyses just shows that synthetic speech matches poorly with naturally produced speech.

The recognition results of speech with added noise are much better with the auditory model than those with the LCE analysis when using undistorted reference templates. The plot of digit error rate versus SNR in figure 4.7 shows that the GSD analysis is better than the LCE analysis at every SNR level. Moreover, the error rate obtained with the LCE analysis starts increasing very rapidly for SNR's lower than 24 dB, while it stays about the same (~17%) for SNR's between 15 and 24 dB in the case of the GSD analysis. This result provides further confirmation of the fact that the lateral inhibition produced in the auditory model is an efficient way to suppress low levels of noise .

The recognition of spectrally tilted speech again shows the superior resistance of the auditory model to signal degradations compared to the LCE analysis. The average digit error rate of 60% obtained for the filter-bank method shows that a method based on a power spectrum analysis is sensitive to linearly distorted speech while a synchrony analysis is only slightly affected, although we must say that the digit error rate obtained with the GSD analysis is somewhat higher that we might have expected.

CHAPTER 5

CONCLUSION

5.1 DISCUSSION

From a fairly brief description in the proceedings of a conference [53], we have implemented a computational model of the auditory system. The filter-bank formulation proposed in the original version has been slightly modified to obtain a better representation of the tuning accomplished in the basilar membrane. Also an interpolating function has been implemented in order to obtain an accurate synchrony estimation of higher frequency bandpass-filtered signals.

Although the computational auditory model presents a kind of frequency masking, we have found by measuring its suppression behavior that it does not accord with that found in mammalian ears. The synchrony detection used in the model has been reformulated and a cross-channel correlation added to its output. In contrast to the earlier version, we have found that the new version reproduced the two-tone suppression behavior measured in mammals.

The new computational auditory model has been tested as the front-end acoustic analysis stage of an existing template-based automatic speech recognition system. Its performance has been compared with that of a conventional filter-bank representation in various cross-speaker experiments. On quiet

undistorted speech the model resulted in slightly worse performance than the conventional representation. However, when noise was added to the test speech, or when it was subjected to linear filtering resulting in a 6 dB/octave spectral tilt, the representation provided by the model showed a clear advantage over the conventional representation.

We are encouraged by the recognition results with this auditory model, since one of the major deficiencies of present-day speech recognition technology is the lack of robustness to noise and linear distortions compared with human listeners. We recognize, however that the robustness we have demonstrated in the model is still far below human performance, and, in addition, there is a need to achieve recognition performance on undegraded material that is at least as good as that obtained with conventional approaches before the model can be considered to be a practical alternative.

5.2 FURTHER RESEARCH

It is known that different representations of the response in the auditory nerve fibers to speech sounds are present in the neurons of the cochlear nucleus [41]. We might therefore consider that the synchronous analysis performed in the computational auditory model would correspond only to those neurons in the cochlear nucleus that extract cues related to steady voiced sounds or sounds with formants varying only slowly in time.

Other analyses are therefore needed to obtain representations describing completely all speech sounds. More particularly, we would need features

representing onsets and offsets of speech energy considering that the present model deemphasizes these features.

It is difficult to say if reproducing the lateral inhibition (two-tone suppression) measured in mammalian ears permits us to obtain a better representation of speech sounds. Since we mostly expect steady voiced sounds to be well represented by the synchronized analysis, recognition tests could be done with only these sounds instead of digits.

As we described in chapter 3, formant peak splitting due to harmonics of the excitation signal or occurring when two formants are close together seems to affect the recognition performance obtained with the model. Therefore, one improvement that could be made to the model would consist of using multiple GSD's for each channel of the computational auditory model in order to detect the dominant frequency in adjacent channels where a formant is present. Such an analysis would agree more with the concept that higher auditory centers analyze dominant frequencies because the responses in auditory nerve fibers are phase-locked to dominant frequencies [32]. Also recent research has found that a recognition system using dominant frequencies gives very low error rates for vowel sounds [6].

REFERENCES

[1] Allen J. B. and Sondhi M. M., "Cochlear Macromechanics - Time Domain Solution.", *J. Acoust. Soc. Am.*, vol. 66, pp. 123-132, Jan. 1979.

[2] Allen J. B., "Cochlear Modeling." *IEEE Acoust., Speech and Signal Processing Magazine*, vol. 2, p. 3-29, Jan. 1985.

[3] Allen J., Carlson R., Granstrom B., Hunnicutt S., Klatt D. and Pinsoni D., "*Conversion of Unrestricted English Text to Speech.*", Massachusetts Institute of Technology, Cambridge, MA, 1979

[4] Atlas L. E. and Hengky L. M., "Cross-channel Correlation for the Enhanecement of Noisy Speech", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, ICASSP-85, pp. 724-727, March 1985.

[5] Bellman R. E., *Dynamic Programming*, Princeton, N.J., Princeton University Press, 1975

[6] Blomberg M., Carlson R., Elenius K. and Granstrom B., "Experiments with auditory models in speech recognition." in : *The representation of speech in the peripheral auditory system*, edited by R. Carlson and B. Granstrom, Elsevier Biomedical Press, pp. 197-201, 1982.

[7] Bridle J. S., Chamberlain R. M. and Brown M. D., "An Algorithm for Connected Word Recognition.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, ICASSP-82, pp. 899-902, May 1982.

[8] Carlson R. and Granstrom B., "Towards an Auditory Spectrograph." in : *The representation of speech in the peripheral auditory system*, edited by R.

Carlson and B. Granstrom, Elsevier Biomedical Press, pp. 109-114, 1982.

[9] Cole R. A., Zue V. W., "Experiments on Spectrogram Reading.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-79, pp. 116-119, April 1979.

[10] Cooke M. P., *A Computer Model of Peripheral Auditory Processing,* National Physical Laboratory, Teddigton, Middlesex, U.K., May 1985.

[11] Dautrich B. A., Rabiner L. R. and Martin T. B., "On the Use of Filter Bank Features for Isolated Word Recognition.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-83, pp. 1061-1064, April 1983.

[12] Davis S. B. and Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.", *IEEE Trans. Acoust., Speech and Signal Processing,* vol. ASSP-28, pp. 357-366, Aug 1980.

[13] DeBoer E., "On Active and Passive Cochlear Models - Toward a Generalized Analysis.", *J. Acoust. Soc. Am.*, vol. 73, pp. 574-576, Feb. 1983.

[14] Dolmazon J. M., Bastet L. and Shupljakov V. S., "A Functional Model of Peripheral Auditory System in Speech Processing.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-77, pp. 261-264, May 1977.

[15] Dolmazon J. M., "Representation of Speech-Like Sounds in the Peripheral Auditory System in Light of a Model." in : *The representation of speech in the peripheral auditory system,* edited by R. Carlson and B. Granstrom, Elsevier Biomedical Press, pp. 151-163, 1982.

[16] Elinius K. and Blomberg M., "Effects of Emphasizing Transitional or Stationary Parts of the Speech Signal in a Discrete Utterance Recognition System.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,*

ICASSP-82, pp. 535-537, May 1982.

[17] Escudier P. and Schwartz J. L., "Pulsation Threshold Patterns of Synthetic Vowels : Study of the Second Formant Emergence and the "Center of Gravity" Effects.", *Speech Communication,* vol. 4, pp. 189-198, Aug. 1985.

[18] Flanagan J. L., *Speech Analysis, Synthesis and Perception,* Springer-Verlag, 2nd edition, 1972.

[19] Fletcher H., *Speech and Hearing,* Van Nostrand, 1929

[20] Goldhor R., "A Speech Signal Processing System Based on a Peripheral Auditory Model.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-83, pp. 1368-1371, April 1983.

[21] Goldstein J. L. and Srulovicz P., "A Central Spectrum Model: a Synthesis of Auditory-nerve Timing and Place Cues in Monaural Communication of Frequency Spectrum.", *J. Acoust. Soc. Am.,* vol. 73, pp. 1266-1276, April 1983.

[22] Holmes J. N. and Segwick N. C., "Noise Compensation for Speech Recognition using probabilistic models.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-86, pp. 741-744, April 1986.

[23] Hunt M. J., Personal communication, 1986

[24] Hunt M. J., "Delayed Decisions in Speech Recognition - the Case of Formants.", *British Pattern Recognition Association,* Third International Conference, Sept. 1985.

[25] Hunt M. J. and Swail C. P., *Visa : A Display Program for Speech and Other Signals* Laboratory Technical Report, National Aeronautical Establishment, National Research Council of Canada, Dec. 1985.

[26] Hunt M. J. and Harvenberg C. E., *A Set of Sofware Tools for Signal Processing,* Laboratory Technical Report, National Aeronautical Establishment,

National Research Council of Canada, Aug. 1985.

[27] Hunt M. J., "The Speech Signal.", Proc. NATO AGARD Lecture Series No. 129, *Speech Processing*, pp. 2.1-2.12, May 1983.

[28] Hunt M. J., "Speaker Differences in Speech and Speaker Recognition.", Proc. NATO AGARD Lecture Series No. 129, *Speech Processing*, pp. 4.1-4.12, May 1983.

[29] Hunt M. J., "Time Alignment of Natural Speech to Synthetic Speech.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, ICASSP-84, pp. 2.5.1-2.5.4, March 1984.

[30] Hunt M. J. and Lefebvre C., "Speech Recognition Using a Cochlear Model.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, ICASSP-86, pp. 1979-1982, April 1986.

[31] Javel E., "Suppression of auditory Nerve Responses: Temporal Analysis, Intensity Effects and Suppression Contours.", *J. Acoust. Soc. Am.*, vol. 69, pp. 1386-1391, June 1981.

[32] Johnson D. H., "The Relationship Between Spike Rate and Synchrony in Responses of Auditory Fibers to Single Tones.", *J. Acoust. Soc. Am.*, vol. 68, pp. 1115-1122, Oct. 1980.

[33] Kernighan B. W. and Pike R., *The Unix Programming Environment* Prentice-Hall Software Series,Englewood Cliffs, N.J., 1984.

[34] Klatt D. H., "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, ICASSP-82, pp. 1278-1281, May 1982.

[35] Klatt D. H., "Speech processing strategies based on auditory models" in : *The representation of speech in the peripheral auditory system*, edited by R. Carlson and B. Granstrom, Elsevier Biomedical Press, pp. 181-196, 1982.

[36] Lowerre B. T., "Dynamic Speaker Adaptation in the Harpy Speech Recognition System.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-77, pp. 788-790, May 1977.

[37] Lyon R. F., "A Computational Model of Filtering, Detection, and Compression in the Cochlea.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-82, pp. 1282-1285, May 1982.

[38] Moore R. K., "Techniques for Automatic Speech Recognition.", Proc. NATO AGARD Lecture Series No. 129, *Speech Processing,* pp. 4.1-4.12, May 1983.

[39] Patterson R. D., "Auditory filter shapes derived with noise stimuli.", *J. Acoust. Soc. Am.,* vol. 59, 640-654, March 1976.

[40] Pfeiffer R. P., "A Model for Two-tone Inhibition of Single Cochlear-Nerve Fibers.", *J. Acoust. Soc. Am.,* vol. 48, pp. 1373-1378, Dec. 1970.

[41] Pickles J. O., *Introduction to the Physiology of Hearing,* Academic Press, Inc., (London) Ltd, 1982.

[42] Pinsoni D. B., "Speech Perception: Somes new Directions in Research and Theory.", *J. Acoust. Soc. Am.,* vol. 78, pp. 381-388, July 1985.

[43] Rabiner L. R. and Levinson S. E., "Isolated and Connected Word recognition - Theory and Selected Applications.", *IEEE Trans. on Comm.,* vol. COM-29, pp. 621-659, May 1981.

[44] Rabiner L. R., Levinson S. E. and Sondhi M. M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition.", *Bell System Tech. J.,* vol. 62 , pp. 1035-1074, April 1983

[45] Rabiner L. R. and Schafer R. W., *Digital Processing of Speech Signals,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.

[46] Rabiner L. R., and Soong F. K., "Single-Frame Vowel Recognition Using Vector Quantization with Several Distance Measures"., *Bell System Tech. J.*, vol. 64., pp. 2319-2330, April 1983

[47] Sachs M. B. and Young E. D., "Effects of Nonlinearities of Speech Encoding in the Auditory Nerve.", *J. Acoust. Soc. Am.*, vol. 68, pp. 858-875, Sept. 1980.

[48] Sachs M. B., Young E. D. and Miller M. I., "Encoding of Speech Features in the Auditory Nerve." in : *The representation of speech in the peripheral auditory system,* edited by R. Carlson and B. Granstrom, Elsevier Biomedical Press, pp. 115-130, 1982.

[49] Sakoe H. and Chiba S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition.", *IEEE Trans. Acoust., Speech and Signal Processing,* vol. ASSP-26, pp. 43-49, Feb. 1978.

[50] Sankoff D. and Kruskal J. B., *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison,* Addison-Wesley, Inc., Reading , MA., 1983.

[51] Schroeder M. R., "Linear Predictive Coding of Speech: Review and Current Directions.", *IEEE Comm. Magazine,* vol. 23, pp. 54-61, Aug. 1985.

[52] Searle C.L., Jacobson J. Z. and Rayment S.G., "Stop Consonant Discrimination based on Human Audition.", *J. Acoust. Soc. Am.*, vol. 65, pp. 799-809, March 1979.

[53] Seneff S., "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model.", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* ICASSP-84, pp. 36.2.1-36.2.4, March 1984.

[54] Seneff S., *Pitch and Spectral Estimation of Speech Based on an Auditory Synchrony Model* PhD Thesis, Dept. of Electrical Engineering, M.I.T., Cam-

bridge, MA.

[55] Warren R. M., *Auditory perception: A New Synthesis,* Pergamon Press Inc., Elmsford, N.Y., 1982.

[56] White G. M., "Speech Recognition: An Idea Whose Time is Coming.", *Byte Magazine,* pp. 213-222, Jan. 1984.

[57] White G. M. and Neely R. B., "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming.", *IEEE Trans. Acoust., Speech and Signal Processing,* vol. ASSP-24, pp. 188-193, April 1976.

[58] Zue V. W., "The Use of Speech Knowledge in Automatic Speech Recognition.", *Proceedings of IEEE,* vol. 73, pp. 1602-1615, Nov. 1985.

[59] Zwicker E. and Terhardt E., "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency.", *J. Acoust. Soc. Am.,* vol. 68, pp. 1523-1525, Nov. 1980.

APPENDIX 1 - BANDPASS FILTER DESIGN

A-1.1 Overview

The bandpass filters have been set to obtain the following response:

1. The shape of each bandpass filter response should be similar to that observed in the basilar membrane. That is, the slope on rise should be small, while the slope on fall should be very steep. Also, the roll-off on fall should be fast.
2. Bandpass filter bandwidths should be set at the 3 dB points to agree with psychophysical tuning curves measured by Zwicker [59] for which the specifications are given in table A-1.1. The filters are spaced by half a bark.
3. The bandpass response of each filter should be adjusted to obtain two-tone suppression contours for each channel of the auditory model that accord with Javel's measurements [31].

To obtain the first two specifications presented above, we have used the filter-bank structure proposed by Lyon [37]. It consists of cascaded notch filters in parallel with resonators. We have added to each notch filter a 12'th order lowpass butterworth filter to obtain a steep slope ($>$ 96 dB/octave) on fall of the bandpass response.

Each of the three filters can be represented by an analog transfer function. The transfer function corresponding to the resonator is :

$$H_r(s) = \frac{fs}{s^2 + 2ds + (d^2 + e^2)} \qquad \text{(A-1.1)}$$

The coefficient f is a scaling factor, which is set to obtain the desired overall amplitude response of the bandpass filter. Coefficients d and e are set to obtain the center of each bandpass filter response at approximately the specified angular center frequency and to obtain an attenuation of 3 dB at the specified lower cut-off frequency multiplied by 2π and of 2.5 dB at the specified upper cut-off frequency multiplied by 2π.

The transfer function corresponding to the notch filter is :

$$H_n(s) = \frac{s^2 + 2bs + (b^2 + c^2)}{s^2 + 2as + (a^2 + c^2)} \qquad \text{(A-1.2)}$$

The value of coefficients a and b are small compared with that of c. They are fixed to obtain a notch attenuation of -40 dB. This corresponds approximately to the notch attenuation shown in the bandpass responses of the BM in chapter 2. The coefficient c is set to the frequency of the notch and is calculated to obtain an attenuation of approximately of 0.4 dB at the upper cut-off frequency multiplied by 2π of the given channel.

The transfer function corresponding to the 12'th-order ($N = 12$) lowpass butterworth filter is :

$$|H_l(s)|^2 = \frac{1}{1 + (s/jw_b)^{2N}} \qquad \text{(A-1.3)}$$

The coefficients in the butterworth filter equation are set to obtain an attenuation of approximately 0.1 dB at the upper cut-off frequency multiplied by 2π (w_b is the angular cut-off frequency of the butterworth filter).

A-1.2 Conversion of the analog filter to digital filter

Each analog filter has been converted into a digital filter using the bilinear transformation. Since this transformation introduces a non-linear shift in the frequency mapping between the s and z planes, the center and cutoff frequencies specified in table A-1.1 have been pre-warped before using the bilinear transformation. Although the filter specifications have been pre-warped in this way, the non-linear shift in the frequency response due to the bilinear transformation will still accentuate the slope on the high-frequency side of a bandpass response. Such an effect is advantageous, since we desire to model the BM by obtaining as steep a slope as possible.

The step-by-step procedure to calculate the coefficients of the digital filter equation from the psychophysical data of table A-1.1 is :

1. Prewarp the lower cut-off (f_l), upper cut-off (f_u) and center (f_c) frequency values using the equation :

$$f = \frac{SF}{\pi} \tan\left\{\frac{\pi f_c}{SF}\right\} \qquad \text{(A-1.4)}$$

where SF is the sampling frequency.

2. Determine the coefficients (α_1, β_1,... etc.) of the analog filter equation for each filter :

$$H(s) = \frac{\alpha_1 s^2 + \alpha_2 s + \alpha_3}{\beta_1 s^2 + \beta_2 s + \beta_3} \qquad \text{(A-1.5)}$$

3. Convert the transfer function of each analog filter to a digital filter formulation using :

$$s = \frac{2(1 - z^{-1})}{T(1 + z^{-1})} \qquad \text{(A-1.6)}$$

4. The filter equation of the resonator and the notch filter is represented by a second-order digital filter equation. The 12'th-order lowpass butterworth consists of six cascaded second-order digital filter equations. A second order digital filter equation is represented by :

$$y[n] = a_1 x[n] + a_2 x[n-1] + a_3 x[n-2] + a_4 y[n-1] + a_5 y[n-2] \qquad \text{(A-1.7)}$$

Table A-1.1 : Bandpass filter specifications

Filter index c	Center frequency (Hz) f_c	Lower cut-off frequency (Hz) f_l	Upper cut-off frequency (Hz) f_u
1	3400	3150	3700
2	3150	2870	3370
3	2900	2700	3150
4	2700	2505	2920
5	2500	2320	2700
6	2325	2160	2510
7	2150	2000	2320
8	2000	1860	2160
9	1850	1720	2000
10	1725	1600	1860
11	1600	1480	1720
12	1485	1375	1600
13	1370	1270	1480
14	1270	1175	1375
15	1170	1080	1270
16	1085	1000	1175
17	1000	920	1080
18	920	845	1000
19	840	770	920
20	770	700	845
21	700	630	770
22	635	570	700
23	570	515	635
24	510	455	570
25	450	400	510
26	400	350	455
27	350	300	400
28	300	250	350
29	250	200	300
30	200	150	250
31	150	100	200
32	100	50	150

APPENDIX 2 - NATURE AND ORIGIN OF SOFTWARE AND SPEECH DATA

A-2.1 Software

Various computer programs developed by different persons and by the author have been used for the investigation of the speech recognition front-end based on the auditory system. The main function of these programs is to simulate the auditory model, the conventional filter-bank analysis and the recognition algorithm. There are also some utility programs that have been used to set parameters of the auditory model and to display its output. In the following paragraphs, we give a brief description of these programs.

Five computer programs, all written in *C*, are used to simulate the preemphasis stage, the filtering stage, the compression stage, the synchrony analysis stage and a raised-cosine window stage of the auditory model. The program simulating the first stage is taken from the digital signal processing package [26] developed by C. E. Harvenberg and others at NRC. The four other programs have been developed by the author. A description of the implementation of these programs is given in appendix 3.

A utility program has also been developed by the author to allow the parameters of the filter-bank analysis to be adjusted. The user specifies the attenuation due to the lowpass butterworth, notch and resonator filters at the

upper cut-off frequency in the bandpass response of each channel. The program shows on the terminal screen the bandpass filter of each channel as a function of frequency and creates a new data file of digital filter coefficients, which is used during the compilation of the bandpass filtering program.

The speech recognition front-end based on a log channel energy (*LCE*) analysis and the speech recognition algorithm accomplishing the cepstrum coefficients transformation and the dynamic time-warping of spectral patterns are *fortran* programs developed by M. J. Hunt at BNR under a DND contract. In order to carry out recognition experiments with data from the auditory model, the recognition program had to be converted by the author to handle the 32 channels of the auditory model instead of 20 the channels of the LCE analysis. A C version of the speech recognition program has also been produced by the author.

Software for displaying speech waveforms and spectrograms has been implemented by C. P. Swail and others [25]. Different computer programs can be used to obtain discrete Fourier transform-based, LCE-based and GSD-based spectrograms. In these spectrograms intensity is represented by colors on a "heat" scale. white corresponds to highest intensity, followed by continuous sequence running through yellow, orange, black and finally blue as the color background.

A-2.2 Speech data

The speech data used in the speech recognition experiment has been taken from various sources. The test speech data was recorded and digitized at 8 kHz and the word boundaries were marked by M. J. Hunt at BNR under a DND contract. The British English reference speech data was spoken by M. J. Hunt in an anechoic chamber at NRC. It was digitized at 10 kHz and the word boundaries were determined by M. Baranowski. The data has been converted to 8 kHz samples using one of the utility programs of the digital signal processing package [26].

In the speech recognition experiments, we have also used a synthetic voice as a source of reference speech. The synthetic reference digits are produced by a *pascal* version of the *MITalk* system developed by J. Allen [3] and others at MIT and converted to pascal by A. Rudnicky at CMU.

APPENDIX 3 - SOFTWARE IMPLEMENTATION DETAILS OF THE AUDITORY MODEL

A-3.1 Overview

As we described in appendix 2, five computer programs written in C are used to implement the auditory model. These programs have been written to take advantage of the function *pipe* (|) of the UNIX operating system [33].

By means of pipe we can divide a huge computer program into different small programs without increasing the burden of writing and reading data into explicit intermediate files. This makes debugging the programs easier because each program is small. It also makes it easy to examine the output at any intermediate stage.

In this way the following statement causes execution of the complete model :

$$delt\ 0.5 < inputfile \mid filt8 \mid comp8 \mid gsd8 \mid raiscos8 > outputfile$$

where *delt* 0.5 is the preemphasis stage with a leak factor of 0.5, *filt*8 is the filterbank analysis, *comp*8 is the non-linear compressing stage, *gsd*8 performs the rectification, interpolation and synchrony analysis and *raiscos*8. is the raised-cosine window stage. The sign "<" and ">" specify an input and an output file

UNIX is a Trademark of AT&T Bell Laboratories.

respectively, while the pipeline sign specifies that the output of one stage is to be the input of the next stage.

Although the pipe function of the Unix system links the output of one stage to the input of the next one, each program must take the record structure of the data into account. The input, output and intermediate stages represent data in binary rather than character form. The input file consists of sequence of short-integer (16-bit) numbers representing a speech waveform sampled at 8 kHz. The output of the preemphasis stage has the same form. The data format for the other stages is much more complicated since we must process data in different channels. The filter-bank program reads a block of 204 short-integer samples and outputs consecutively 32 blocks of 204 long-integer (32-bit) numbers corresponding to the data processed in each channel starting with the highest frequency channel. This record structure is preserved between all stages in the auditory model up to the last (*raiscos8*) stage, where data is output in records of 32 short-integer numbers with one such record being output for every 51 records at the input of the model.