UNIVERSITY OF CALIFORNIA

Los Angeles

# A Model of Dynamic Auditory Perception and its Application to Robust Speech Recognition

A thesis submitted in partial satisfaction of the

requirements for the degree Master of Science

in Electrical Engineering

by

**Brian P. Strope**

1995

The thesis of Brian P. Strope is approved.

Greg Pottie

John Villasenor

Abeer Alwan, Committee Chair

University of California, Los Angeles

1995

# Dedication

For those with enough patience to read it.

# Table of Contents

# List of Figures

# List of Tables

# ACKNOWLEDGMENTS

My sincere thanks go to the subjects who unselfishly gave their time to help me and this work. Further, I thankfully acknowledge my Advisor, Prof. Abeer Alwan, for stressing the importance of exacting, careful research, and for critically challenging each step of this effort. In addition, I appreciate Prof. John Villasenor's, clear, enthusiastic instruction of signal processing, which provided me fundamental concepts useful for this project. I also acknowledge Richard Lyon of Cal. Tech. and Apple Computer, whose papers and discussion provided much of the original motivation for this effort. This work was supported in part by NIH-NIDCD Grant No. 1 R29 DC 02033-01A1.

## ABSTRACT OF THE THESIS

# A Model of Dynamic Auditory Perception and its Application to Robust Speech Recognition


by

Brian Strope

Master of Science in Electrical Engineering
University of California, Los Angeles, 1995
Professor Abeer Alwan, Chair



A model that predicts dynamic auditory perception is presented. Physiological and psychophysical results suggest the qualitative model for the structure presented, a simple set of perceptual forward masking experiments determine quantitative model parameters. The model uses a parallel filter bank combined with carefully parameterized logarithmic automatic gain control on each band to model adaptation in human audition. This structure re-creates several measurable aspects of dynamic, or time varying, audition including: emphasis of onsets and transitions, level-dependent effective filter shapes, a relatively small short-term dynamic range, and reduced sensitivity to slowly varying channel characteristics. This model, as the front end for DTW-based speech recognition, together with a perceptually motivated spectral distance measure, improves system resilience to background noise by 5-10 dB, when compared to LP-cepstral representations using weighted-cepstral distance measures.

# Chapter 1

# Introduction

## 1.1 Psychophysical and Physiological Auditory Modeling

Fletcher [1940], and later Zwicker *et al.*, [1957, 1980] quantified the frequency-selectivity of the human auditory system through a series of static perceptual masking experiments. These observations and resulting models, which define the critical bandwidths of effective auditory filters, continue to provide significant insight for speech perception models, and perceptually-based speech signal processing.

Bekesy [1953, 1960] directly observed the mechanical operation of the cochlea, and how the non-uniform basilar membrane leads to mechanical frequency selectivity (and won the Nobel Prize). More recently, precise measuring techniques have shown that cochlear mechanics, and therefore the frequency selectivity of the basilar membrane, is extremely active and fragile, adapting to the input waveform and deteriorating quickly with the health of the specimen [Sellick *et al.*, 1982] [Ashmore, 1987].

In addition to static (or simultaneous) masking, where one static sound masks the perception of another, researchers also measure dynamic (or forward) masking, where a loud sound masks the perception of a following quiet sound [Plomb, 1964], [Jesteadt *et al.*, 1982]. Further, perceptual experiments with speech stimuli show that our auditory system is acutely sensitive to transitions and onsets in speech [Furui, 1986]. Together with evidence of an adapting, active cochlea, these experiments firmly support a dynamic model of speech perception.

Although there are exceptions (e.g. Hermansky *et al.,* [1992], Ghitza [1991], Seneff [1990], Lyon [1988, 1982]), the predominant model of hearing used for speech recognition directly incorporates static critical-bandwidth frequency-selectivity and does not include an explicit model of dynamic auditory perception.

The critical bandwidth model of auditory frequency selectivity predicts static auditory masking, provides significant insight into auditory perception, and has important implications for speech processing applications. Following this example, a well-quantified model of dynamic auditory adaptation, which predicts dynamic auditory masking, should provide additional insight into auditory perception, and may similarly improve speech processing applications. This thesis quantifies a dynamic auditory model which predicts both static and dynamic auditory masking, and evaluates the model in a simple speech recognition system.

## 1.2 Speech Recognition Overview

Most speech recognition systems divide the recognition problem into two parts: a short-term spectral estimation, followed by a comparison of a collection of short-term spectral estimations with previously observed spectral estimations. In some sense, the short-term spectral estimation models our ears, changing pressure waves into an internal time-frequency representation, and the comparison of these spectral estimations models higher-level brain functions, comparing the sequence of internal representations to previously observed sequences associated with some meaning. A single short-term spectral estimation is often called a feature vector, and a collection of them ordered in time is the observation sequence. Figure 1.1 shows a block diagram for this overview.

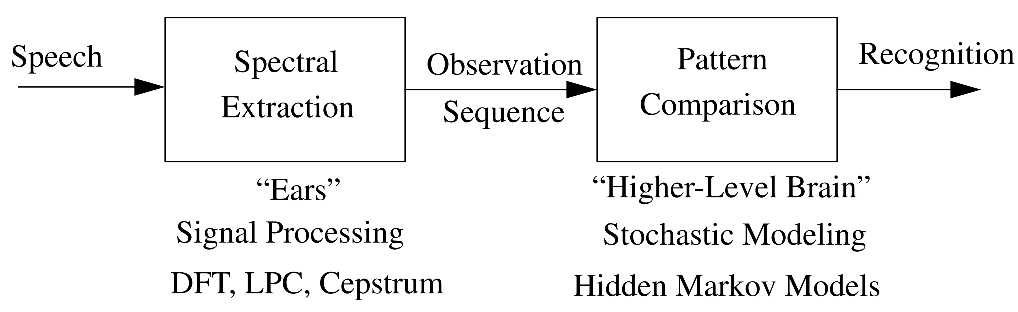


Figure 1.1 Speech Recognition Overview

Several signal processing techniques have been used for spectral extraction for speech recognition, although currently most systems use some type of cepstral representation [Rabiner and Juang 1993]. There are computationally efficient algorithms to obtain cepstral representations, these representations have relatively few elements per feature vector, and they retain a subset of the available acoustic information which has proven to be suitable for recognition. Further, there are techniques to incorporate aspects of static auditory perception in the cepstral representation, which have improved recognition performance.

There are several common approaches for pattern comparison. A dynamic programming technique provides a deterministic solution to measure the best alignment, and thereby, the closest match between an observation sequence and a set of stored template observation sequences, subject to imposed alignment constraints. Neural Networks have been used in different parts of pattern comparison algorithms, making the comparison processes more flexible and trainable.

However, recent advances in the application of stochastic modeling, using Hidden Markov Models, have greatly improved speech recognition systems' performance [Rabiner and Juang 1993]. These techniques provide a stochastic framework for rigorous statistical training from observation sequences. Applications of HMMs have successfully lead to speaker-independent, connected-speech, and with higher-level grammar and context models, large vocabulary recognition systems [Lee *et al.*, 1991].

## 1.3 The Problem and Proposed Solution

Despite advances from stochastic modeling in the areas of connected speech, speaker-independence, and increased vocabulary size, recognition systems typically do not work well with even slight changes in the acoustic environment. Different background noise from that present in the training data, or even a different microphone can greatly reduce recognition performance. Without a robust solution, wide-spread speech recognition applications do not yet exist.

We assume speech recognition systems will be more robust if they use spectral estimations more consistent with those estimated by the human auditory system. A well-quantified model of short-term auditory dynamics provides the next level of modeling detail after critical band frequency-selectivity.

Although several models incorporating auditory adaptation have been proposed [Hermansky *et al.,* 1992, Kates, 1991, Seneff, 1990, Lyon, 1988 and 1982, Goldhor, 1985], choosing parameters for the adaptation remains a consistent challenge. In perhaps the most complete derivation of these, Seneff [1990], using the adaptation structure from Goldhor [1985], chooses adaptation parameters for a specific (and single) stage of the auditory system based on detailed physiological data. By definition, this approach neglects significant adaptation at higher-levels of the auditory system. Others typically use parameters either loosely motivated by perceptual or physiological observations, or simply those parameters which provide good results for the specific application.

The significant contribution of this work is the ‘closed-loop’ quantification of a dynamic auditory model. Specifically, the shape of the I/O curves and the logarithmic adaptation time constants are obtained through a complete set of perceptual forward masking experiments.

## 1.4 Thesis Overview

This thesis relies on evidence from cochlear mechanics and psychoacoustics to develop a relatively simple time-varying model which captures several prominent auditory phenomena. The model is evaluated as the front end for a speech recognition system. Original forward masking experiments across a wide range of frequencies, levels, and delays provide model parameters.

Chapter 2 describes the physiological and perceptual motivation for the dynamic model and includes a qualitative description of its implementation and functionality. Chapter 3 summarizes the perceptual experiments and results which lead to model parameters. Chapter 4 translates the experimental results into model parameters. Chapter 5 is a description of the speech recognition system. Finally, Chapter 6 discusses incorporating the model as the front end to the recognizer, and the impact the model has on improving recognition performance in noisy environments. The summary includes discussions of other applications of the model, possible extensions to the model, as well as future developments of the recognition system.

# Chapter 2

# Time-Varying Spectral Extraction

## 2.1 Motivation

There is significant physiological and psychoacoustical evidence for a time-varying active adaptation in the human auditory system. Physiologically, there is evidence that outer hair cells provide varying amounts of amplification to the wave traveling down the basilar membrane in the cochlea. Psychophysically, we can measure how quickly the auditory system adapts to changing input levels. These measurements combined with static frequency-selectivity data quantify a first order model of adaptive auditory perception.

### 2.1.1 Physiological Review

The ear is divided into an outer, middle and inner ear. The outer and middle ear are often modeled as a single passive system. The inner ear contains the cochlea and is the interface between the pressure wave and the neural signals sent to the brain. Mechanically, the cochlea separates signals by their frequency components, while outer hair cells within the cochlea provide active, time-varying amplification. This thesis proposes a model of these last two functions leading to a dynamic spectral representation of sound.

#### 2.1.1.1 Outer and Middle Ear

A pressure wave traveling through the human auditory system first passes through the outer ear which consists of the pinna and the ear canal. The pinna and head provide different frequency responses as a function of the relative position of the sound source, while the ear canal provides a relatively broad and fixed resonance around 1.5 kHz. The middle ear is mostly an impedance matching device between air and the fluid within the cochlea. The tympanic membrane, or ear drum, is attached to three bones forming a lever which, in turn, is attached to the oval window opening to the cochlea. The middle ear changes the small pressure over the relatively large area of the ear drum into a large pressure over the relative small area of the oval window. There are muscles attached to the bones of the middle ear which can change their relative position reducing the amount of energy passed to the cochlea (when the listener speaks for example), but typically this effect is not significant. Therefore the entire outer and middle ear are mostly passive devices, often modeled as a single linear system [Pickles, 1988].

#### 2.1.1.2 The Cochlea

The cochlea converts the acoustic pressure wave from the middle ear into sequences of neural spikes sent to the brain. The pressure wave travels from the oval window at the base of the fluid-filled cochlea, along the basilar membrane toward the apex of the cochlea. At the apex of the cochlea, the chambers above and below the basilar membrane are connected. The lower chamber continues back from the apex to the round window near the base of the cochlea. Inner hair cells along the basilar membrane respond to the traveling pressure wave by firing neural spikes sent along the seventh cranial nerve to the brain. For every inner hair cell, there are approximately three larger outer hair cells connected to neurons travelling *from* the brain *to* the outer hair cells [Pickles, 1988].

The basilar membrane is not uniform. At its base it is broad and stiff; toward the apex it becomes narrower and less rigid [von Bekesy, 1960]. Because of this change in shape and stiffness, pressure waves travelling on the basilar membrane slow exponentially by a factor of about 100 as they travel from base to apex [Lyon and Mead, 1988]. As the wave speed slows, the energy per period from the wave concentrates over a shorter distance and the displacement per unit length of the basilar membrane increases. Inner hair cells fire in response to the change in displacement with length [Lyon and Mead, 1988]. Eventually, depending on the period (or frequency) of oscillation, the wave concentrates the deformation along a short enough distance of the basilar membrane, that the losses due to the deformation of the membrane dominate the propagating wave. After this point, the wave propagation becomes two and three dimensional, and the wave energy per unit length drops abruptly as it moves toward the apex. This mechanism effectively separates the frequency components of the input signal. Higher frequency components concentrate per unit length and dissipate earlier, closer to the base, while lower frequencies concentrate and dissipate later, closer the apex. The collection of neural spikes from inner hair cells along the length of the basilar membrane form a spectral estimation, as well as a detailed frequency-divided time domain representation, sent to the brain. Figure 2.1 [von Bekesy, 1960] shows the concentration and dissipation of a sinusoidal wave along the basilar membrane in the cochlea.

Figure 2.1 Displacement of the basilar membrane in response to a sinusoid at 4 stages in its propagation (after von Bekesy, [1960]).

G. von Bekesy[1953, 1960] pioneered cochlear mechanics, by directly observing the vibrations of the basilar membrane in human and animal cadavers using mechanical stimulation and an optically-based microscopic stroboscope. We now understand that the mechanical vibrations in the cochlea change dramatically with the condition of the cochlea and the animal. However, understanding the passive mechanical process of gradually concentrating and then quickly dissipating as a function of stimulus frequency is the first step toward physiologically defining auditory frequency selectivity. Further, this process suggests individual band-pass filters which are sloped more gradually toward lower frequencies and more abruptly toward higher frequencies. Lower frequency components concentrate slowly as they propagate past, and partially stimulate, “high-frequency” inner hair cells, while higher frequency components dissipate quickly and do not propagate past many “low-frequency” inner hair cells. Figure 2.2, after Ghitza [1991], shows the frequency response of specific inner-hair cells along a healthy cat’s basilar membrane.

Figure 2.2  Frequency response of neural fibers along a cat's basilar membrane (after Ghitza 1991).

Using the Moessbauer technique, Sellick *et al.* [1982] and Johnstone *et al.* [1986] have more precisely measured the displacement of the basilar membrane in healthy guinea-pigs as a function of input frequency and level. Figure 2.3.a from Johnstone *et al.* [1986] shows the change of basilar motion, at a point corresponding to an 18 kHz center frequency, with input frequency and level. Along the lowest input, 20 dB SPL, notice that the basilar motion is extremely sharply tuned: a small deviation from the center frequency results in a large reduction of basilar motion. However, as the level of the input increases, the curves become increasingly broader: small deviations in frequency are no longer as significant. Figure 2.3.b

[Sellick *et al.*, 1982] records the level of an input, of varying frequencies, necessary for a constant amplitude basilar motion. The dotted line indicates the threshold of a neuron attached to the basilar membrane with an 18 kHz center frequency. Again, notice the tuning is sharpest with lower inputs. A frequency response that changes with input level necessitates at least a non-linear system, and may suggest level-dependent active amplification.

Figure 2.3 Measured at 18 kHz center frequencies: A) Basilar motion in response to constant level inputs of varying frequency (after Johnstone *et al.*, 1986). B) Constant basilar motion in response to inputs of varying frequency and level: the dotted line marks the threshold of an associated inner hair cell (after Sellick *et al.*,1982).

### 2.1.2 Evidence Supporting a Mechanically Active Cochlea

There is significant direct evidence that the cochlea is not a mechanically passive device. Cochlear emissions may provide the strongest evidence for active amplification in the cochlea. Responding to a quiet click, the cochlea can generate more acoustic energy in the ear canal, than the click itself [Kemp, 1978]. More obviously, during tinnitus, we can acoustically measure the ringing produced by the cochlea [Wilson, 1980], most likely the result of an instability introduced by active amplification. In addition, healthy outer hair cells change their shape (with time constants near 200 *u*sec) in response to electrical stimulation, suggesting a suitable mechanism for active amplification, perhaps even on a cycle by cycle basis [Ashmore, 1987]. Further, experiments that selectively disable outer hair cells in the cochlea show increased thresholds and broader tuning curves [Evans and Harrison, 1976].

Ignoring influences from outer hair cells, a passive mechanical model does not explain extremely sharp tuning curves of healthy hearing, nor does it explain the enormous usable dynamic range. Careful measurements and simulations of the passive mechanical system predict relatively broad (fixed) filter shapes [Viergever and Diependaal, 1986], similar to those measured with dead cochlea. By adjusting model parameters for less damping, the passive mechanical model can approximate the sharp tuning, if not the level-dependence, however in order to produce the magnitude of the tuning in Figure 2.3.a, the tip of the resonance becomes too narrow [Viergever and Diependaal, 1986]. Instead, evidence supports an adaptive influx of energy from the cochlea to the traveling wave just before the wave reaches its peak displacement, as shown in Figure 2.4 from Pickles [1988].

Figure 2.4 An active mechanism explaining basilar motion: dotted lines shows passive mechanics, after Pickles [1988].

Models incorporating such an active mechanism reproduce the level-dependent basilar membrane responses and neural tuning curves of Figures 2.3.a,b [Neely and Kim, 1986]. Analysis of the firing rates of inner hair cells indicates that even though some cells fire in response to low level inputs, and others fire in response to higher-level inputs, 80% of these cells have thresholds in the lowest 20 dB [Liberman, 1978], suggesting at least difficulty encoding our 100+ dB usable range. Interestingly, a reduced dynamic range, and wider filter shapes are more characteristic of sensorineural hearing loss.

This evidence, combined with number of outer hair cells and the neural connection from the brain to the outer hair cells, suggests an active feedback loop from the brain to the outer hair cells [Pickles, 1988]. The outer hair cells add energy to the propagating wave, increasing stimulation of specific regions of inner hair cells. When a sound at a specific frequency is barely audible, outer hair cells at the corresponding region provide their maximum amplification. As the sound increases in level, the outer hair cells apply increasingly less amplification. Above some input level, and perhaps below some lower input level, outer hair cells provide little or no amplification.

Adaptive amplification of the input increases the usable dynamic range; otherwise inaudible sounds are amplified to audibility, while sounds that would not need amplification are not amplified. This adaptation compresses a wide dynamic range into a smaller one. More subtly, adaptive amplification, a non-linearity, makes the effective frequency response of each filter level-dependent. The adaptive change of amplification, presumably controlled or at least influenced through neurons from the brain, is a primary physiological mechanism for the dynamic auditory model presented in this thesis.

### 2.1.3 Psychoacoustical Evidence Supporting Adaptation

There are three categories of evidence supporting auditory adaptation: more directly measurable psychoacoustic phenomena, secondary evidence of perceptual effects consistent with a model of auditory adaptation, and comparisons of hearing impaired perception with that of healthy hearing. The first provides an opportunity to quantify model parameters, while the second and third provide confirmation that ramifications of the model are consistent with other measurable effects; the model is not specific to a single perceptual experiment nor phenomenon. In fact, this relatively simple model, provides a framework that explains a surprising amount of non-linear auditory perception.

#### 2.1.3.1 Direct Psychoacoustic Evidence

The most direct psychoacoustic effects supporting an underlying model of adaptation are forward- or post-masking, and level-dependent perceptual tuning curves. Forward masking is relatively quantifiable, however, due to non-linearities in the cochlea, perceptual tuning curves are slightly more elusive.

Forward masking reveals that even though our auditory system may have a 100+ dB dynamic range, over short durations the actual dynamic range is much smaller and largely dependent on previous stimuli. When a probe signal follows a masking signal with similar spectral characteristics, the probe signal is less audible than a similar probe signal following silence. As the duration between probe and masker decrease, the threshold of the probe is increasingly a function of the level of the proceeding masker, and decreasingly a function of the probe signal threshold in silence. We interpret this phenomena as the auditory system adapting to the masker. Once the system adapts to the level of the masker, it takes time for the system to adjust so that the lower-level probe is audible. Over short durations, the audible dynamic range below a masker is far less than that predicted by static threshold measurements. In fact, we show over the range of durations and levels significant for speech perception, this is an enormous effect which shifts thresholds by as much as 40-50 dB from their static levels. Further, the shift is a function of the duration of the masker [Zwislocki *et al.*, 1959], reflecting the time necessary for the auditory system to adapt completely to the masker. Forward masking experiments provide the opportunity to measure the rate of adaptation as a function of input level and frequency, in addition to the magnitude of this phenomenon.

To a first order, psychophysical tuning curves approximate physiological tuning curves. Figure 2.5 shows estimates of psychophysical tuning curves at three center frequencies [Zwicker, 1974]. The subject adjusts the level of a narrow-band masker until a low level sinusoid held fixed at the center frequency is just audible. Notice the general similarity to the physiological curves of Figure 2.3.

Figure 2.5 Psychophysical Tuning Curves, from Zwicker [1974].

However, Houtgast [1977] explains that because psychophysical tuning curves are usually measured with simultaneous masking of tones by noise, the presence of the masker affects the perception of the tone; both signals are processed through a non-linear cochlea. Instead, Houtgast proposes measuring psychoacoustic filter shapes with non-simultaneous (forward) masking, by probing the after-effects of a single tone. A short (10-20ms) probe tone of varying frequency immediately follows a longer masking tone. The frequency-dependent threshold of the short probe indicates how much neighboring filters, or neurons, are stimulated by the tonal masker. Collections of these data indicate filter responses significantly narrower, less symmetric, and more similar to the physiologically estimated filters in Figure 2.3, than those from simultaneous masking data [Moore *et al.*, 1984]. Figure 2.6 compares the tuning curves derived from simultaneous and non-simultaneous masking, highlighting the differences in bandwidth and symmetry.

Figure 2.6 Tuning curves from simultaneous and non-simultaneous (forward) masking experiments, after Moore [1978].

Clearly, we can not define a single, ideal, cochlear filter frequency response. Such a definition reflects the imposition of an idealized linear model on a fundamentally non-linear system. Nonetheless, the perceptual and physiological evidence do support filters of: increasing bandwidth with increasing center frequency, more abrupt high frequency cut-off than low, and increasing sharpness with decreasing level.

#### 2.1.3.2 Secondary Perceptual Evidence Supporting Adaptation

There is a second group of perceptual evidence supporting auditory adaptation which does not provide readily measurable model parameters, but does confirm that ramifications of auditory adaptation are consistent with other measurable perceptual phenomena.

The human auditory system is relatively insensitive to fixed, or even slowly changing, channel frequency-characteristics. The location of spectral peaks, and not so much their relative strength nor bandwidth, provide the most salient cues for static vowel perception [Klatt series: 1979, 1980, 1981, 1982, 1986]. An adaptive auditory system adjusts to the channel frequency-characteristics so that fixed average energy differences across frequency are less significant.

Perhaps equally important for the perception of speech, onsets and spectral transitions are more salient than static sounds [Furui, 1986]. As the input to an adaptive model transitions from low-level (or near silence) to higher-level, the adaptive model takes time to adjust to the high-level input. At the onset the model applies large amplification to a higher-level input, greatly emphasizing the onset. A spectral transition provides a sequence of “onsets” in different frequency bands, just as running your finger across the keys of a piano provide a sequence of onsets of each note. Channel-insensitivity and emphasis of onsets and transitions are listed as the primary motivation for the relative spectral representation of RASTA [Hermansky *et al.*, 1992], but are also natural ramifications of an adaptive auditory model.

#### 2.1.3.3 Evidence From Sensorineural Hearing Impairment

There is more secondary evidence of auditory adaptation from the perception of hearing impaired by sensorineural loss. These losses, usually characterized by outer hair cell damage, result in increased tone thresholds in quiet, reduced usable dynamic range, and broader tuning curves [Zwicker 78]. If the outer hair cells play an essential role in auditory adaptation, their loss implies reduced adaptation, consistent with increased tone thresholds in quiet: the auditory system can no longer amplify lower level signals; and reduced dynamic range: the auditory system can not reduce the amplification with increasing input level to compress a wide dynamic range. In fact, to compensate for these hearing losses, hearing aids with different types of automatic gain control, off-loading part of the auditory adaptation are prescribed [Dillon and Walker, 1982]. Similarly, if adaptive amplification plays a key role in sharpening the tuning curves, otherwise available from a passive cochlea, the loss of adaptation implies broader- than normal- tuning curves.

Finally it seems reasonable that adaptation, common to the rest of our senses (if not our thoughts), is perhaps a primary lesson of evolution, which naturally extends to our auditory system.

## 2.2 A Dynamic Auditory Model

We model a dynamic auditory system with a simple filter-bank followed by carefully-parameterized individual automatic gain control on the output of each filter. Figure 2.7 shows a schematic overview of the proposed model.

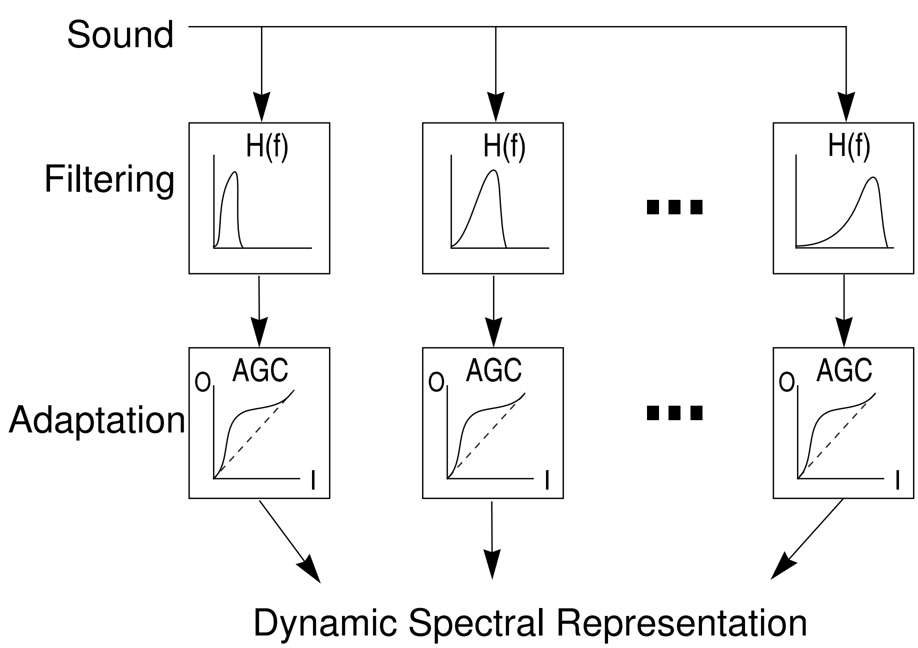


Figure 2.7 Overview of the dynamic model.

The filter bank models the frequency selectivity from cochlear mechanics, the automatic gain control models adaptation. The parallel filter bank is unambiguous once frequency responses are specified, however the non-linear automatic gain control (AGC) requires more careful explanation.

Typical AGC slowly adapts to maintain the output near a target level when the input changes level. A time constant determines how quickly the model adapts, and the output target level sets the final level. More sophisticated AGC, typical of those used in compression hearing aids, adjust its amplification only above an input threshold, called the compression threshold. Once the input is above threshold, the target output may slowly increase with increasing input. The reciprocal of the slope of this increase defines the compression ratio. Plotting target output level as a function of input level, on a log/log scale, defines the static I/O curve for the AGC. Figure 2.8 is an example of a typical I/O curve for a hearing aid.

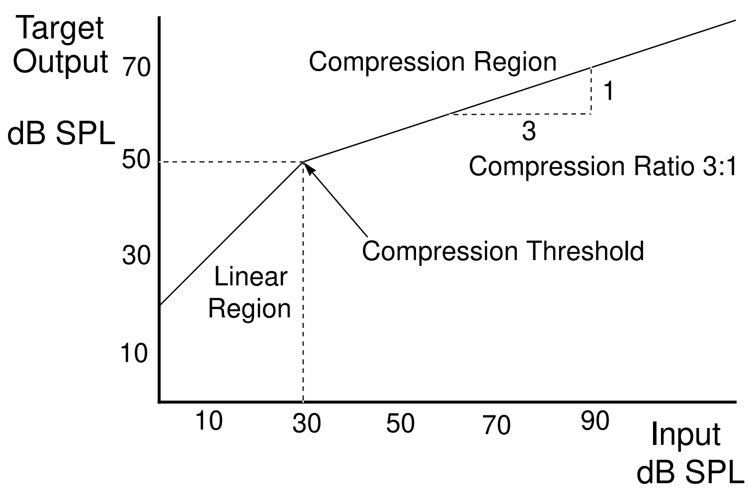


Figure 2.8 A typical AGC static I/O curve.

Notice inputs below 30 dB are linearly amplified by 20 dB, however inputs above 30 dB are amplified increasingly less. Notice also, above the compression threshold, the 60 dB input dynamic range (30 to 90 dB) is compressed into a 20 dB dynamic range (50 to 70 dB). On log/log IO curves, a slope of one implies linearity, slope less than one implies compression, and a slope greater than one implies expansion. Linear amplification translates to a vertical shift. The model developed in this thesis uses independent AGC on the output of each filter, each with I/O curves and time constants derived from forward masking experiments.

Lyon described a filter bank followed by automatic gain control cochlear model first in 1982; Lyon and Mead [1988] and Kates [1991] continue to evolve this model along with models of higher-level processing. Other cochlear models [Duifhuis, 1973, Seneff, 1990] incorporate explicit, often probabilistic, adaptive neural models after the filtering. Instead of modeling the physiology in detail, we focus on the two primary processes suggested by perceptual data: frequency selectivity and adaptation. The model described here differs from Kates' and Lyon's in the structure and implementation of the filter bank, but perhaps more significantly, we derive perceptually-based I/O curves and time constants for the automatic gain control, quantitatively tying the adaptation of the model to measured perceptual phenomenon. Finally, we evaluate the model by incorporating it as the front-end for a simple speech recognition system.

## 2.3 Ramifications of the Dynamic Model

The model incorporates a simple adaptation mechanism with frequency selectivity. As such, it reproduces several of the physiological and perceptual phenomena listed above as evidence for frequency-dependent adaptation.

### 2.3.1 Wide Dynamic Range

Figure 2.9 shows a possible I/O curve for a single automatic gain control block in the model.

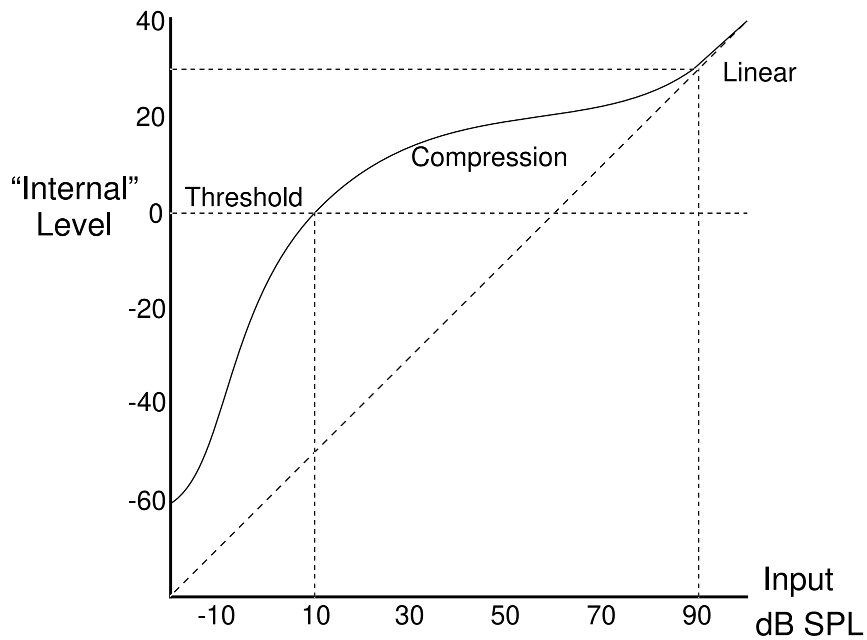


Figure 2.9 Prototypical I/O curve for a single AGC block.

The model provides maximum amplification for inputs near threshold, decreasing amplification with increasing input level, and little or no amplification above some upper limit. AGC with the I/O curve of Figure 2.9 will compress the wide input dynamic range from 10-90 dB SPL into a smaller 0-30 dB internal dynamic range.

### 2.3.2 Dynamic Relative Loudness and Dynamic Thresholds

I/O curves for AGC provide the static output for a given *static* input. Time is not included on an I/O curve. For dynamic inputs, the I/O curve provides the corresponding target output level, implying a target amplification. Time constants control how quickly the AGC adjusts to the target values. Over short time scales relative to the time constants of the model, the amplification does not adjust, and the system is nearly linear. On I/O curves, short time scale changes are viewed as instantaneous motion along a diagonal line, followed by relatively slow adjustment to target values. Figure 2.10 shows the prototype I/O curve with output trajectories corresponding to an input transitioning from 80 to 30 dB SPL at three different rates relative to the model time constant.

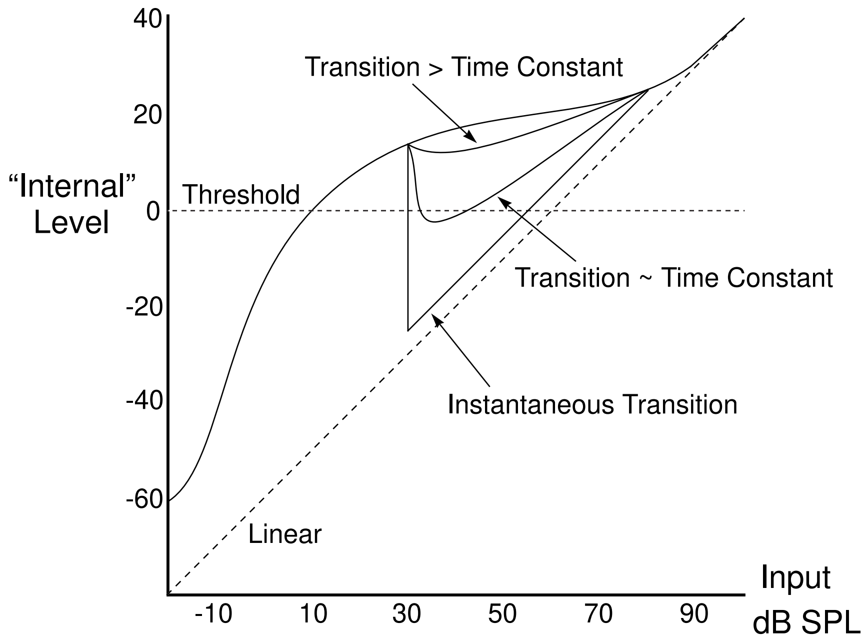


Figure 2.10 Output trajectories with inputs transitioning from 80 to 30 dB SPL at three different rates.

When the input transitions instantaneously from 80 to 30 dB, the trajectory falls linearly, and the internal level correspondingly falls by 50 dB. Eventually the adaptive amplification of the AGC increases, and the trajectory rises to meet the static target output for the new lower input. When the input transition is much slower than the adaptation time constant, the trajectory follows the I/O curve more closely. In between, with transition times on the order of the adaptation time constant, there is a continuum of possible trajectories.

Notice, it is quite possible for rapid transitions between levels which are both above static thresholds, to lead to momentary dips below threshold. Whenever the trajectory dips below threshold, the model predicts the sound is inaudible. Therefore, with carefully chosen I/O curves and time constants, this structure predicts dynamic thresholds, or forward masking. The amount of forward masking will be a function of the level of the masker, the delay between masker and probe, and the length of the masker (until the masker length is long relative to the time constant). Further, the continuum of super-threshold trajectories predicts dynamic relative loudness. That is, the model predicts the perceived, internal loudness to be a function of the preceding stimuli.

### 2.3.3 Level-Dependent Filter Shapes

AGC is non-linear: different level inputs result in varying amplification. Therefore, the system of a linear filter followed by AGC has a level-dependent effective frequency response. Figure 2.3 shows that the frequency response of the basilar membrane at a fixed position is a function of the level of the input. More specifically the sharpest frequency response occurs with inputs near threshold. A simple example will show how the proposed model at least qualitatively reproduces this phenomenon.

Assume one filter in our model has a 20 dB drop 200 Hz from its center frequency, and that we use the prototypical I/O curve for the AGC. A static 15 dB sinusoid at the center frequency on the input, passes the linear filter with unity gain, and is then multiplied by the AGC such that the internal level is just above threshold. A sinusoid of the same level, shifted by 200 Hz, passes the linear filter with a 20 dB drop. The I/O curve places the internal level roughly 30 dB below the level of the sinusoid at the center frequency. The non-linearity of the AGC expands the 20 dB linear difference into a 40 dB internal difference. If we consider the same two sinusoids at an increased amplitude of 70 dB, we see the 20 dB difference from the linear filter is compressed into a 5 dB difference internally. Therefore, the shape of the I/O curve defines the level-sensitivity of the effective frequency response. Figure 2.11 shows these two examples. With the prototypical I/O curve, the sharpest filter response occurs around threshold, and filter shapes are increasingly broad with increasing level.

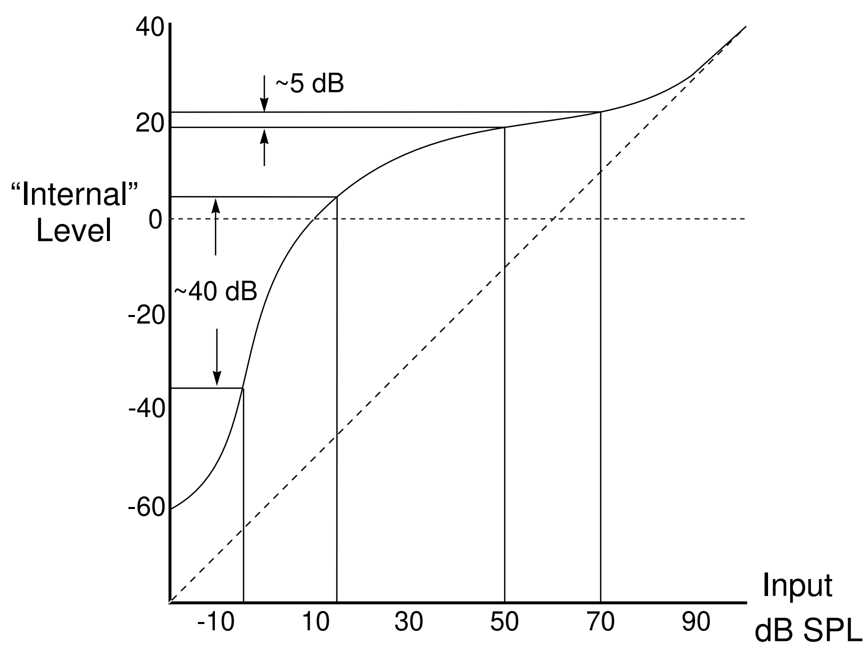


Figure 2.11 Examples of how AGC leads to level-dependent filter shapes. After

linear filtering in our model, a 20 dB change in frequency response translates to a 5 or 40 dB internal level change, depending on the level of the input.

### 2.3.4 Emphasis of Onsets

Previously we considered how the model responds when the input transitions from high to low levels. Even if the low level is above a static threshold, depending on the rate of transition, the model may predict momentary inaudibility. A similar analysis of transitions from 30 to 80 dB SPL, with trajectories for different transition times appears in Figure 2.12. The model is nearly linear when the transition is much less than the model time constant. Corresponding trajectories move diagonally instantaneously, and then settle vertically to the corresponding point on the I/O curve. As the transition time increases, the trajectory stays closer to the I/O curve. As before, there is a continuum of possible trajectories depending on the rate of transition.

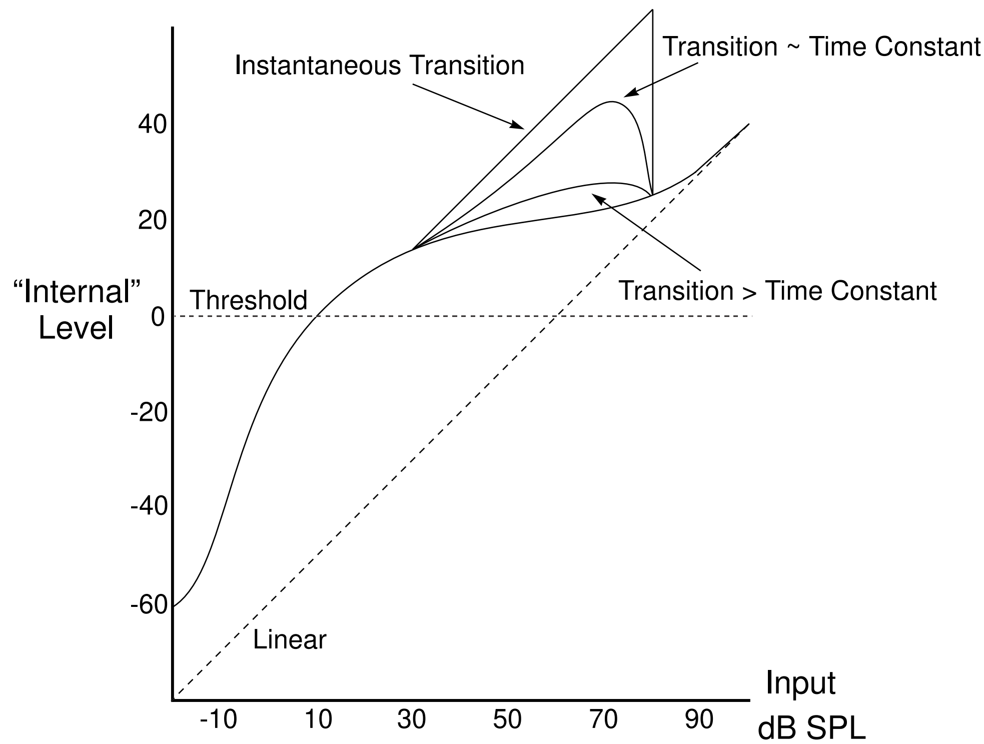


Figure 2.12 Input transition from 30 to 80 dB SPL with three different transition times.

The start of a sound, or the onset, may include low to high-level transitions on several bands simultaneously. Clearly, the adaptive model strongly emphasizes such onsets.

### 2.3.5 Emphasis of Spectral Transitions

A sine-wave sweep is a simple example of a spectral transition. More generally, a spectral transition is a motion in frequency over time of some component of a sound. Spectral analysis of conversational speech shows that spectral transitions may be more the rule than the exception. The adaptive model uses independent AGC on the outputs of the parallel filter bank. As a signal transitions in frequency from low to high, the signal is an effective onset for each filter as it moves into the filter's pass band. Therefore, a spectral transition creates a sequence of "onsets" at different frequencies, and the model emphasizes these onsets as described above. Notice a single tone will generate only one emphasized onset, while a transition will generate a sequence of emphasized onsets. Figure 2.13 schematically describes the model emphasizing a spectral transitions.

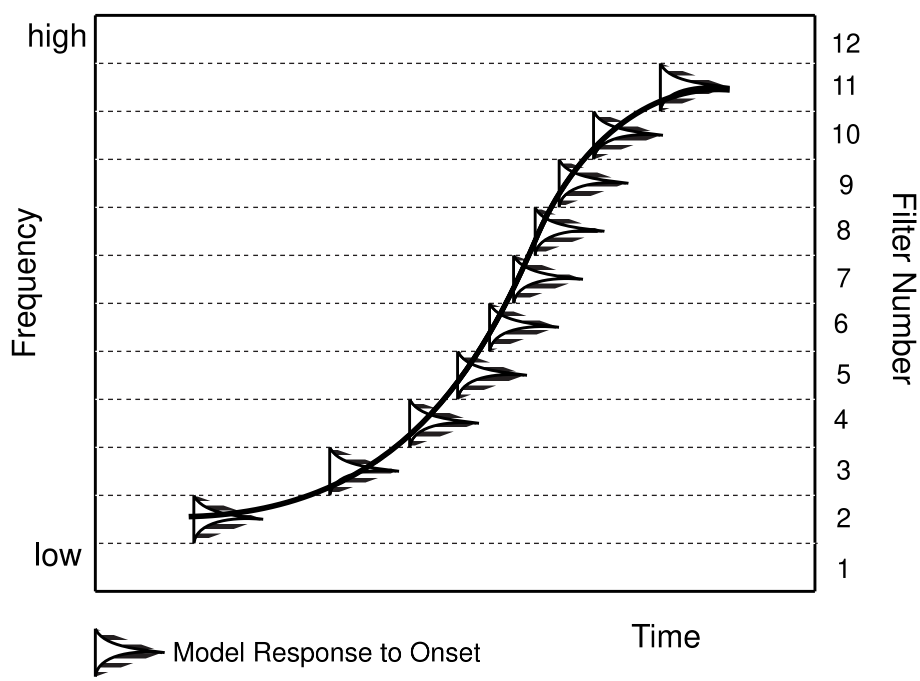


Figure 2.13 Spectral transitions create a sequence of onsets from each filter/AGC pair.

### 2.3.6 Reduced Sensitivity to Channel Shapes

The model compresses a large input dynamic range into a smaller internal range, it predicts dynamic thresholds, dynamic relative loudness, emphasis of onsets, and emphasis of spectral transitions. Each of these are natural ramifications of a system adapting to its input and emphasizing changes. Further, with this model, we expect less sensitivity to the frequency characteristics of the channel, than with more classic static filter bank models. The proposed model adapts to the relatively static channel frequency shape, and emphasizes signal changes.

### 2.3.7 Static Perception

In addition to predicting several aspects of dynamic auditory perception, the proposed structure models several static auditory phenomena. Specifically, the filters group signals of similar frequency into a single band. If one of the signals with spectral components in this band is louder, the energy in that band reflects the louder signal. This is the primary mechanism of simultaneous masking, where loud signals mask quiet signals of similar frequency. Perhaps more subtly, the shape of the filters determine how simultaneous masking varies as a function of the difference between the masker and probe frequencies.

We have already seen physiological and psychoacoustical evidence for filter shapes with gradual transition toward lower frequencies, and more abrupt cut-off toward higher frequencies. These filter shapes predict above-band masking; lower frequency signals of sufficient energy mask higher frequency signals. A single filter shape, imposed over a low frequency narrow-band noise and a higher frequency tone appear in Figure 2.14.

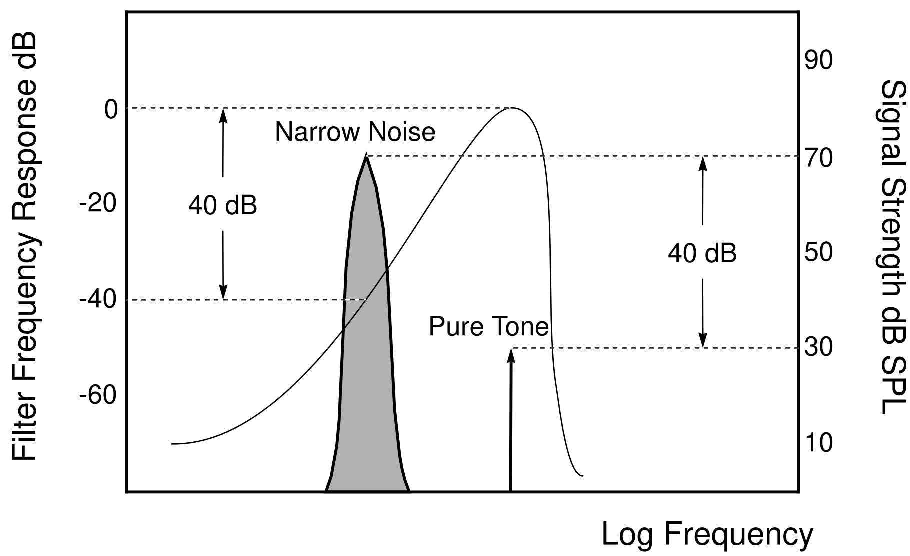


Figure 2.14   Filter shapes predict above-band masking.

Even though the noise and tone are of different frequencies, the energy at the output of the filter will be identical for the noise and the tone; the noise will just mask the tone. Notice the model does not predict significant masking when the noise is at a higher frequency than the tone. Therefore, with reasonable filter shapes which slope more gradually toward lower frequencies than higher, the model predicts the first-order detail of simultaneous and above-band masking.

Finally, the model predicts static constant loudness contours. These contours show the level of a signal necessary to maintain constant perceived loudness as function of signal frequency. The break points and shapes of the I/O curves, and the offset of the internal thresholds, for each filter/AGC pair, define the constant loudness contours of the model. More specifically, I/O curves for middle frequency bands have lower thresholds, and flatter compressive regions than those at lower frequencies. The position and shape of the static I/O curves determine the model's sensitivity to static as well as dynamic inputs.

## 2.4 Need for Model Parameters

The model uses filter shape approximations from the physiological and psychoacoustical evidence described previously. To fully specify the model, we derive I/O curves and time constants of adaptation for each frequency band from psychoacoustic forward masking experiments. The time constants and I/O curves completely specify the model's dynamic thresholds. Matching the dynamic thresholds to the forward masking data provides a fully parameterized first-order adaptive auditory model.

# Chapter 3

# Perceptual Experiments

## 3.1 Why Forward Masking?

The human auditory system is active and adapting. This thesis describes the development and evaluation of a dynamic auditory model which includes active adaptation and linear frequency selectivity. Our present task is to choose the parameters for the adaptation.

In Chapter 2, we described several physiological and psychophysical phenomena as evidence of auditory adaptation. Carefully quantifying any of these effects implies parameters for our model. However, many of the phenomena are ramifications consistent with active adaptation, while few provide direct estimates of model parameters. Perceptual emphasis of onsets and transitions, immunity to slowly-varying channel frequency characteristics, and compression of a wide dynamic range, are all natural ramifications of an adaptive model but do not illuminate specific model parameters. On the other hand, level-dependent frequency responses, and dynamic relative thresholds provide a more direct description of the underlying auditory adaptation.

Previously, we described the difficulties of measuring individual filter responses psychoacoustically. In fact, forward masking experiments provide filter shapes more consistent with those from physiological experiments, than do filter shapes obtained from simultaneous masking tests [Houtgast, 1977]. Further, static measurements of level-dependent frequency responses only provide a measurement of the magnitude of adaptation, and do not provide a measure of the adaptation rate.

However, dynamic relative thresholds revealed in forward masking experiments provide a glimpse of both the rate and magnitude of auditory adaptation. Also, the rates and magnitudes of the adaptation revealed in forward masking are significant for speech perception. Therefore, for the proposed first-order model, we select the parameters of active adaptation after each auditory filter, from the dynamic thresholds measured in forward masking experiments. Measured as a function of masker level and frequency, forward masking data provide both the I/O curves and time constants for our model.

## 3.2 Review of Existing Forward Masking Data

Moore [1989] summarizes five basic points from forward masking data: 1) the amount of forward masking increases as the duration between the probe and masker decreases; 2) forward masking does not increase by an amount equal to the increase in masker level (as simultaneous masking does); 3) forward masking increases with the duration of the masker, at least through masker durations of 50 ms; 4) the amount of forward masking decreases when the masker and probe are of differing frequencies; and 5) the rate of recovery from forward masking is faster for higher-level maskers, than for lower-level maskers-- that is, the rate of the decay of the amount of forward masking is greater for higher-level maskers which create more forward masking.

Moore's summary points 1-4 are fairly obvious and readily modeled, however the last point that the rate of recovery varies with masker-level may make any "simple-minded" modeling effort seem hopeless. Nonetheless, Chapter 4 includes a description of how our relatively simple model of auditory adaptation predicts Moore's five summary points of forward masking.

### 3.2.1 Previous Forward Masking Data

Duifhuis [1973] describes forward masking as the result of two processes: one that occurs over very short times (<20 ms) and another that occurs over longer times (~75 ms). Further, he explains short time masking in terms of time domain interactions on the basilar membrane between the probe signal, and the 'ringing' of narrow bandwidth auditory filters. He suggests that the longer term masking is due to adaptation through neural saturations and latencies at several levels of the auditory system. Viewing forward masking data as a function of the log of the delay between masker and probe, Duifhuis fits a time constant of 75 ms to the rate of decay of the amount of masking. (Note that a single time constant is contrary to Moore's observation that the rate of decay varies with the amount of masking, strongly supported by Plomb's data [1964].)

Despite the vast empirical data on forward masking, we had difficulty finding a reasonably complete data set describing forward masking of short probe tones by long narrow band maskers. These data are necessary to quantify each adaptation block in our model individually. Perhaps the most complete data set of forward masking of tones following tones is from Jesteadt *et al.* [1982]. Although this data includes a range of frequencies and masker levels, the longest delay measured between masker and probe is only 40 ms. At this delay there is still significant forward masking, making it difficult to characterize complete auditory adaptation. The majority of published forward masking data involves the masking of impulses or clicks. Obviously, impulses maximize the time-domain granularity of the measurements, at the expense of fundamentally no frequency-domain resolution. For our model, we need a map of adaptation at each frequency, to maskers at that frequency, varying across the majority of the auditory dynamic range, and as a function the range of delays significant for auditory (and speech) perception.

Therefore, to derive the adaptation parameters for our model, we devise a simple, but complete, set of forward masking experiments to quantify auditory adaptation as a function of time, input level, and frequency.

## 3.3 Forward Masking Experiments

Our forward masking experiments use long tone maskers followed by short tone-like probes of similar frequencies. Tonal maskers and probes of the same frequency provide measurements of the adaptation in the auditory system at one center frequency at a time. We derive the parameters for the adaptation blocks after each filter in our model from the data of these experiments. Because we use tonal maskers and probes at the same frequency, translating the data into model parameters is not complicated by the frequency selectivity of audition, nor by possible differences in the neural processing of tones and noise. In some sense, we *attempt* to measure auditory adaptation in response to the stimulation of a single point along the basilar membrane.

The masker tone is long enough to ensure that the auditory system has completely adapted before the masker is shut off, and the probe is short enough to measure the response of the auditory system at a relatively specific point in time after the masker. The experimental paradigm is 2AFC. A decaying 60 ms probe tone follows one of two 300 ms maskers, separated by 500 ms. The subject chooses which masker the probe followed. We require a range of levels, frequencies and delays to completely quantify the magnitude and rate of auditory adaptation across reasonable hearing ranges. The magnitude and rate of auditory adaptation imply the I/O curves and time constants of our model.

### 3.3.1 Stimuli

We select masker frequencies from 250 through 4000 Hz at octave intervals, three masker levels, and four exponentially spaced delay times between masker and probe from 15 to 120 msec. The static threshold of the probe, corresponding to an infinite delay time from the masker is also measured. Both the sinusoidal masker, and the sinusoidal probe are gated on and off with one half period of a raised-cosine function. On and off times, defined as the length of the half-period of the raised-cosine, are 5 msec for both signals.

We require measurements of auditory adaptation as a function of time and frequency. Our goal is to measure the auditory response to a specific frequency tone at a specific delay after the masker. Unfortunately, time and frequency granularity can only be traded for one another. Shorter probe durations improve time granularity, but as closer approximations of impulses (or clicks) they tend toward equal energy at all frequencies. With longer probe durations we are no longer certain of when the subject started hearing the probe. To improve this trade-off for these specific measurements, after the raised cosine onset, the probe decays exponentially with a 20 ms time constant (20 ms after onset, the amplitude of the probe has decayed to $1/e$ of its original value). Therefore, much of the energy in the probe occurs just after its onset: the energy is not spread evenly throughout the duration of the probe as it would be for a pure tone gated on and off. To ensure that subjects respond to the onset of the probe, and not its tail, the masker without a probe in the 2AFC experimental paradigm (described below) also decays exponentially with the same time constant. The subject must detect the onset of the probe, and not only its decay. To further reduce the spectral splatter of the onsets and offsets, the entire stimuli is filtered through a linear phase, 201 tap, FIR filter with bandwidth equal to one critical band, centered at the (masker and) probe frequency.

Duihfuis [1973] found that phase shifts between masker and probe can be significant at short enough masker-to-probe delays; any phase difference implies different time-domain interactions between the probe and the ringing of the basilar membrane. Phase shifts can also be perceived as momentary pitch shifts. Therefore, the phase of the probe is identical to that of the masker. Figure 3.1 describes the stimuli in the experiment in more detail.

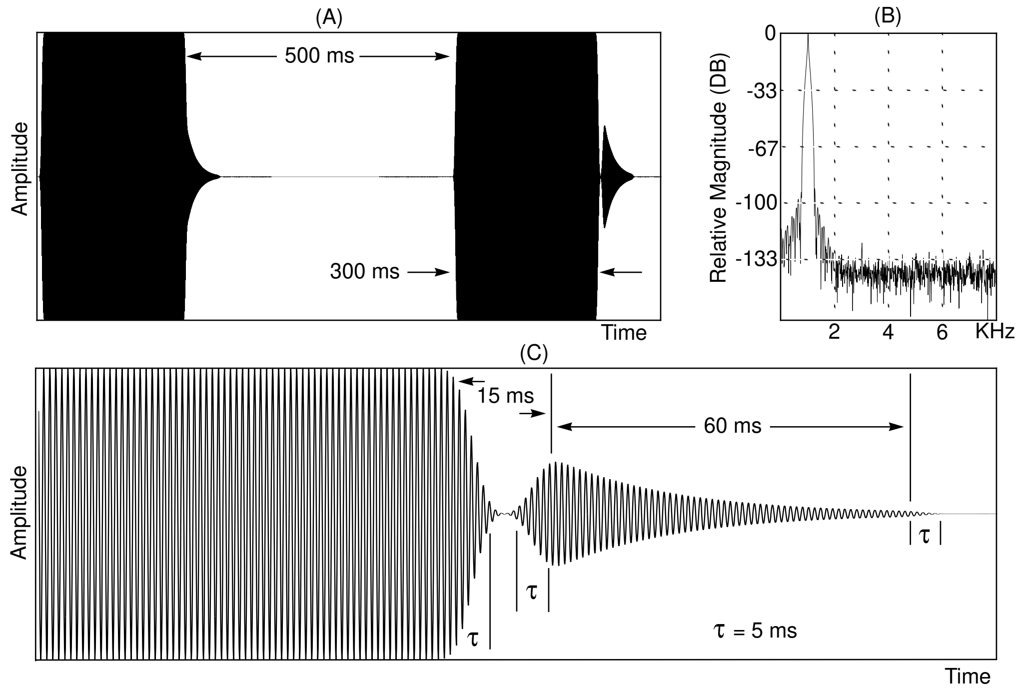


Figure 3.1 Forward masking stimuli: A) Large time scale view of a single 2AFC trial; B) Fourier transform of the probe signal (128 ms rectangular window); C) Smaller time scale view of the probe following the masker by 15 ms.

Notice the exponential decay of the masker without a following probe, and the narrow spectral shape of the probe. In this example, the delay from the masker to the probe is 15 ms (measured in figure 3.1.c), the probe is 8 dB quieter than the masker, and both probe and masker oscillate at 1kHz.

### 3.3.2 Subjects

Five subjects, BS (the author), JG, JH, PB, and SC, participated in the experiments. All are graduate students at UCLA, all are native speakers of American English. JG is female, and the other subjects are male. Their ages range from 23 to 28 years. Static hearing thresholds for each were at, or below, 20 dB HL at frequencies from 250 to 8000 Hz, with the exception of JH's left ear which has a 25 dB HL threshold at 8 kHz. All subjects, except the author, were paid.

### 3.3.3 Methods

As mentioned earlier, the experiment paradigm is 2AFC. The subject hears two masking tones, with a probe signal following one. The subject responds by choosing which masker the probe followed. For a given masker level and frequency, and masker-to-probe delay, the probe level is varied to determine a threshold.

We implement an adaptive "transformed up-down" procedure [Levitt, 1971, 1992] to adjust the level of the probe. An adaptive up-down procedure decreases the level of the probe when the subject responds correctly, and increases the level when the response is incorrect. Obviously, this concentrates stimuli near threshold, improving the convergence to threshold, and decreasing total test time. A transformed adaptive procedure requires more than one trial to decide whether the subject responded correctly to a given input. In these experiments we require correct responses to three consecutive trials before the group-response is considered correct. An incorrect response on any of the three trials defines an incorrect group-response (and terminates that group). Computer software terminates the test if adapting stimuli exceed 90 dB SPL.

A psychometric function is the subjects' percentage correct choice as a function of probe level. When the probe is inaudible, the subject is forced to guess, and will be correct 50% of the time. When the probe is clearly above threshold, the subject will be correct nearly 100% of the time. Increasing the level after every incorrect response, and decreasing the level after every correct response converges stimuli to the 50% point. This is of little use in a 2AFC paradigm. However, requiring three correct responses in the transformed procedure converges stimuli to the 79% point on the psychometric function. The subject has to hear a louder level before he or she will respond correctly three times in a row. We define the probe level at the 79% point on the psychometric function for our 2AFC test as the probe threshold.

A reversal defines the point when the adaptive test changes from increasing stimuli to decreasing stimuli (or vice versa). The step size defines the change of level as the stimuli is increased or decreased. In these experiments we use an initial step size of 4 dB. After the first reversal the step size reduces to 2 dB, and after the third reversal, the step size reduces to 1 dB. The experiment continues until nine reversals occur. The stimuli levels at the last six reversals are averaged to determine the final threshold value.

There are six sessions for each subject, including a first session for static hearing threshold tests, and training with example forward-masking experiments. Each session, including the first, required 45 minutes to an hour. Subjects respond to the trials through a computer terminal. Software controls the experiment by adapting levels, regenerating stimuli, and summarizing results.

### 3.3.4 Equipment and Calibration

Computer software generates the test tokens in digital form as the experiment runs. The sampling rate is 16 kHz, and the samples are represented as 16-bit numbers (linearly quantized). An Ariel ProPort 656 converts the digital samples into an analog waveform. The pre-amp of a Sony 59ES DAT recorder, drives TDH-49P headphones. The subject hears tokens through the headphones in a double-walled sound isolation chamber. The stimuli is presented binaurally with identical waveforms to each ear. The Sony pre-amp is necessary to place the low end of the 16-bit digital dynamic range just below static threshold. Without it, the quietest tones are so coarsely quantized that harmonic distortion is perceptible.

We calibrate the system by playing digitally synthesized sine-waves through the headphones attached to a Larson Davis 800B Sound Level Meter with a 6cc coupler. Calibration occurs in two stages. The first ties an internal digital dB scale to measured dB SPL, the second is linear equalization to correct for the frequency response of the total system.

First, the computer generates a 1KHz reference tone at 80 dB on an internal digital scale. The gain of the pre-amp is adjusted until the Sound Level Meter measures 80 dB SPL. Second, sine-waves at third octave intervals from 125 to 7500 Hz are generated at 80 dB on the internal digital scale. A final 401-tap linear-phase FIR equalization filter corrects for any differences from 80 dB SPL across these frequencies as measured by the Sound Level Meter. Software redesigns the final

FIR filter by windowing the inverse DFT of the desired equalization. Convolution of the data with the FIR filter is performed using an FFT-based overlap and add technique for near real-time performance. Calibration at both stages is iterative. The difference from the 80 dB SPL measured during the first calibration step is corrected within $\pm$ 0.2 dB SPL, the equalization is corrected within $\pm$ 0.5 dB SPL, so that reported dB SPL are within $\pm$ 0.7 dB SPL.

All software described above: test token generation, experiment adaptation and control, and calibration, is written in C and compiled on HP and Sun workstations. The FFT routine used is from “Numerical Recipes in C” [1992].

### 3.3.5 Experiment Results

There are several ways to view forward masking results. In our data, the threshold of the probe is a function of three variables: the level of the masker, the delay between masker and probe, and the frequency of the masker and probe. Further, instead of viewing the absolute threshold, often the shift in threshold from that of a probe with no masker, defined as the amount of forward masking, is viewed. Figure 3.2 shows the average amount of masking as a function of the level of the masker at 1 KHz. The four contours correspond to the four delay times between masker and probe. Vertical lines at data points indicate standard deviation.

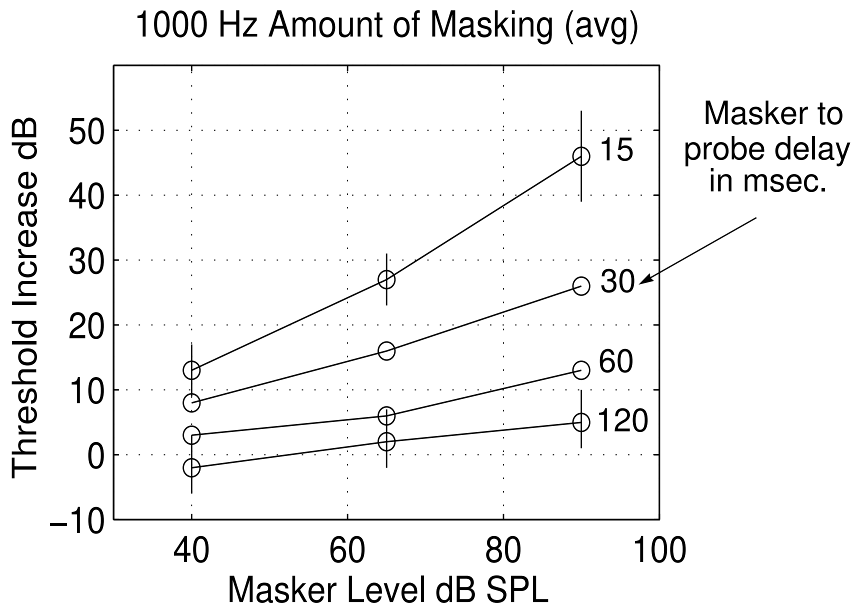


Figure 3.2 Amount of masking at 1KHz as a function of masker level, averaged across five subjects. Vertical lines indicate standard deviation.

As expected, the amount of masking increases with increasing masker level and decreasing masker delay.

As discussed previously, forward masking can also be viewed as the dynamic range below masker. This defines a reference relative to the masker level instead of the static threshold. At sufficiently short masker-to-probe delays, the dynamic range below masker specifies the distance of the I/O curve from threshold in our model. Figure 3.3 shows the data of figure 3.2 with a reference relative to the masker level.

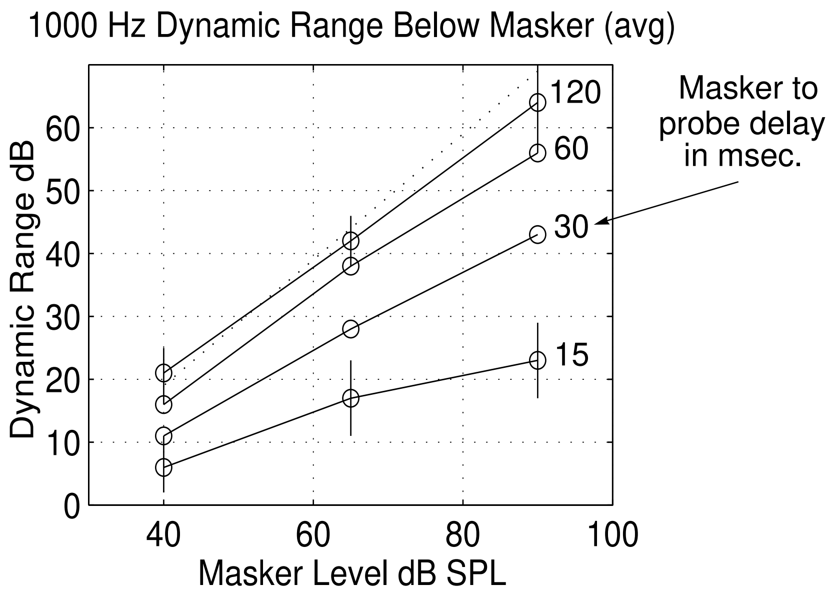


Figure 3.3 The dynamic range below masker at 1 kHz, averaged across 5 subjects.

A 45 degree diagonal on this graph indicates no forward masking; thresholds are only a function of their absolute level. A flat line would indicate purely relative thresholds; that is, thresholds are simply a fixed level below the masker level.

Finally, we view the amount of masking on a log time scale of the delay between masker and probe. Now the log of the delay time is the explicit independent variable. Constant masker levels form the contours on the graph. Figure 3.4 shows the 1 kHz average data in this form.

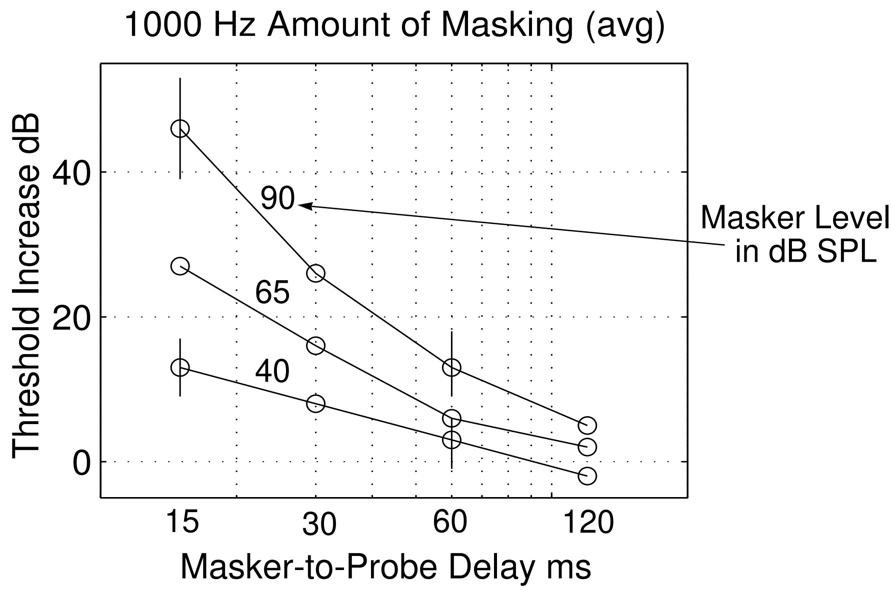


Figure 3.4 Average results for the amount of masking at 1 kHz as a function of the log of the delay between masker and probe.

Consistent with Plomb [1964], Jesteadt *et al.*, [1982], and Moore and Glasberg [1983], we measure an increase in the rate of decay of masking with increased masker levels. The slope of the lines with larger amounts of masking is steeper on this log-log graph, especially between 15 and 60 msec.

The following figures depict our experimental results in detail. Figure 3.5 shows the average amount of masking as a function of the level of the masker, across all frequencies measured. Figures 3.6 - 3.10 show the probe thresholds for each subject as a function of the level of the masker, across all frequencies. In these graphs, the static threshold is indicated by a dotted line.

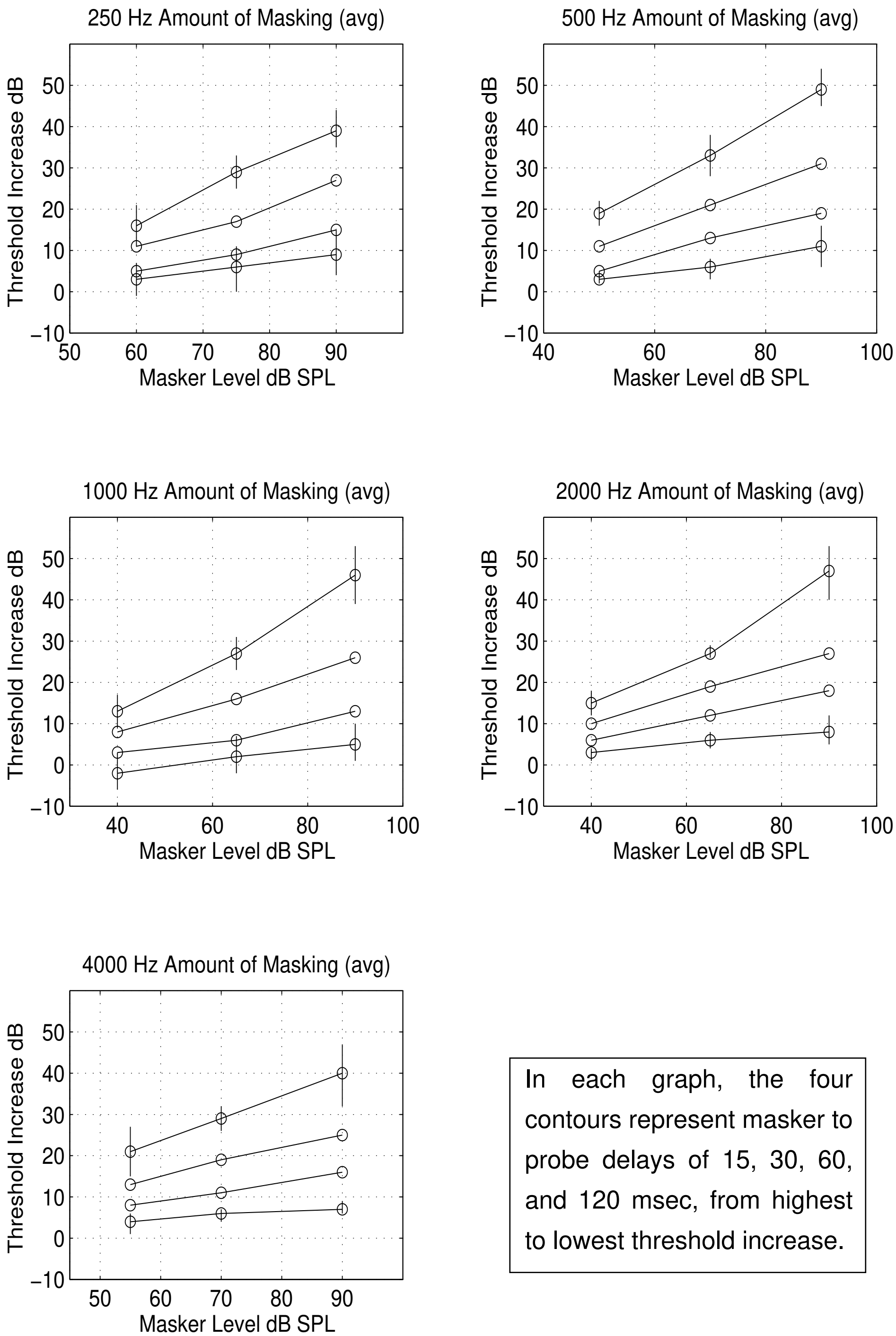


Figure 3.5 The amount of masking as a function of the level of the masker, averaged across 5 subjects.

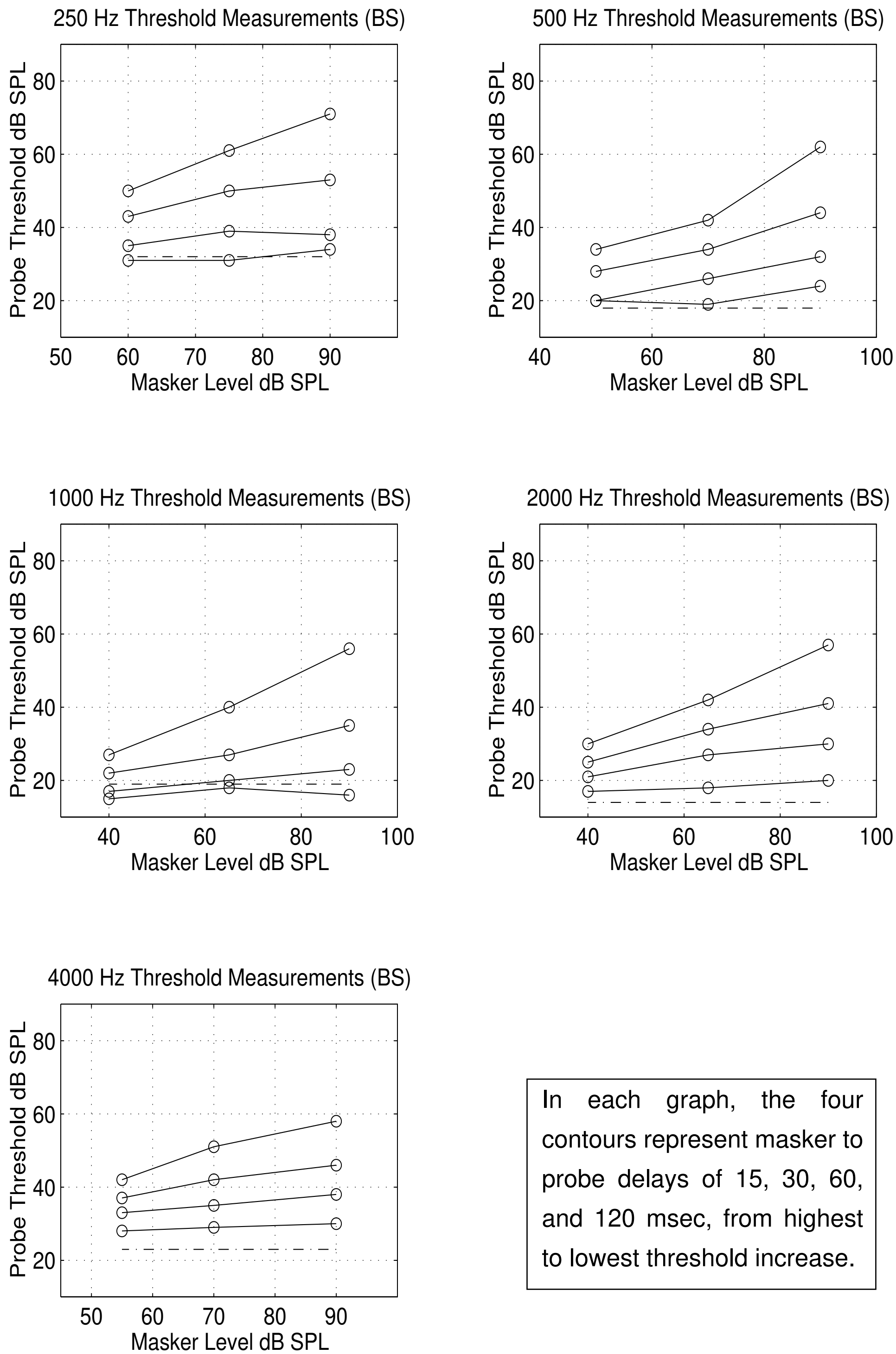


Figure 3.6 Subject BS: Probe threshold as a function of masker level.

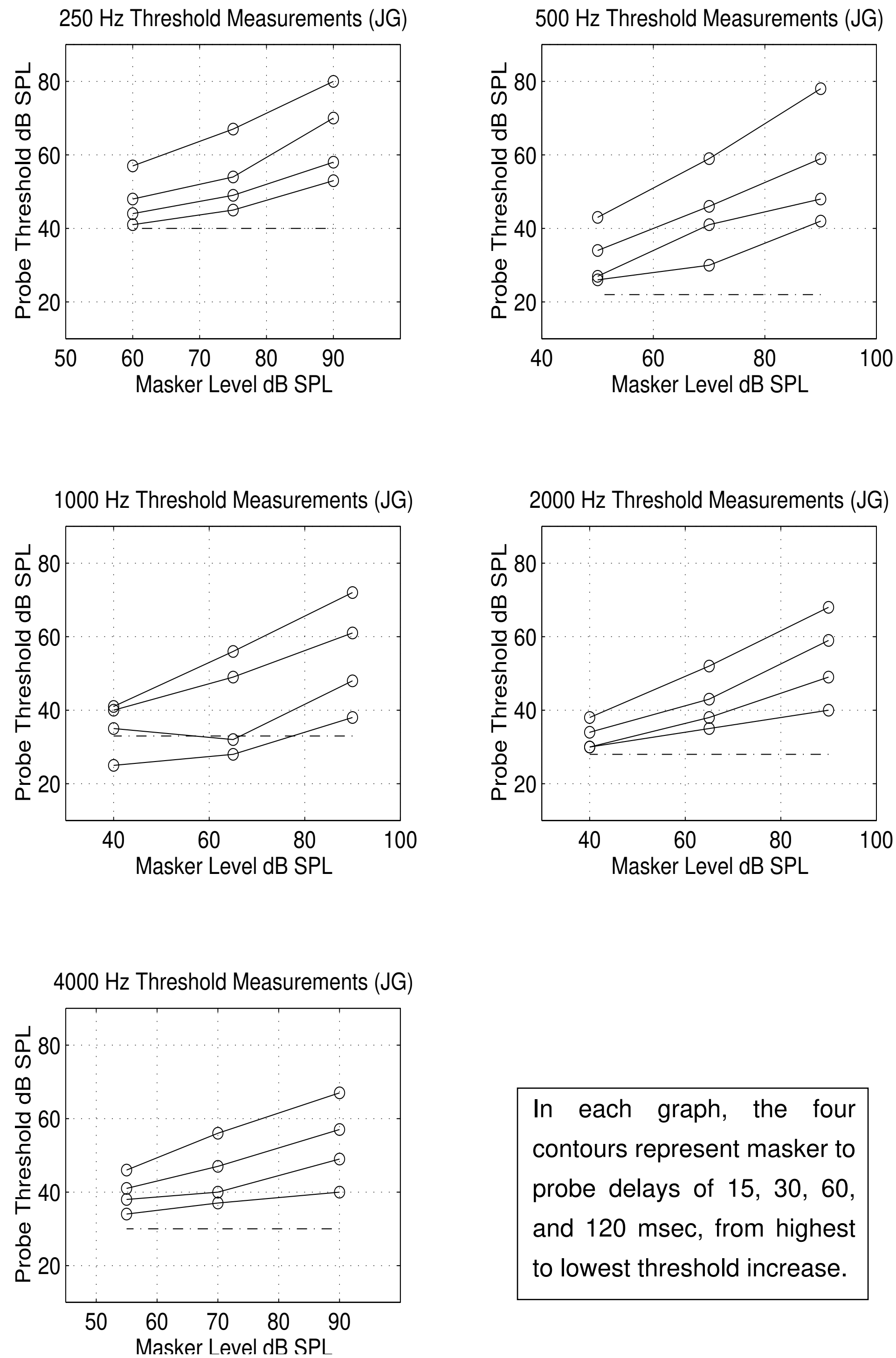


Figure 3.7 Subject JG: Probe threshold as a function of masker level.

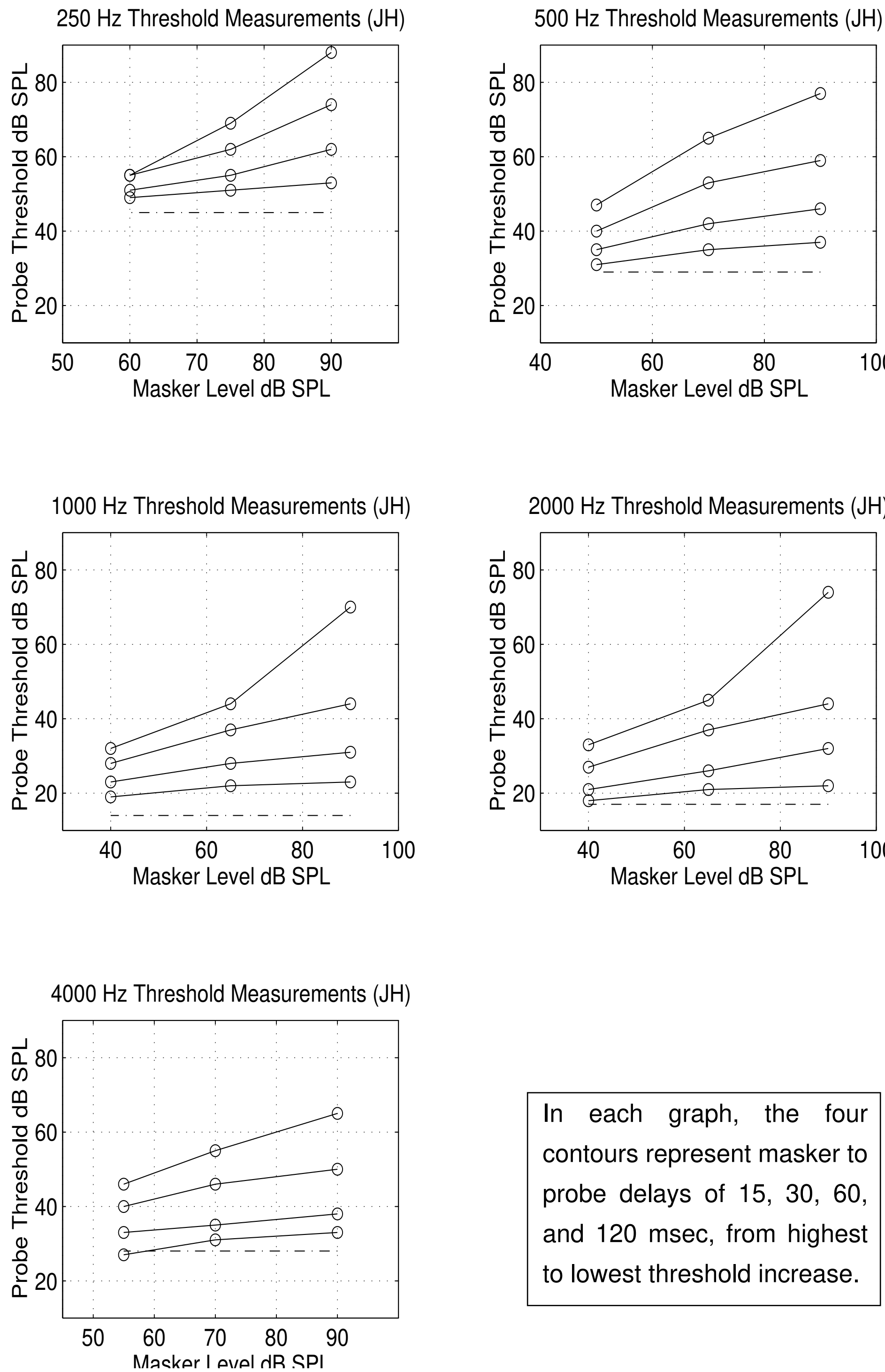


Figure 3.8 Subject JH: Probe threshold as a function of masker level.

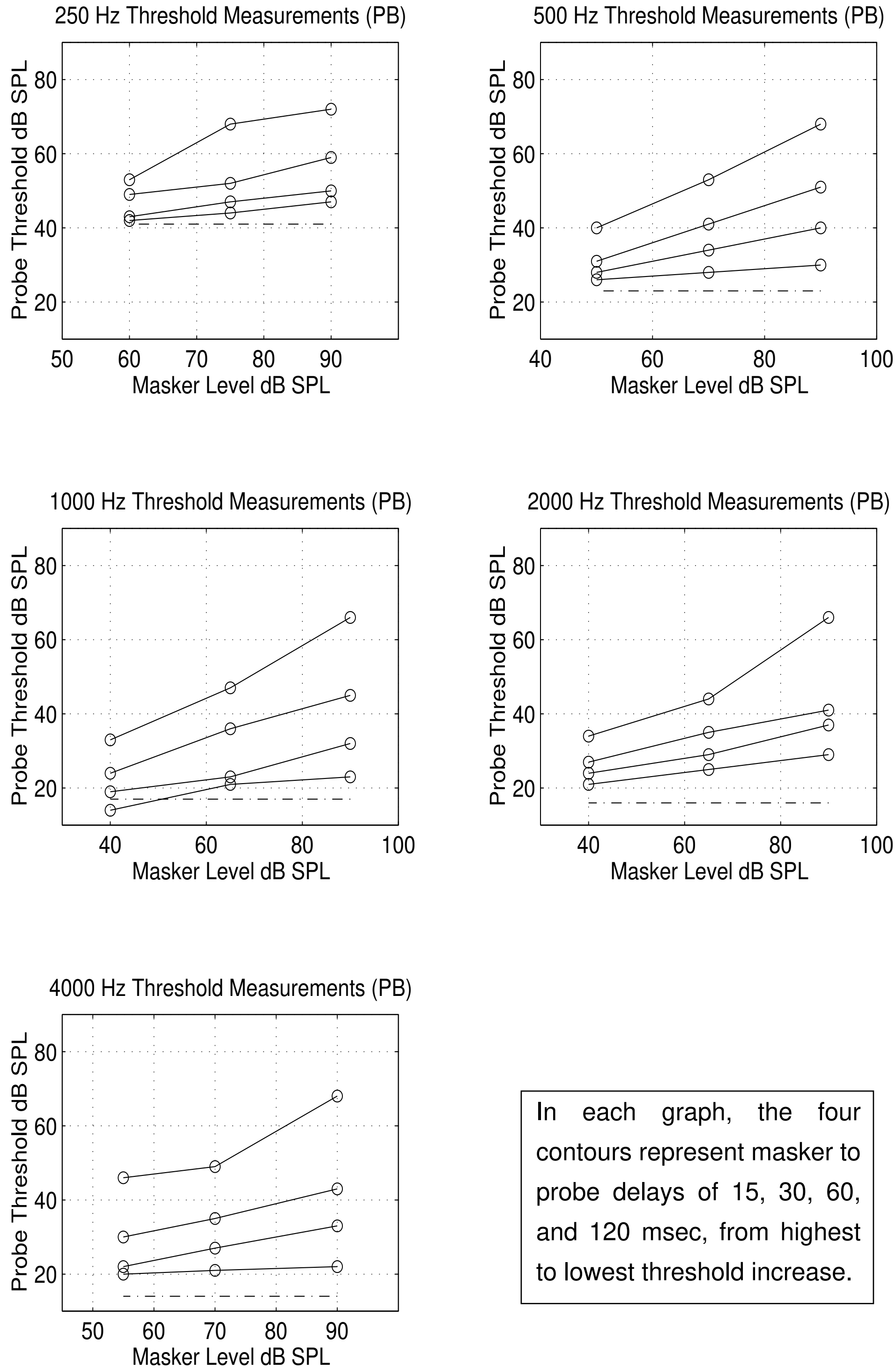


Figure 3.9 Subject PB: Probe threshold as a function of masker level.

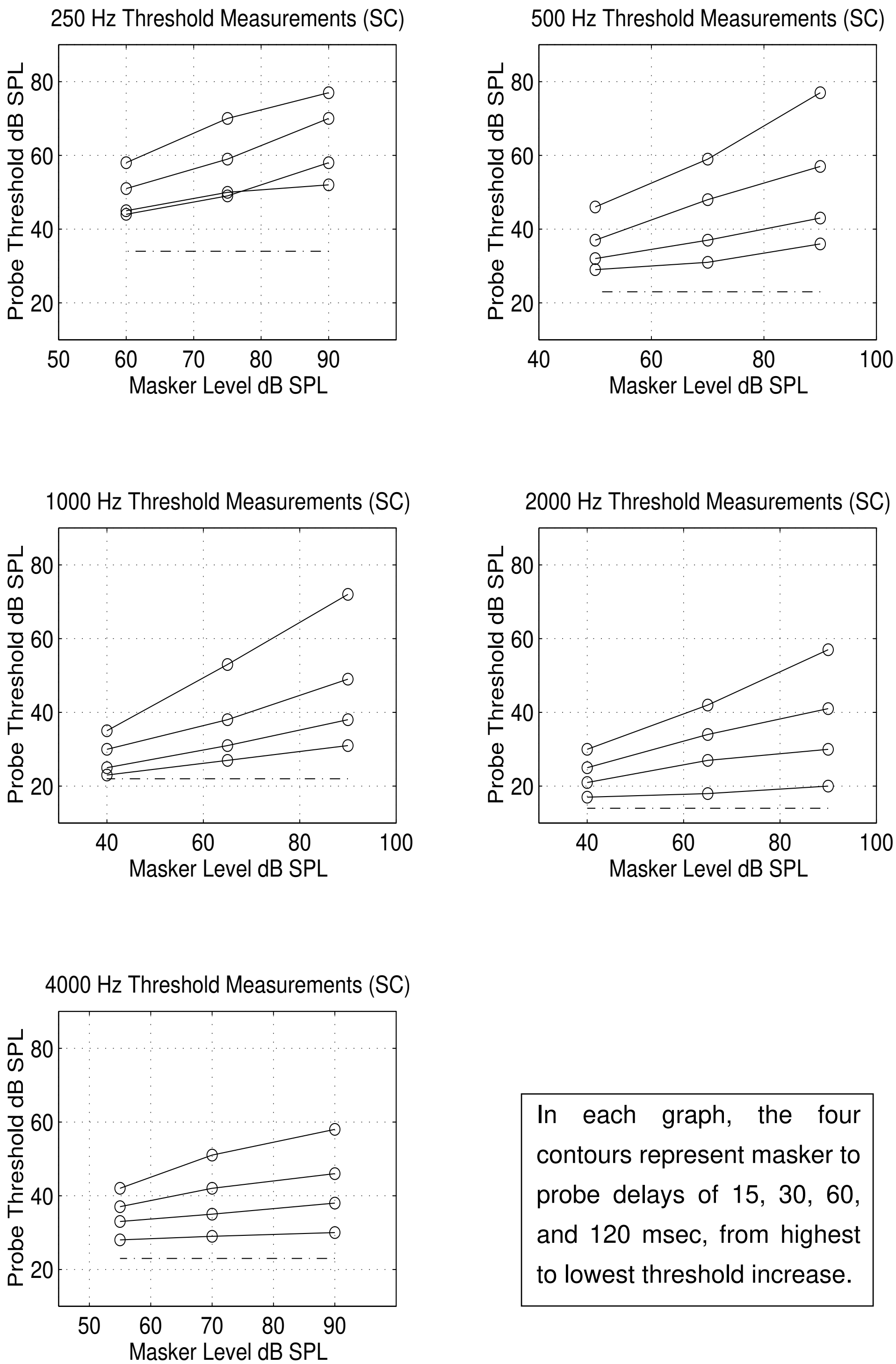


Figure 3.10 Subject SC: Probe threshold as a function of masker level.

## 3.4 Discussion of Forward Masking Results

Generally, the forward masking results here are very similar to those reported in other studies [Moore and Glasberg 1983], [Jesteadt *et al.*, 1982], [Plomb 1964], etc. The amount of masking increases as a fraction of the level of the increase in the masker, and as the delay between masker and probe increases, larger amounts of masking decrease more quickly than smaller amounts of masking. Figure 3.2 shows the fractional increase in the amount of masking, and Figure 3.4 shows the change of the amount of masking with masker to probe delay.

More specifically, data points in common with those from Jesteadt *et al.* [1982] are very closely matched. At 1 kHz, with an 80 dB SPL masker, and an effective probe delay of 15 ms, Jesteadt reports an average amount of masking of 39 dB. Interpolating an 80 dB SPL masker from the 90 and 65 dB SPL maskers in Figure 3.2, the data above reports an average amount of masking just under 40 dB. Further, with a 40 dB SPL masker, and a 30 ms probe delay, both data sets report an amount of masking just under 10 dB.

Jesteadt *et al.* [1982] proposes the following equation to predict the amount of forward masking as a function of masker level, and masker to probe delay:

$$A_m = a\,(b - \log \Delta t)\,(M - c)$$

where $A_m$ is the amount of masking, $\Delta t$ is the masker to probe delay, M is the level of the masker above threshold, and a, b, and c, are parameters chosen to fit the data. This equation predicts straight lines of increasing slope when forward masking data is viewed with logarithmic masker-to-probe delay as the independent variable. Figure 3.4 shows our average forward masking data in this format. Notice, if we only consider shorter masker to probe delays (to 30 or 60 ms), straight lines of increasing slope fit the data well. Jesteadt *et al.* [1982] does not consider masker to probe delays beyond 40 ms.

Figure 3.5 shows the average forward masking results at the 5 frequencies measured. Notice the amount of masking is greatest at center frequencies with the greatest dynamic range. If adaptation enables a large usable dynamic range, and adaptation necessitates short-term adjustment (forward-masking), it is consistent that middle frequencies with the greatest dynamic range also show the greatest amount of forward-masking. Also notice the variations from one octave to the next are relatively small. To choose adaptation parameters at each center frequency in the model proposed in this thesis, we interpolate between the adaptation parameters implied by the forward masking data above. If there were more abrupt discontinuities across frequency, we would require forward masking measurements at more frequencies.

Figure 3.5 also shows a rather abrupt increase in the amount of masking for 15 ms delays and 90 dB SPL masker levels at 1 and 2 kHz. Such an increase is consistent with the slope of an I/O curve in our model flattening, or increasing its compression, at high levels before becoming linear. Or perhaps, these points mark the beginning of a second, faster, adaptation process.

Finally, it is important to note the relatively large variance across subjects in the forward masking data. Moore and Glasberg [1983] noticed higher variance in tone after tone experiments when compared to tone after noise. He suggests subjects may have difficulty separating the masker and tone when they are both sinusoidal. The decaying sinusoids used in the experiments above may have added to that confusion.

In Chapter 4, we translate these experimental results into model parameters.

# Chapter 4

# From Experimental Results to Model Parameters

## 4.1 The Model and Forward Masking

As described in Chapter 2, normal human hearing has a usable dynamic range at middle frequencies of over 100 dB. However, over short time periods the dynamic range is much smaller. Forward masking experiments detailed in Chapter 3 measure short term dynamic range below the masker level. A mask signal precedes a probe signal each of a similar frequency. The threshold of the probe signal is a function of the time between the masker and probe, the level of the masker, and the duration of the masker, at least until that duration is greater than 100-200 msec [Zwislocki *et al.*, 1959].

The model proposed in this thesis consists of an adaptation block after each auditory filter. AGC with carefully chosen I/O curves models adaptation. Figure 2.10 (in Chapter 2) shows a prototypical I/O curve imposed over “output trajectories” corresponding to various rates of input level change. I/O curves have no time axis; however, changes of input level with time imply output trajectories. When the input changes slowly, the output trajectory follows the I/O curve. When the input changes instantaneously, the model is linear, the output trajectory moves diagonally, and then eventually drifts back to the target output. When a rapid decrease of the input level creates a trajectory that momentarily drops below threshold, the model predicts forward masking. Specifically, the model predicts that inputs at levels below the point where the output trajectory crosses the threshold, will be masked, until the trajectory drifts back above threshold. Figure 4.1 shows the adaptive model predicting forward masking following an 80 dB masking tone.

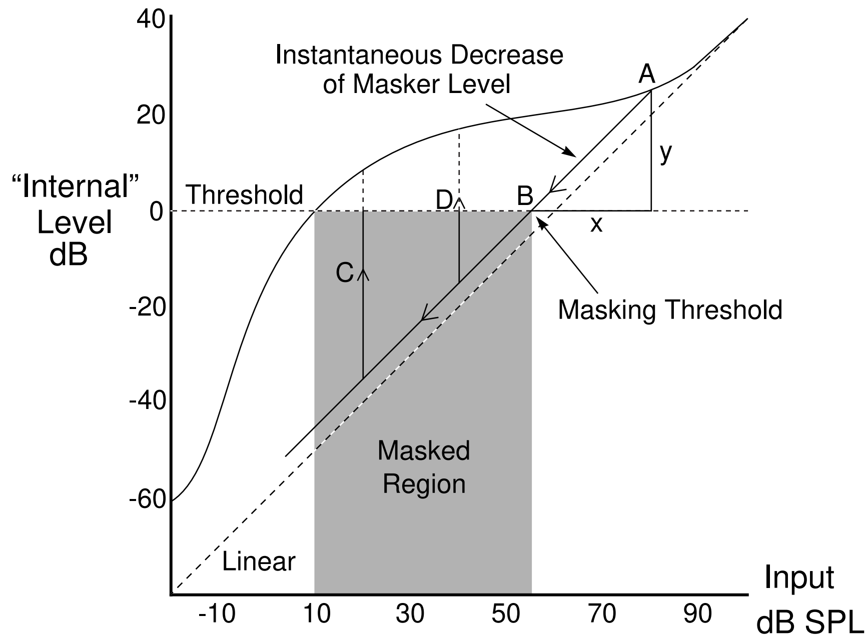


Figure 4.1 Model predicting masking of inputs below 55 dB SPL by a preceding 80 dB SPL masker.

Initially, the model has settled to the static target output corresponding to the 80 dB SPL masker (point A). When the masker shuts off abruptly, the output of the model drops linearly, and the trajectory moves instantaneously along the diagonal. Inputs at levels below the intersection of the diagonal with the internal threshold (see point B) are masked. Inputs immediately following the masker, exactly at the input level of this intersection, are at threshold. This defines an instantaneous dynamic range below masker as the difference between the masker level and the quietest sound still audible instantaneously after the masker. Because the model is instantaneously linear, these trajectories move diagonally, and the vertical distance y in Figure 4.1 is identical to the horizontal distance x. Therefore, the vertical distance from the I/O curve to the internal threshold defines the instantaneous dynamic range below masker of the model.

Figure 4.1 also includes two trajectories corresponding to probe tones of 20 and 40 dB SPL following the 80 dB masker. The rate of motion of the trajectory from the diagonal to the target on the I/O curve is proportional to the distance from the current position to the target. Points C and D are the halfway points from the diagonal to the target for the trajectories in response to the 20 and 40 dB probe signals, respectively. Notice it takes longer for the trajectory from the 20 dB SPL input to rise above threshold than for the trajectory from the 40 dB input. After half of the adaptation is complete, the 40 dB input is audible, however, the 20 dB input is still below threshold. As these trajectories rise through the threshold, they define the model's prediction of forward masking thresholds, as a function of the delay between the masker and probe.

Unfortunately I/O curves have no time axis, and we have resorted to discussing "output trajectories" to describe the model. The following figures show time as the independent variable. Although these figures do not completely characterize the model, they illuminate its functionality through examples. Figure 4.2 shows the responses of the model to a tone that abruptly drops from 80 dB SPL to a series of lower values. Time is linear on the left figure, and logarithmic on the right. The time origin is the abrupt change of the input. Before the abrupt drop of the input, the model has completely adapted, and reached the target on the I/O curve. At the abrupt transition, the output trajectory in Figure 4.1 drops instantaneously along the diagonal; in Figure 4.2 the model's output falls abruptly by an amount equal to the change in input level. After the transition, the output trajectory in Figure 4.1 drifts back to the target on the I/O curve; in Figure 4.2 the output exponentially adapts toward a new target output. When the output falls below the internal threshold (0 dB on the internal dB scale), the model predicts forward masking. As an output rises through the threshold, the model predicts a forward masking threshold. Figure 4.3 shows the response to a series of starting input levels and one ending input level.

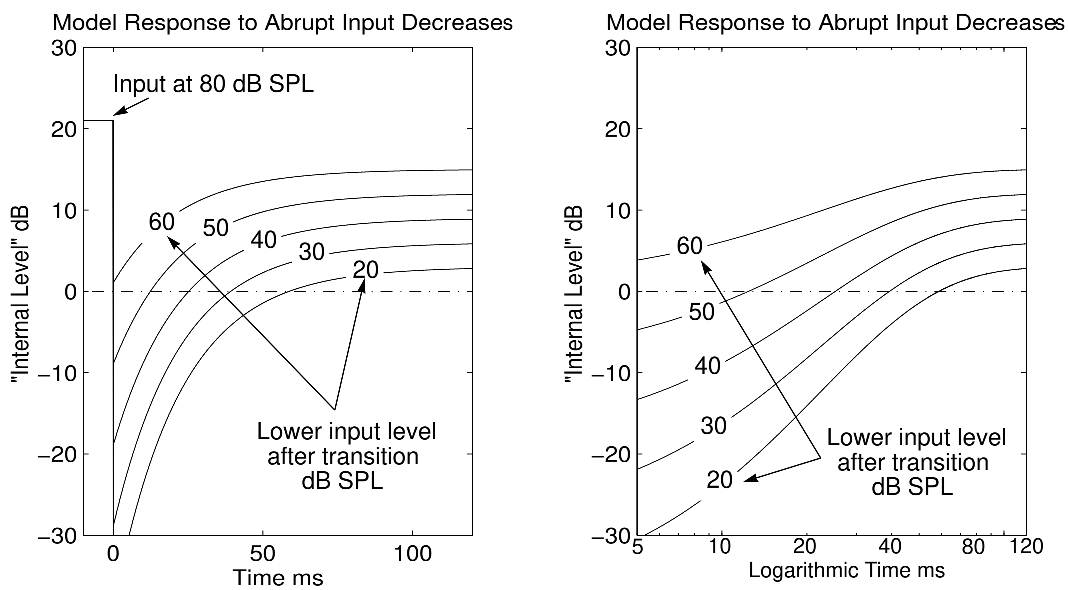


Figure 4.2 The input to the model abruptly drops from 80 dB SPL to a series of lower levels from 60 to 20 dB SPL. The model adapts to the lower level. While the model output is below the internal threshold, the model predicts forward masking. The left is a linear time scale, the right is logarithmic.

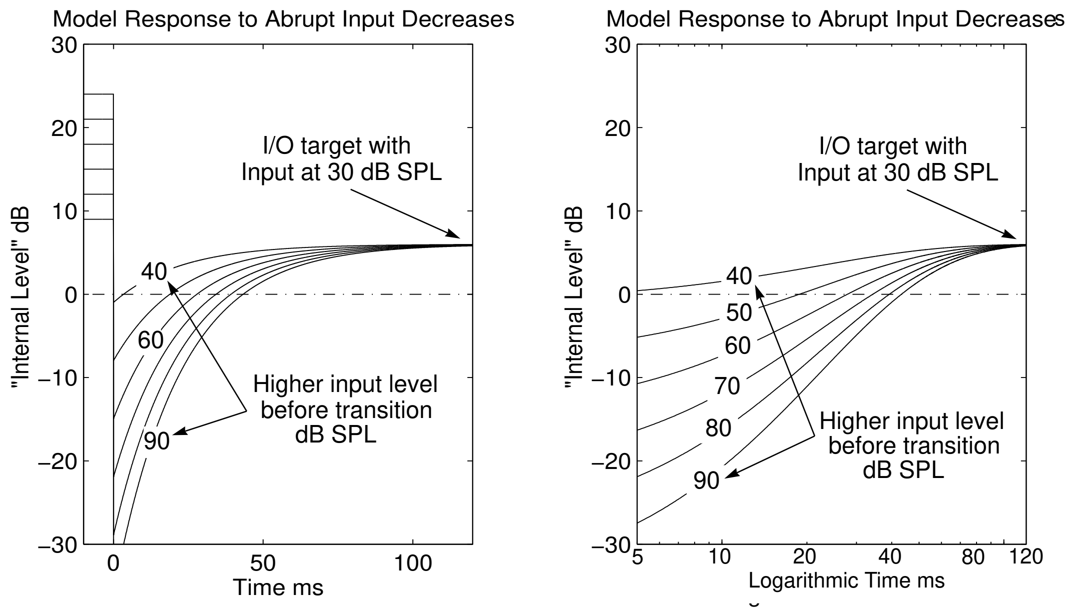


Figure 4.3 Model response to abrupt decreases in input level. A series of starting input levels from 90 to 40 dB SPL abruptly decrease to 30 dB SPL. The left is a linear time scale, the right is logarithmic.

## 4.2 Derivation of Model Parameters

The empirical forward masking data presented in Chapter 3 along with an understanding of how our proposed model predicts forward masking specify the model's I/O curves and time constants across frequencies. This first model uses piecewise-linear I/O curves, and one time constant for each adaptation block. Figure 4.4 shows the piecewise linear I/O curve with an output trajectory corresponding to an input instantaneously transitioning from 80 to 30 dB SPL.

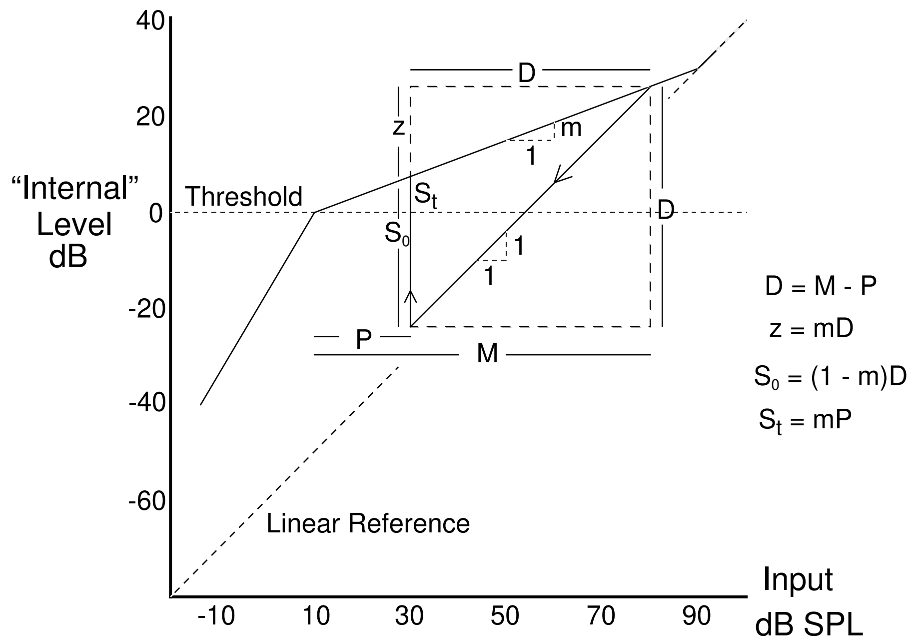


Figure 4.4  Piecewise linear I/O curve with an output trajectory corresponding to an instantaneously decreasing input.

When the output trajectory is below threshold, the model predicts masking, as the output trajectory crosses threshold, at a distance $S_t$ from the static output target, the model predicts the forward masking threshold. The model incrementally adjusts its amplification based on the difference between the current output and the target output for the current input. We implement the adaptation as a first order difference equation. Therefore, the distance between the output trajectory and the target on the I/O curve decays exponentially, at a rate determined by the amount of incremental adaptation. Figure 4.4 labels the geometry necessary to relate incremental adaptation and the slope of the I/O curve to the forward masking data.

Immediately after the abrupt drop of the input (by an amount D in Figure 4.4), the output is a distance $S_0$ from target. $S_n$ decays with discrete time $n$ as:

$$S_n = (1 - \alpha)^n S_0,$$

where $\alpha$ is the amount of incremental adaptation at each discrete point in time, and $S_0$ is the initial distance to target. When $S_n$ decays to $S_t$ as defined in Figure 4.4, the trajectory crosses the internal threshold. The time of this decay is the model's prediction of delay necessary such that a 30 dB SPL probe will just be audible after an 80 dB SPL masker. In Figure 4.4, $P$, and $M$ are the distance above threshold of the probe and masker signals. Substituting $S_t$ for $S_n$, noting that $S_t = mP$ and $S_0 = (1 - m)(M - P)$, we find the following relation:

$$mP = (1 - \alpha)^n (M - P)(1 - m) \qquad (1)$$

$M$, $P$, and $n$ are specified by the forward masking data. The two model parameters, $m$ and $\alpha$, are unknown. Therefore, two (different) forward masking data points are sufficient to define the model parameters. Defining $\beta = (1 - \alpha)$, we solve the equation above for the slope $m$,

$$m = \frac{(M - P)\beta^n}{P + (M - P)\beta^n} \qquad (2)$$

and the incremental adaptation parameter $\beta$.

$$\beta = \left(\frac{mP}{(M - P)(1 - m)}\right)^{1/n} \qquad (3)$$

Two forward masking points theoretically specify the model parameters m and β. However, an analytic solution of equation (1) from two arbitrary $\{P, M, n\}$ points is non-trivial. If $n$ is not equivalent for the two points, the resulting equations from equation (1) are of two, different, high powers of $n$. If $n$ is equivalent for the two points, we can use equation (3) with two values of $\{P, M\}$ to find an algebraic solution for $m$, and then β. Unfortunately, this leads to the trivial solution of $m$=1, for any two pairs of $\{P, M\}$. Instead, we find the solution iteratively. Equations (2) and (3) provide successive estimations of $m$ and β, given current values of β, $m$, and the data points $\{P, M, n\}_1$, and $\{P, M, n\}_2$. Starting with initial values of $m$=0.4, and β=0.998, the solution for m and β converges within 6 decimal places after roughly 20 iterations. Table 4.1 lists average solutions for $m$ and α across frequencies, using the forward masking data from Chapter 3. α is referenced to a 16 kHz sampling rate.

Table 4.1 Model Parameters

| Frequency Hz | Slope m | Adaptation α | "Time Constant" |
|---|---|---|---|
| 250 | 0.20 | 0.0026 | 55 ms |
| 500 | 0.26 | 0.0023 | 61 ms |
| 1000 | 0.26 | 0.0029 | 48 ms |
| 2000 | 0.31 | 0.0021 | 64 ms |
| 4000 | 0.34 | 0.0019 | 69 ms |

## 4.3 Model Performance

The time constant of Table 4.1 is defined as the amount of time for the output to settle within 2 dB of the target on the I/O curve, after an abrupt 25 dB change of the input [Dillon and Walker, 1982]. This time constant is not the rate of decay of the amount of masking. As mentioned above, we propose a general model of adaptation that predicts forward masking. The model predicts a changing rate of decay of masking by modeling a fixed rate of adaptation. Figure 4.5 shows the model prediction of the amount of masking as a function of masker to probe delay, with several different masker levels, and model parameters m=0.30, and $\alpha$=0.0028, referenced to a 16 kHz sampling rate. The left figure is a linear time scale, the right is logarithmic.

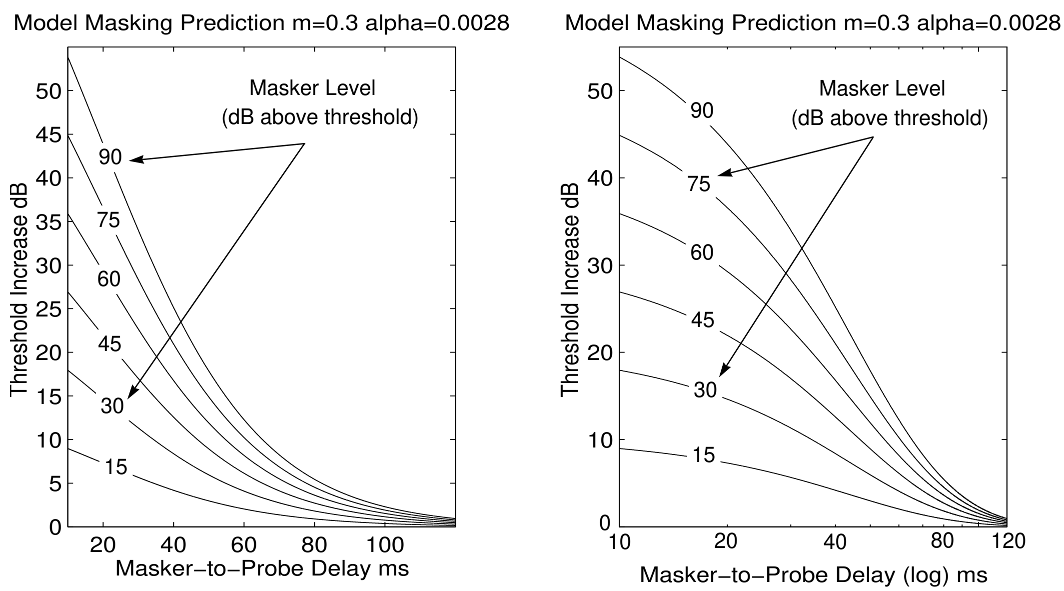


Figure 4.5 Model prediction of the amount of masking as a function delay and masker level (m=0.30, and $\alpha$=0.0028).

As expected, and consistent with our data, as well as that from Plomb [1964], and Moore and Glasberg [1983], the model predicts a faster decay of the amount of masking with increased masking. From 20 to 80 ms after the masker, the magnitude of the slope of the lines on this log/log scale increases with increasing masker level. Regardless of the initial amount of masking, after 100-200 ms, the amount of masking is negligible. However, as the delay between masker and probe approaches zero, the model has little time to adapt. Therefore, for short delays, the model predicts the difference between the I/O curve and the internal threshold, or the instantaneous dynamic range, as the amount of masking. Unfortunately, this "saturation" of the amount of masking to the instantaneous dynamic range below masker, is not consistent with forward masking data.

Figures 4.6 and 4.7 on the following pages impose the model's prediction of forward masking over the average forward masking data. Figure 4.6 shows masker level as the independent variable with constant masker to probe delay contours, Figure 4.7 shows delay as the independent variable with constant masker level contours. Across frequencies, the standard deviation of the model prediction error ranges from 2.5 to 3.3 dB.

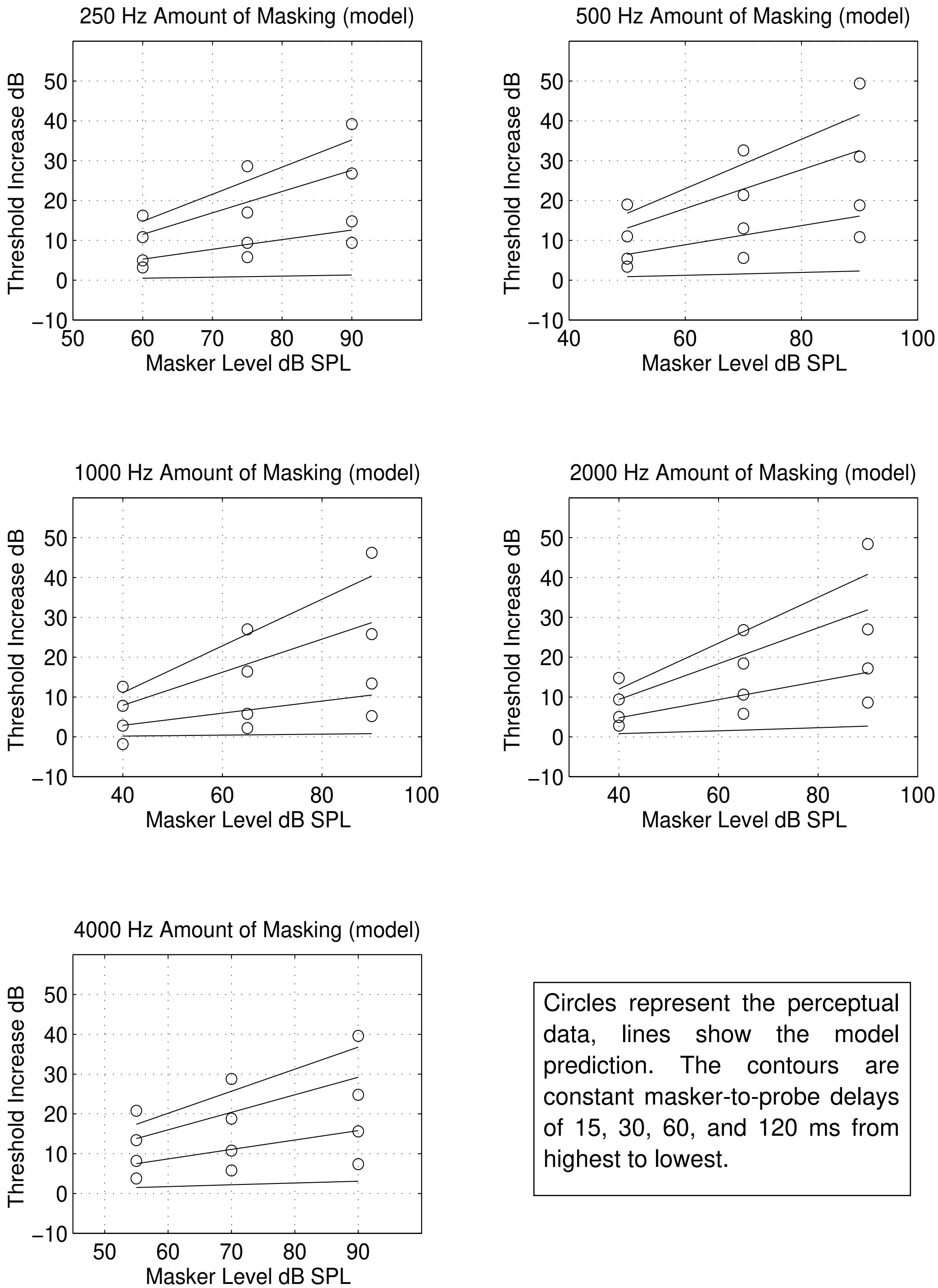


Figure 4.6 Model prediction of forward masking compared to averaged perceptual data.

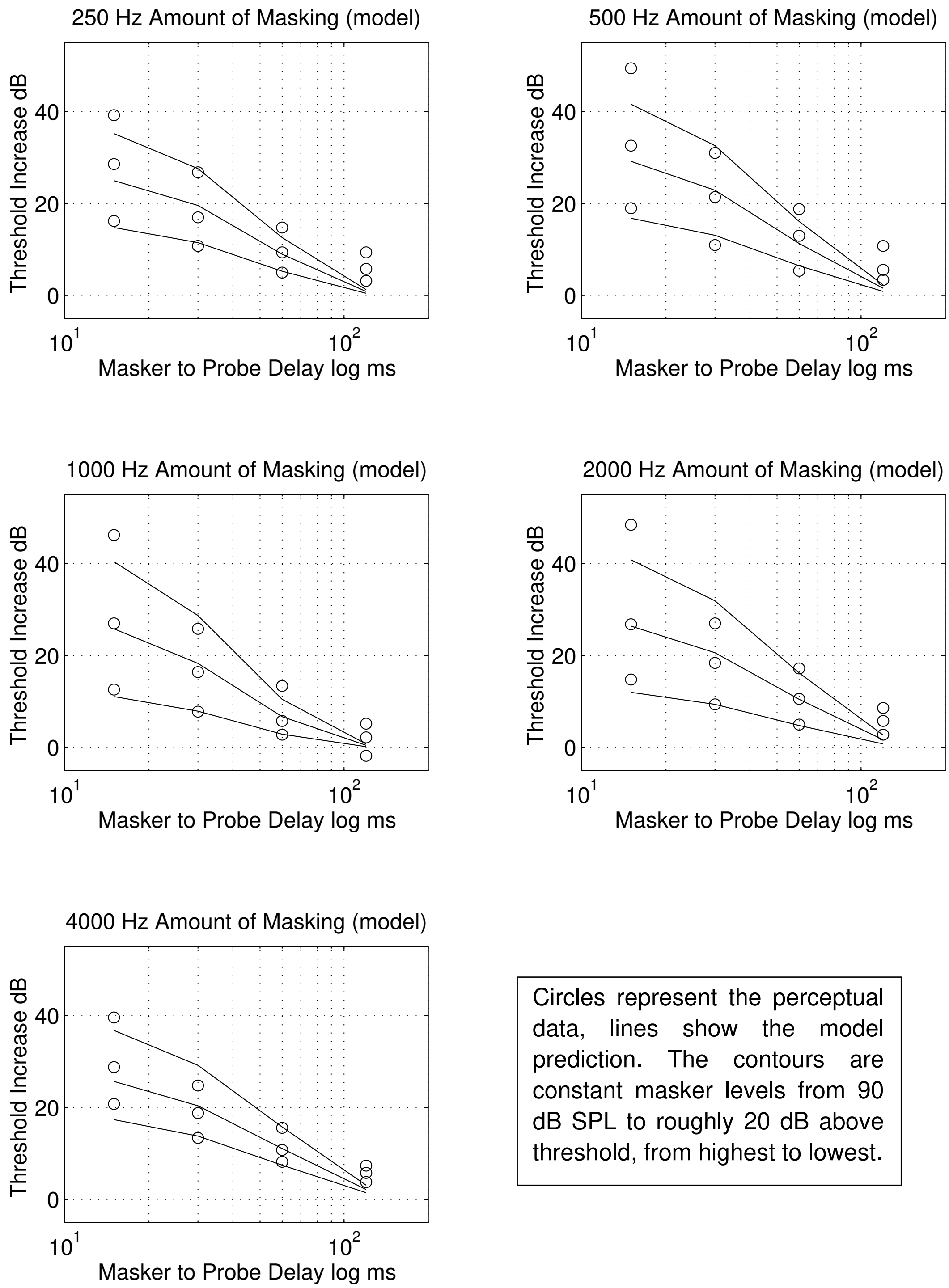


Figure 4.7 Model prediction of forward masking compared to averaged perceptual data.

## 4.4 Discussion

As shown in Figure 4.6 and 4.7, the quantified model fits the experimental data within a standard deviation of 3.3 dB. The model's deviation from the perceptual data, reflects the compromise of a single adaptation stage and a constant slope linear I/O curve. Specifically, Duihfuis [1973] recognized forward masking as the result of at least two processes, one of adaptation with a time constant near 75 ms, and a second with a time constant near 2 ms. This second process is responsible for the relatively sharp increase of the amount of masking when the delay between masker and probe is less than 20 ms. Further, Duihfuis [1973] shows a relation between the short-time process, and time-domain interactions between the probe and the ringing of the mechanical filtering of the basilar membrane.

Therefore, if we assume a two-stage process, and choose the model parameters based on forward masking data with masker to probe delays from 30 to 120 ms, the model's prediction is much closer to this data. Potentially then, multi-stage adaptation could increase the precision of the model's match to the data. Figure 4.8 compares the model prediction to the perceptual data, using only the longer-delay data to choose model parameters.

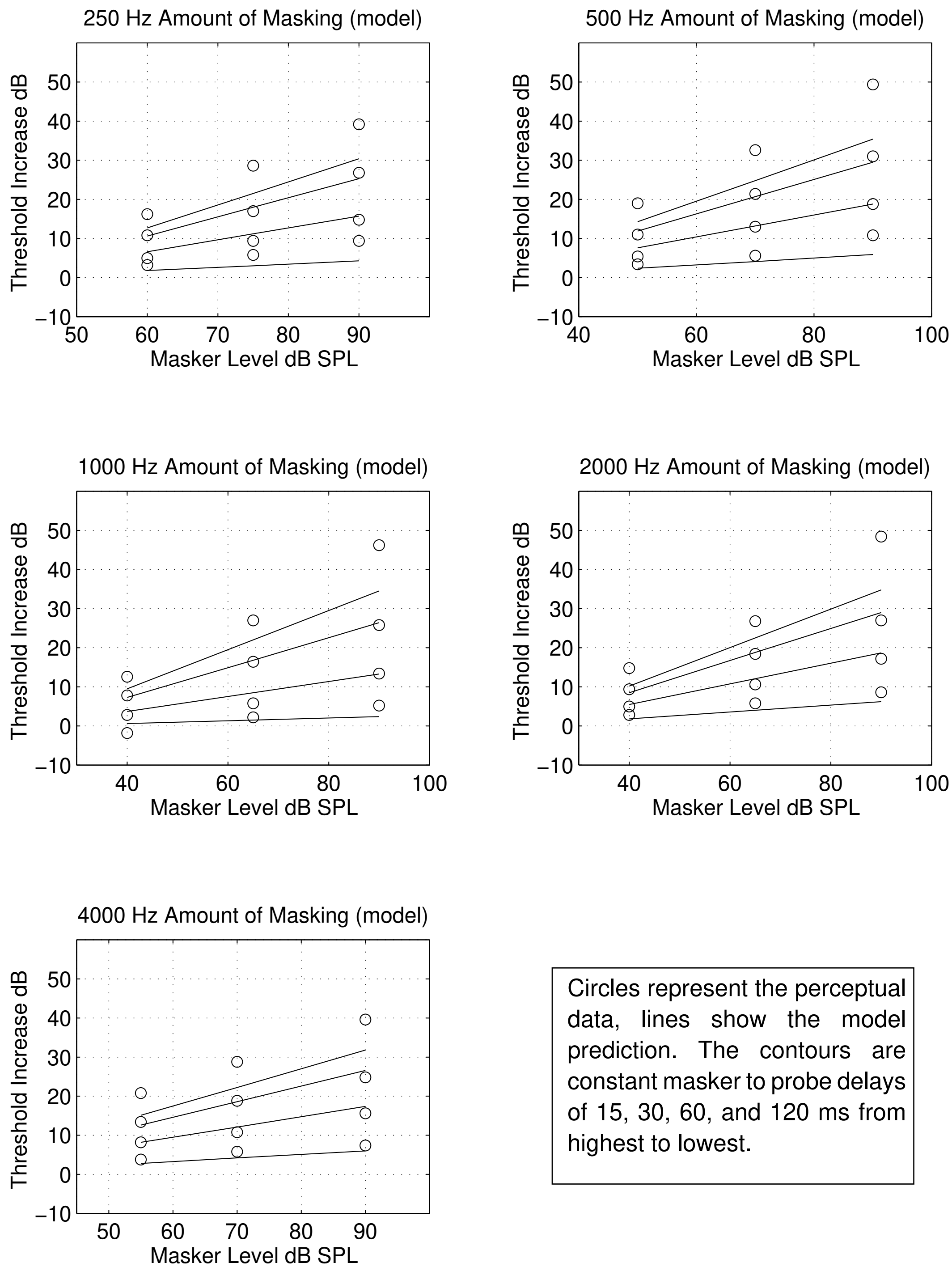


Figure 4.8 Model prediction, using only masker to probe delays from 30 to 120 ms.

## 4.5 The Model and Five Points of Forward Masking

Chapter 3 includes a description of five points summarizing forward masking data [Moore, 1989]. Forward masking: 1) increases as the masker-to-probe delay decreases, 2) increases as a fraction of the masker level, 3) increases as the duration of the masker increases, if that duration is sufficiently short, 4) decreases when the masker and probe are of differing frequencies, and 5) decays more quickly for higher-level maskers than lower-level maskers. In the following, we discuss these five points with respect to the model.

1) *Masker-to-Probe Delay:* When short-term linearity causes output trajectories to fall below threshold, the model predicts forward masking. As the masker-to-probe delay decreases, the model has less time to drift back above threshold, implying increased masking with decreasing masker-to-probe delay.

2) *Masker Level:* When the model predicts forward masking, the amount of masking is the distance from the masking threshold to the static threshold (the point where the I/O curve crosses the internal threshold). Higher masker levels shift the predicted masked threshold higher, increasing the amount of masking with masker level. However, the increased amount of masking with incremental masker level is controlled by the slope of the I/O curve. When the I/O curve is locally flat, a masker increment translates to an equal increment in the amount of masking, when the I/O curve is exactly diagonal (slope of 1), a masker increment translates to no increment in the amount of masking, in between, a masker increment translates to a fractional

increment in the amount of masking, consistent with Moore's second point.

3) *Masker Duration:* So far, we have only considered cases where the masker is long enough so that the model completely adjusts to its level. That is, the output, before the drop in level at the offset of the masker, has reached the target on the I/O curve. If the duration of the masker is sufficiently short relative to the time constant of the model, the model will not have completely adjusted to its level, by the time of the masker offset. The I/O curve specifies decreasing amplification with increasing input level. Therefore, as the model adapts to a masker level from a lower level, it slowly decreases the amount of amplification. Until the model has completely adapted to the masker, it amplifies the masker above its target output. If the masker shuts off before the model finishes adapting, the linear diagonal drop starts from the point above the target on the I/O curve. The diagonal extension of this line crosses the internal threshold at a lower point than it would have, had the trajectory started from the target on the I/O cure. Therefore, the model predicts less masking for shorter maskers, at least until the duration of the masker exceeds the time constant of the model.

4) *Frequency Difference:* The model proposed here uses independent adaptation on the output of each filter. If the masker and probe differ sufficiently in frequency, the masker will not affect the model's response to the probe. Therefore, the model predicts forward masking is dependent on the difference in frequency between the masker and probe.

5) *Changing Rate of Decay:* Finally, Moore describes how masking decays at different rates for different masker levels. Specifically, high-level maskers lead to large amounts of masking which decay at a faster rate than the smaller amount of masking from lower-level maskers. This phenomenon appears to require different time constants of adaptation for different level inputs-- hardly a time constant. However, this requirement is only necessary if we assume a model that explicitly tracks the amount of masking. Such a model would monitor the input, determine how much masking should occur for a given situation, and then reduce that amount of masking at a rate dependent on the input level. This complexity is the result of modeling the specific phenomena of forward masking as opposed to the more general process of adaptation.

Instead, our model adapts to changes in input level by slowly adjusting the amount of amplification to move the output closer to target values specified by the I/O curve. Forward masking occurs when output trajectories fall below threshold; this is a natural ramification of our model of adaptation. The discussion of how our adaptive model predicts varying decay rates is more complex. Figure 4.1 may be helpful for this discussion.

Initially, the amount of masking is the horizontal distance from the intersection of the diagonal trajectory with the internal threshold, to the static I/O curve intersection with the internal threshold. The model does not explicitly keep track of the amount of masking as a function of time. However, with increasing delay, probes of decreasing level lead to output trajectories that cross the internal threshold. The model predicts a time-varying amount of masking as the horizontal distance from the (last) probe signal that causes the output trajectory to just cross the internal threshold, to the static threshold. The adaptation of the model is exponential, in that at each incremental time step, the amount of adaptation is proportional to the distance from the current output point to the target on the I/O curve. Therefore, the greater the distance from the I/O curve, the higher the incremental adaptation. Because the I/O curve is slowly increasing for most of the audible range, at higher input levels the I/O curve is further from the internal threshold. Immediately after masking at higher levels, high-level probe signals lead to output trajectories that just cross threshold at points relatively far from the target on the I/O curve. The model has greater incremental adaptation to these high level probes further from the I/O curve. (Remember, the amount of masking is the distance from the last probe to cause an output trajectory to just cross threshold, to the static threshold.) As output trajectories corresponding to higher-level inputs cross the internal threshold, the model is adapting more quickly than it is when lower-level inputs cross the internal threshold. Therefore, the amount of masking decays more quickly from a higher-level masker, even though our model has a single (fixed) time constant.

## 4.6 Summary

We propose a first-order model of auditory adaptation consisting of a linear filter bank with independent AGC on each filter output. Static frequency selectivity data determine the filter bank parameters. A simple but complete set of sinusoidal forward masking experiments provide data necessary to quantify the adaptation for each AGC block in our model. Despite the compromise of a single adaptation stage, the model's predictions provide a reasonable first-order fit to the perceptual data (Figures 4.6 and 4.7). In addition, ramifications of this non-linear adaptive model are consistent with a wide range of physiological and psychophysical phenomena. We evaluate the benefit of the adaptive model, by incorporating it as the front-end to a speech recognition system.

# Chapter 5

# Recognition System

## 5.1 Overview

Speech recognition systems typically consist of a signal processing section followed by a pattern comparison and classification section. The signal processing section divides the speech signal into an observation sequence (usually a sequence of spectral estimations), and the pattern comparison section compares the observation sequence to previously observed sequences. Dynamic Programming, Neural Nets, Hidden Markov Models, and several variations and combinations of these, have been used as pattern comparison solutions, with HMM-based recognition systems as today's most popular choice. To evaluate the adaptive model presented in this thesis, we compare the performance of a dynamic programming-based recognition system using either the adaptive model signal processing, or other more common spectral extraction techniques. The dynamic programming solution provides a relatively simple and, perhaps more importantly, deterministic pattern comparison solution suitable for initial evaluation of the proposed model.

This chapter describes our baseline speech recognition system and provides a review of the signal processing techniques and considerations for speech recognition. Chapter 6 describes how we incorporate the adaptive model derived in Chapter 4 into the speech recognition system described here.

## 5.2 Common Spectral Estimation Techniques

Rabiner and Schafer [1978] provides detailed reviews and derivations of common short-time spectral estimation techniques used to analyze speech. These techniques involve windowing short overlapping segments of the input data stream, and then analyzing the windowed segments using the Discrete Fourier Transform, Linear Prediction (or Auto Regression), or more often, the corresponding cepstral representations.

### 5.2.1 Windowing

Traditionally, speech is viewed as a sequence of short-term static, or stationary, sounds. Any close examination of real speech waveforms, however, shows that the spectral shape of speech changes at least as often as it stays constant. Nonetheless, to simplify analysis, we usually assume speech is nearly stationary over a 10-30 ms segment, and analyze these individual segments, or frames. Speech is then represented as a sequence of spectral estimations. We define the following terminology: the input is the direct (un-windowed) data stream; the window is the shape multiplied by the input to extract a smaller quasi-stationary piece of data called a frame, or a time slice; the spectral estimation from each frame is a feature vector; the sequence of feature vectors forms the observation sequence.

Clearly, there is time/frequency resolution trade-off when we choose the window length. A shorter window provides good temporal resolution; a longer window provides finer frequency resolution. The shortest sounds in speech (stops) are as short as 5 ms, while vowels range from 40 to 200 ms (or longer). Typically speech recognition systems use window lengths between 10 and 30 ms [Rabiner and Juang, 1993]. The starting point for the next frame is usually 1/4 to 1/2 of the window length ahead of the current frame starting point, so that adjacent frames overlap significantly.

Multiplying the input by a window in the time domain is identical to convolving the spectrum of the window with that of the signal in the frequency domain. Rectangular windows in time are sinc functions in frequency; they have a narrow main-lobe, but high side-lobe peaks. After convolution in frequency, a rectangular window implies the least amount of local "blurring," from the narrow main lobe, and the most amount of "splatter" from the high side-lobe peaks. Raised-cosine (Hamming or Hanning) windows increase the width of the main-lobe and decrease the height of the peak of the side-lobes. After convolution in frequency, these trade more local “blurring” for less “splatter.” Most recognition systems use overlapping raised-cosine windows to extract time slices from the continuous input stream. Figure 5-1 shows an example where successive windows overlap by half of their length:

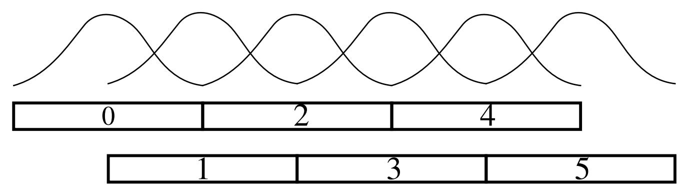


Figure 5.1   Overlapping Raised-Cosine Windows

### 5.2.2 DFT, LPC, and Cepstral Representations

Fant [1960] proposes a simple linear model of voiced speech production: a periodic driving function from the glottis excites a series of resonances from the vocal tract. For unvoiced speech a random-noise generator, from the turbulence associated with a narrow constriction in the vocal tract, excites the resonances of the cavities in front of the constriction.

The driving function of voiced speech approximates an impulse train in time, and therefore also approximates an impulse train in frequency. The spacing of the impulse train in time implies the fundamental frequency, or pitch; the evenly spaced impulses in frequency are the harmonics of the fundamental. A male speaker with a pitch near 100 Hz, will generate harmonics at integer multiples of the fundamental (200 Hz, 300 Hz, 400Hz...). The voiced driving function excites the vocal tract resonances. In time, this implies convolution of the driving function with the impulse response of the vocal tract, and in frequency it implies the corresponding multiplication. Therefore, the harmonics of the fundamental are weighted by the frequency response of the vocal tract.

If the analysis window length is long enough to resolve the harmonics of the driving function, the spectral estimation of the DFT represents both the driving function harmonics, and the vocal tract spectral envelope. The following figure shows the DFT spectral estimation of a single time-slice from the “ee” in “three” from a female speaker. In the time domain, notice the weighting of the Hanning window, in addition to the general periodicity of the input. In the frequency domain, the harmonics are spaced roughly 5 per 1000 Hz, or near a fundamental frequency of 200 Hz. The general spectral envelope, reflecting the resonances of the vocal tract, has peaks (or formants) at roughly 400, 2400, and 4000 Hz. The speech is sampled at 11025 samples/second, and the Hanning window length is 23.2 ms (256 point).

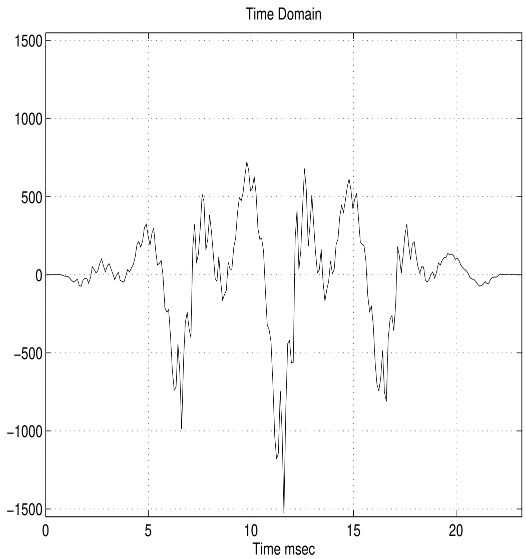


Figure 5.2 Time and frequency-domain representations of a windowed time-slice of “ee” in “three” from a female speaker.

With first-order accuracy, LP (linear predictive) analysis provides an estimation of the vocal tract transfer function. LP analysis imposes the assumption of a flat spectrum driving function, and then specifies an all-pole system that resonates the flat-spectrum driving function to produce the resulting output. Our linear model of speech uses either impulses or white-noise as a driving function. Therefore, LP analysis provides a crude form of deconvolution, where the LP coefficients specify the transfer function of the vocal tract, and the LP error, or residual, provides an estimate of the driving function. This is a key separation for speech recognition. In most languages, the resonances of the vocal tract, and not the variances of the driving function, provide the primary cues for speech recognition.

Cepstral, or homomorphic, analysis is also a form of deconvolution.

Practically, real cepstral coefficients are computed as the inverse DFT of the log magnitude of the DFT of the time domain data. The logarithm before the IDFT changes the convolution operation in time, and multiplication operation in frequency, into an addition operation in the cepstral domain. Therefore, provided the representations of the driving function and vocal tract are reasonably well separated in the cepstral domain, cepstral analysis provides a mechanism to isolate the driving function from the transfer function.

There is a second valuable interpretation of cepstral analysis. As an approximation of the KL-Transform, cepstral analysis takes the DCT of the highly correlated log magnitude spectrum shown in Figure 5.2. The log magnitude (in frequency) is even about the origin, so only the cosine terms in the IDFT will have non-zero contributions. Jain [1989] shows that the DCT of a well correlated Markov-1 sequence approaches the KL-Transform. Therefore, the IDFT of the log of the DFT provides an orthonormal transformation into a vector space where transform coefficients are highly decorrelated, and where the variance of transform coefficients is only significant for the first few coefficients. Euclidian distances are consistent through orthonormal transformations. Therefore, the energy compaction of a cepstral representation significantly reduces the computational complexity of comparing different spectral estimations. It's much easier to compare the first 10-20 cepstral coefficients than an entire spectral vector of 100-200 points.

More pragmatically, the definition of the IDFT is within a complex conjugate of the definition the DFT. The log spectral energy in Figure 5.2 shows the summation of a slowly varying (across frequency) spectral envelope and a ripple-like, rapidly varying component. The DFT of the log DFT would, therefore, reflect a low frequency envelope with a few terms near DC, and the high frequency ripple with a spike at the frequency of the ripple. The IDFT of the log DFT, or the real cepstrum provides a similar decomposition.

Finally, emphasizing or weighting cepstral coefficients can improve the performance of speech recognition systems [Rabiner and Juang, 1993]. If cepstral representations achieve good energy compaction, contributions from high-order coefficients are insignificant and are therefore discarded. Further, low order cepstral coefficients reflect the contribution of the corresponding basis functions of the DCT in the log frequency domain: the first coefficient represents total energy (DC level in log frequency), the second represents overall spectral tilt (1/2 oscillation over log frequency), etc. Because the absolute level, spectral tilt, and other lowest-order cepstral coefficients may not be significant for spectral comparison for recognition, these are often de-emphasized or discarded. The process of weighting, or choosing which cepstral coefficients are significant for recognition is called cepstral liftering. A raised-sin cepstral lifter of the first 10-20 cepstral coefficients gradually de-emphasizes the lowest and highest coefficients, providing a reasonable representation for speech recognition [Rabiner and Juang, 1993]. Implications of cepstral liftering are described in more detail in Chapter 6.

Many speech recognition systems use cepstral coefficients from LP-based spectral estimation, instead of those from DFT-based spectral estimation. That is, they use the cepstral representation of the transfer function estimation provided by LP analysis. Rabiner and Schafer [1978] describe a computationally efficient recursion that provides LP-based cepstral coefficients directly from LP coefficients. Together with Durbin's recursion for LP analysis, this provides a computationally tractable technique to transform a time slice into a cepstral representation suitable for recognition.

To view the spectral estimation implied by cepstral representations, we transform back from the cepstral domain to the log frequency domain through the DFT. Figure 5.3 shows the same "ee" in "three" as Figure 5.2, and includes the spectral estimation from 12th-order LP analysis, as well as the spectral estimation implied by the truncation of the first 16 LP-cepstral coefficients. The LP spectral estimation clearly marks the spectral peaks of the vocal transfer function. Further, the spectral estimation from the truncated LP-cepstrum is a "smoothed" version of the LP estimation; after cepstral truncation, any rapid changes across frequency are removed.

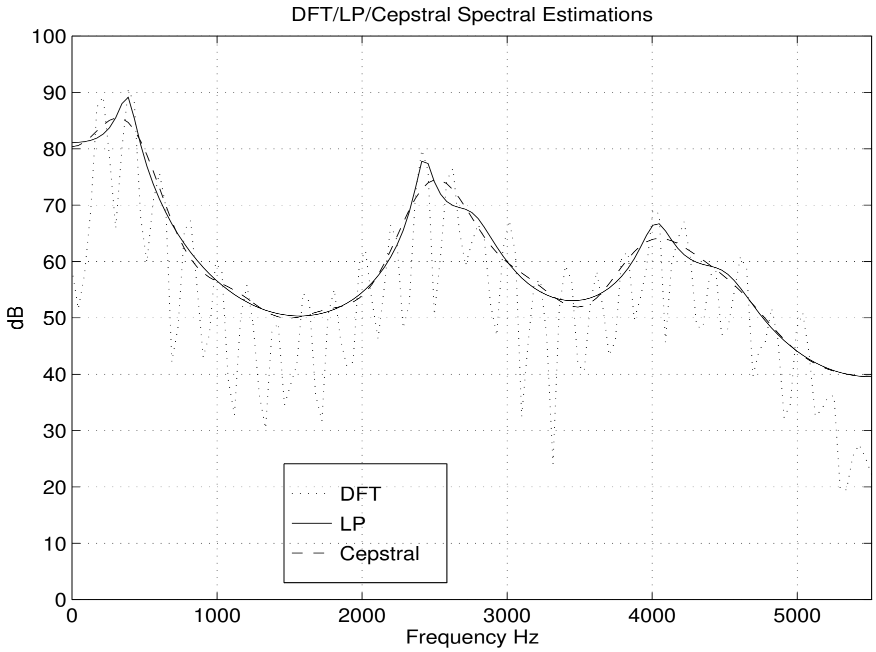


Figure 5.3 Comparison of DFT, LP, and LP-Cepstral spectral estimations.

Figure 5.4 shows the sequence of spectral estimations, for the entire word "three" from a female speaker. Time is horizontal, frequency is vertical, and intensity is mapped to darkness. Notice that the LP and LP-cepstral representations are increasingly "blurred" vertically. As in Figure 5.3, only the DFT analysis retains voicing information, and the LP-cepstral estimation is a smoothed version of the LP estimation. Spectral estimations are computed every 11.6 ms, using the same 256-point raised-cosine window. The grey scale is normalized for each: the highest point in each spectrogram is scaled to pure black, and pure white thresholds are set to a consistent level across spectrograms.

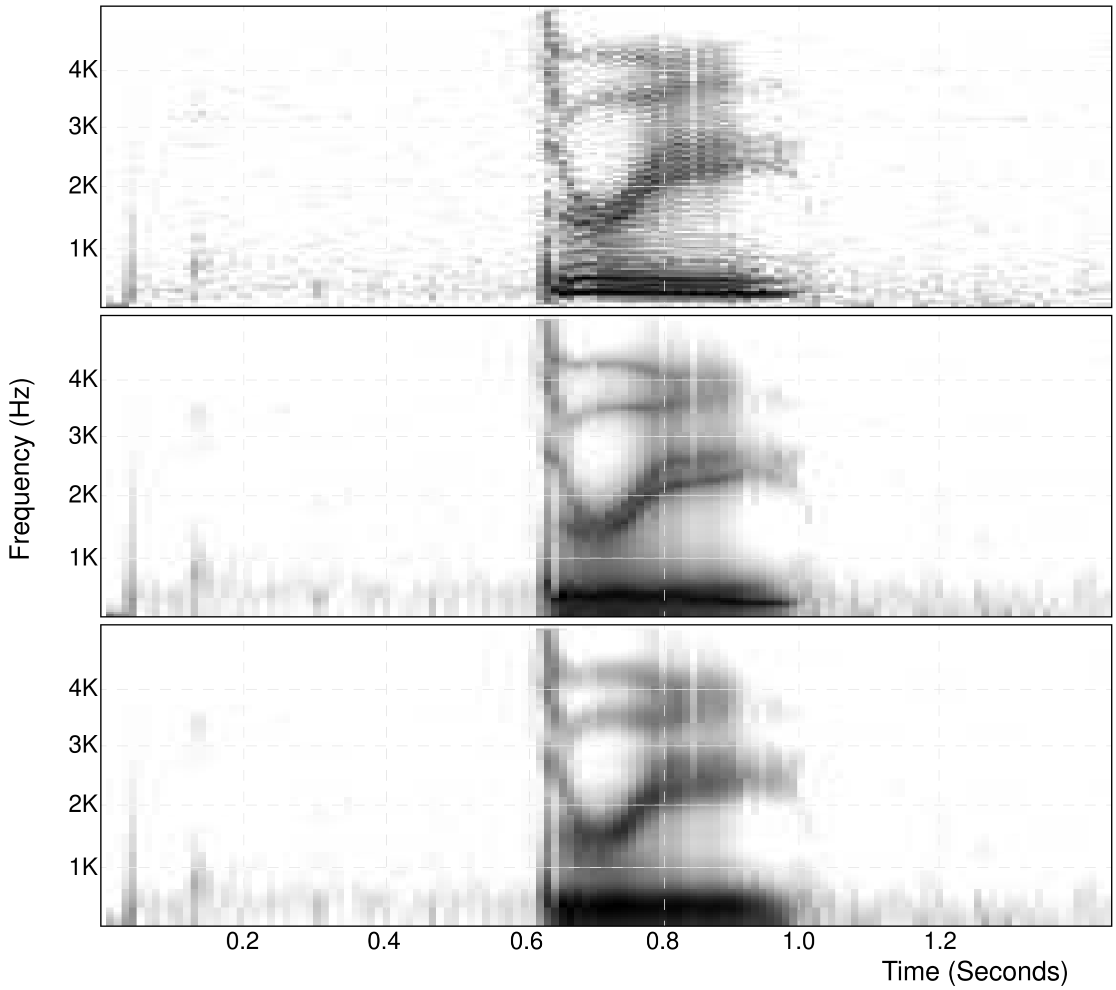


Figure 5.4 Spectrograms from DFT, LP and LP-Cepstral analysis of the word "three" from a female speaker.

The recognition system uses these sequences of spectral estimations, or spectrograms, as the input for higher-level pattern comparison and classification. Dynamic Time Warping accumulates minimum slice-by-slice spectral distances to find the closest spectrogram match. Chapter 6 describes how perceptually-relevant representations from our adaptive model improve recognition performance.

## 5.3 Comparison with Templates

Our recognition system compares the current sequence of spectral estimations (the candidate) to those from previous recordings of the vocabulary set (templates). Pattern comparison with Dynamic Time Warping [Rabiner and Juang, 1993] involves two levels of comparison: a slice-by-slice comparison called the local distance, and the accumulation and propagation of local distances to form the accumulated or total distance. The slice-by-slice comparison measures the similarity between two spectral estimations. The accumulated distance reflects how closely the sequence of spectral estimations matches those from a template. Dynamic Time Warping finds the horizontal (or time) “stretching or compressing” of the template spectrogram that minimizes the accumulated distance to the candidate.

### 5.3.1 Local Distance Metric

Perhaps the most common local distance metric is the $L_2$ norm, also called the Euclidian distance. The Euclidian distance between two K-dimensional spectral estimations $S_1(k)$ and $S_2(k)$ is defined as:

$$D = \sum_{k=0}^{K-1} [S_1(k) - S_2(k)]^2$$

The Euclidian distance is mathematically convenient, however, it may have little perceptual relevance.

Mathematically, one of the nicest properties of a Euclidian distance, is that it is consistent across transformations between orthonormal vector spaces [Rabiner and Juang, 1993]. That is, to find the spectral distance implied by two cepstral vectors it is not necessary to transform back to the spectral domain. We simply find the equivalent Euclidian distance in the cepstral domain.

Perceptually, a Euclidian distance will certainly grow for two widely varying spectral shapes, however it is not at all clear that the growth of the Euclidian distance with diverging spectral vectors is even remotely consistent with the growth of the perceptual difference. Most notably, at least for vowels, the frequency position of vocal tract resonances is much more salient than their amplitude, bandwidth, spectral tilt, overall level, and other spectral differences to which the Euclidian distance is sensitive [Klatt series: 1979-1982]. Cepstral liftering, by de-emphasizing insignificant spectral characteristics, before the Euclidian distance, provides a more perceptually-relevant distance metric. Chapter 6 discusses this in more detail, however the equation for the raised-sin cepstral lifter [Rabiner and Juang, 1993] is:

$$w(n) = 1 + \frac{L}{2}\sin\left(\frac{n\pi}{L}\right)$$

for $0 < n < L + 1$ and,

$$w(n) = 0$$

otherwise (where L is 10 ~ 20).

### 5.3.2 Accumulated Distances

Not only do speaking rates vary but rates of the individual parts of speech also vary significantly. Therefore, in addition to choosing a measure of distance between two static spectral estimations, we must also choose a method to stretch or compress the template sequences of estimations, so that the local distances are comparing “meaningful” elements in the sequence of estimations. Dynamic Time Warping, dynamically warps the template spectrogram (horizontally) to minimize the total distance from the candidate.

Figure 5.5 motivates path propagation, or accumulated distances, and the dynamic programming solution to the problem of time alignment. The vertical axis is the time index into the template sequence of spectral estimations. Similarly, the horizontal axis represents the time index of the candidate sequence of spectral estimations. The values listed on the graph are the local distances between the two associated spectral estimations. Point (2, 3) is the local distance (8 on the graph) between the second spectral estimation in the candidate sequence, and the third spectral estimation in the template sequence.

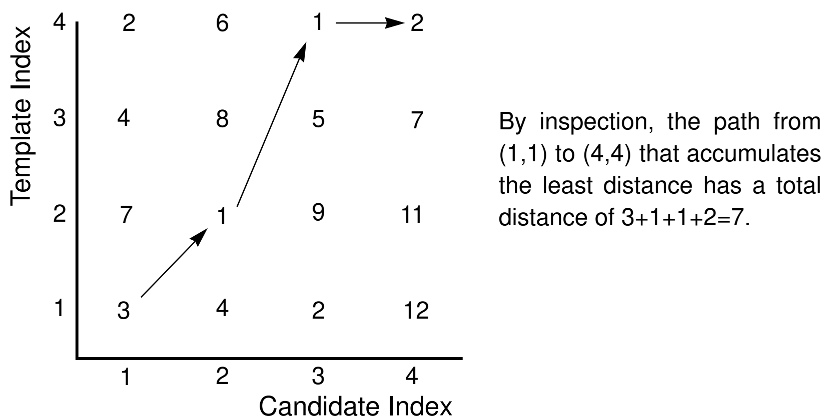


Figure 5.5 DTW finds the path through the field of local distance that accumulates the least distance, subject to a path propagation constraint.

Obviously, we must also choose a meaningful path propagation constraint, otherwise, the path with the minimum distance is simply a direct jump from the (1,1) point to the (4,4) point. Our recognition system uses the "Itakura" constraint [Rabiner and Juang, 1993], which allows for template stretching by a maximum of 2, and template compression by a minimum of 1/2. Viewed locally, paths must propagate to the next candidate index, but the template index can either stay the same, increase by one, or increase by two. Further, the template index can not stay the same twice in a row.

On the surface, it may appear that if the path must start at (1,1) and end at (4,4) the obvious solution is to propagate one path from (1,1) and simply choose the smallest next point within the propagation constraint. This approach would work well for the example in Figure 5.5. However, in other fields, propagating a single forward path and choosing the minimum distance at each point can restrict the path from propagating through smaller local distances in the future. For example:

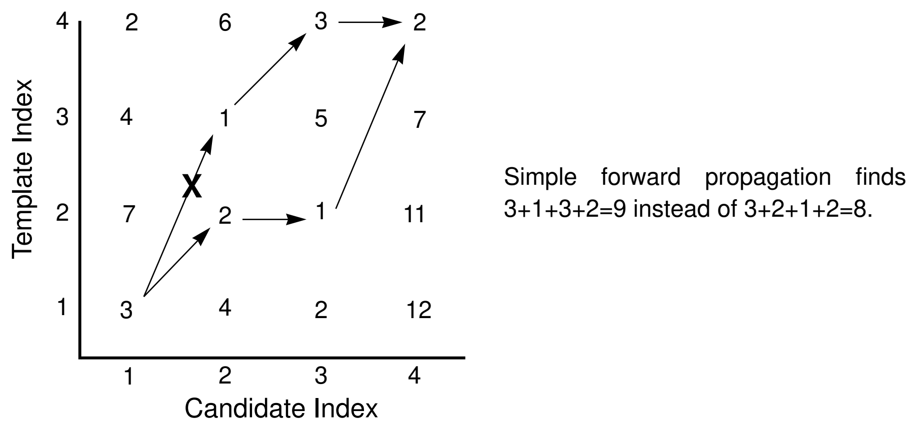


Figure 5.6 Single path, forward propagation misses the minimum path through the field.

### 5.3.3 Dynamic Programming Solution

DTW propagates all possible paths within the propagation constraint. Because the candidate time index must increment by one, it is only necessary to keep two arrays (of length equal to the number of elements in the template sequence) of accumulated distances: the "current" and "last" arrays. The minimum path to any point in the "current" array is the local distance at that point added to the lowest accumulated distance at the "last" array which can legally propagate to that point in the current array. Therefore, although the "current" and "last" arrays move forward in time, the path propagation choice is made looking backward from the "current" array to the "last" array.

In our recognition system, templates are discrete isolated words, and candidates are words within silence or background noise. There is a new path start point at each index point of the candidate (the x-axis in the path propagation fields). Similarly, at every index point in the candidate, a minimum path corresponding to a possible word endpoint propagates out the top of the field. If we keep track of the start point associated with each propagating path, when we choose the minimum path that propagates through the top of the field, we have the endpoint, start point, and total minimum accumulated distance for the associated word. The minimum total accumulated distance across templates specifies the word recognized.

### 5.3.4 Best Path Back-Tracking

To ensure the algorithms are working correctly, and to gain insight into how the system fails, we also keep track of the path propagation selection at each point in the path propagation field. After finding the minimum path that propagates through the top, we back-track through the field and assemble the dynamically time-warped template that best matches the candidate. This step has provided considerable insight into how these algorithms work and perhaps more importantly, how they break down.

# Chapter 6

# Incorporating the Model with the Speech Recognition System

## 6.1 Implementation of the Model

Figure 2.7 (in Chapter 2) shows an overview of the adaptive auditory model, implemented as a parallel filter bank, followed by independent automatic gain control (AGC) on each filter output.

### 6.1.1 Filter Responses

The behavior of the filter bank before the independent AGC is completely specified by the individual frequency responses. The responses chosen for this model are almost directly from Goldhor [1985]. In general, these responses reflect three basic premises derived from psychoacoustic masking data [Zwicker, 1974, 1978, and Houtgast 1977]:

1) Critical bands impose a limited frequency resolution.

2) High-frequency transitions (or filter skirts) are more abrupt than low-frequency transitions, on either a logarithmic or bark frequency scale.

3) Adjacent filters overlap significantly; not only are the skirts non-abrupt, but there are multiple filters per critical band.

The first two points are discussed in Chapter 2, the third point deserves more attention here. Typically, critical band-based analysis of speech consists of non-overlapping equivalent rectangular bands (ERBs) as in Patterson *et al.* [1994]. Each band is a critical band wide, and roughly 20 span the frequency range from 0-5kHz. However, there is no perceptual, nor physiological motivation for a small set of fixed-center frequency filters. That is, if we simply span the frequency range with non-overlapping ERBs, the system is incorrectly sensitive to the absolute locations of speech harmonics and resonances. A resonance at precisely the transition between two ERBs will pass through both bands, but a resonance slightly higher will pass primarily through one. When we examine spectral transitions in speech, this problem gets worse. Therefore, to reduce the model's sensitivity to absolute frequencies, we use multiple filters per critical band, each a critical band wide, and each overlapping adjacent bands significantly. This solution is consistent with a physiological interpretation of the mechanical filtering of the cochlea described in Chapter 2. Seneff [1990] and Lyon and Mead [1988] also implement filter structures with these general characteristics, including multiple center frequencies per critical band.

For each filter in our bank, the -3dB points are spaced by 1 critical band. There are roughly 4 filters, and therefore 4 center frequencies, per critical band (the next higher center frequency is the current center frequency plus 1/4 of the current critical bandwidth). Low-frequency skirts drop as 10 dB/Bark and high-frequency skirts drop as 25 dB/Bark [Goldhor 1985]. Analytic expressions for the bark scale and for the critical bandwidth (in Hz) as a function of linear frequency $F$ (in Hz) are from Zwicker and Terhardt [1980].

$$Bark = 13\,\mathrm{atan}\left(0.76\frac{F}{1000}\right) + 3.5\,\mathrm{atan}\left[\left(\frac{F}{7500}\right)^2\right]$$

$$CB = 25 + 75\left[1 + 1.4\left(\frac{F}{1000}\right)^2\right]^{0.69}$$

All filters are designed by taking the 4096-point IDFT of the desired frequency response, and then windowing the resulting impulse response with a 255-point raised-cosine window. The resulting frequency responses, evaluated on 800 points around half the unit circle, and plotted on a log frequency scale are shown in Figure 6.1. The filters are linear phase FIR.

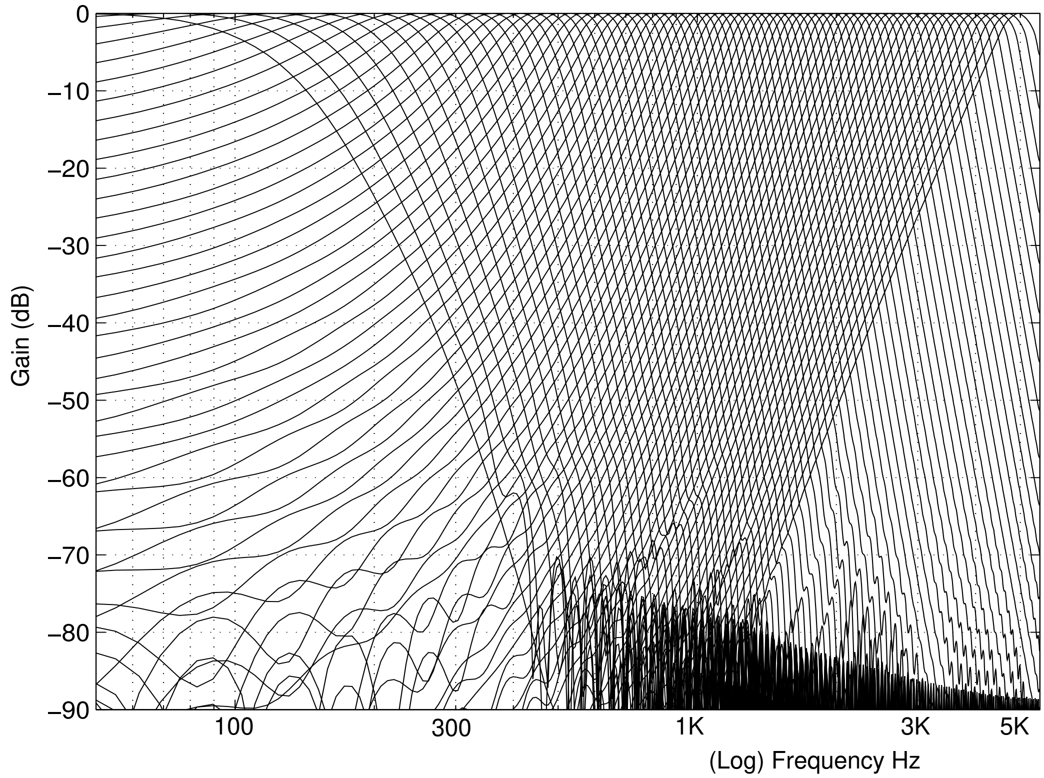


Figure 6.1 Frequency responses of the filter bank: 72 filters, 4 filters per critical band, -3dB bandwidth equal to one critical band, 10 dB/Bark low-frequency skirt, 25 dB/Bark high frequency skirt.

The model convolves each of the filter responses with the input waveform. After convolution, there are 72 data streams each at a 11025 samples/second rate. The model implements the convolution at the full data rate, without down-sampling. (Future versions of this model will exploit the time domain detail of these signals.) The convolution is implemented using the FFT and an “overlap and add” technique. The window length for the overlap and add FFTs are 2048 point, and only 2048-255-1 points from the input data are used in each overlap and add window to ensure no time domain aliasing.

Empirically, using an HP715 workstation, direct convolution of 2 seconds of input data (at 11025 samples/sec) with 72 255-point filter responses required just over 80 seconds. Using the overlap and add FFT technique, the same convolution requires 6.3, 6.1, and 7.0 seconds for 1024, 2048, and 4096-point FFT windows respectively. Therefore a 2048-point window empirically optimizes the trade-off between a large enough window to make sufficient progress through the data for each window, and a small enough window that the computation of the large-window FFTs do not dominate computation time. Unfortunately, the filter bank implementation alone is still greater than 3x real time.

### 6.1.2 Implementation of the AGC

Parameters for the AGC for each octave from 250 Hz to 4000 Hz are derived from the forward masking experiments in Chapter 3 and summarized in Table 4.1 of Chapter 4. Specifically, Table 4.1 describes the AGC I/O curves and adaptation time constants as a function of frequency. The parameters change gradually with frequency, therefore filters with center frequencies between two measured frequencies use a weighted average of the two adjacent parameter sets.

As described in Chapter 4, the AGC is implemented as first order difference, where the gain incrementally adjusts to place the output closer to the I/O curve target. That is, if the input is 30 dB, the gain is currently 0 dB, and the I/O curve target is 60 dB, at that sample the gain in dB would increase by a small $\alpha$ (~0.002) times the difference between 60 and 30 dB. Statically, the system always approaches the I/O curve target.

The AGC is implemented at the full data rate, on each sub-band. This provides smoothly varying adaptation, and more precise time domain detail. However, it requires a computational burden similar to the convolution. (Including time for the DTW after the signal processing, the total system runs at slightly greater than 10x real time.)

After implementing the AGC at the full data rate on each band, the model averages a sliding window for each filter/AGC output to generate an energy per frequency per time representation of the input signal, or a "perceptual spectrogram." The window length is 40 ms, the window increment is 11.6 ms, and the energy average is weighted by a raised-cosine. By reducing the time-domain detail to an energy per time value, the model down-samples the output of each filter/AGC pair by a factor of 128. The effective down-sampling provides a suitable feature vector rate for speech recognition.

## 6.2 Defining a Local Distance Metric

To this point, the model provides a perceptually-parameterized dynamic (or input-dependent) spectral estimation of the speech signal as a function of time. We must also define a perceptually-relevant method to compare the resulting spectral estimations. Although the Euclidean distance is mathematically convenient, the frequency difference of local spectral peaks is far more perceptually relevant [Klatt series 1979-1982]. Therefore, from the perceptually-parameterized spectral estimation, we derive a perceptually-relevant distance metric using cepstral manipulations.

### 6.2.1 Cepstral Liftering

Chapter 5 describes why typical speech recognition systems use cepstral representations (LP-based, or DFT/Mel-based) as the feature vector which forms the observation sequence. The cepstral representation is the DCT of the well-correlated log spectrum, and therefore approximates its KL-Transform. As an approximation of the KL-Transform, the cepstral representation has significant "energy compaction"-- only the lowest cepstral coefficients have significant variance, the coefficients themselves are largely uncorrelated, and the Euclidean distance between two cepstral vectors is consistent with the Euclidean distance between the corresponding spectral vectors. Therefore, cepstral representations provide efficient representations of speech, and reduce the computational burden of comparing spectral estimations.

Perhaps, more importantly, cepstral representations also provide the opportunity for simple but perceptually-relevant manipulations. Truncating a cepstral representation removes any "high-frequency" ripple from the corresponding log spectrum. With DFT-derived cepstral representations, this removes any remnants of voicing, or of the pitch. With LP-derived cepstral representations, cepstral truncation softens, or rounds, the spectral peaks whose absolute magnitude (and sharpness) may contribute significantly to the Euclidean distance, but may be more of a numerical artifact of LP analysis [Rabiner and Juang, 1993] and less of a perceptually-relevant distinction.

In addition to cepstral-truncation, recognition systems often weight the cepstral representation. The lowest cepstral coefficient corresponds to the total level of the spectral vector (DC offset of the log frequency vector), and the next few coefficients represent the spectral tilt and other slow changes across the log frequency vector. Because low order cepstral coefficients are assumed to represent characteristics of the channel (total level and average spectral tilt), and of the speaker (spectral tilt of the individual driving function), they are gradually de-emphasized with decreasing order. Similarly above some middle-order cepstral coefficient, coefficients begin to represent insignificant artifacts of LP-analysis, or of the voicing information present in DFT-based representations. Therefore, higher-order coefficients are also gradually de-emphasized with increasing order. The net effect of the de-emphasis is a type of filtering of the log spectrum in the DCT (IFFT) transform domain called cepstral liftering. As described in Chapter 5, a half-period raised-sin function that starts and ends rather abruptly at zero, and peaks smoothly over the mid-order (~5th to 10th) cepstral coefficients is often used.

Viewed in the log spectral domain, cepstral liftering is a type of "low-frequency" band pass filter with a sharp zero at DC. On a spectrogram, raised-sin cepstral liftering implies vertical "low-frequency" band pass filtering.

### 6.2.2 What remains after raised-sin cepstral liftering?

After de-emphasizing low and high-order cepstral coefficients, it seems reasonable to ask what remains. After cepstral liftering, to what spectrum do the resulting cepstral vectors correspond? Figure 6.2 shows our perceptual model's spectral estimation for a segment of the "ee" in "three" from a male speaker, in addition to the spectral estimation implied by cepstral truncation, and by raised-sin cepstral liftering (including 16 cepstral coefficients). Cepstral coefficients are the DCT of the perceptual spectral estimation (or the IFFT of the even spectral estimation), and the spectrum implied after cepstral manipulation is the FFT of the manipulated cepstral vector.

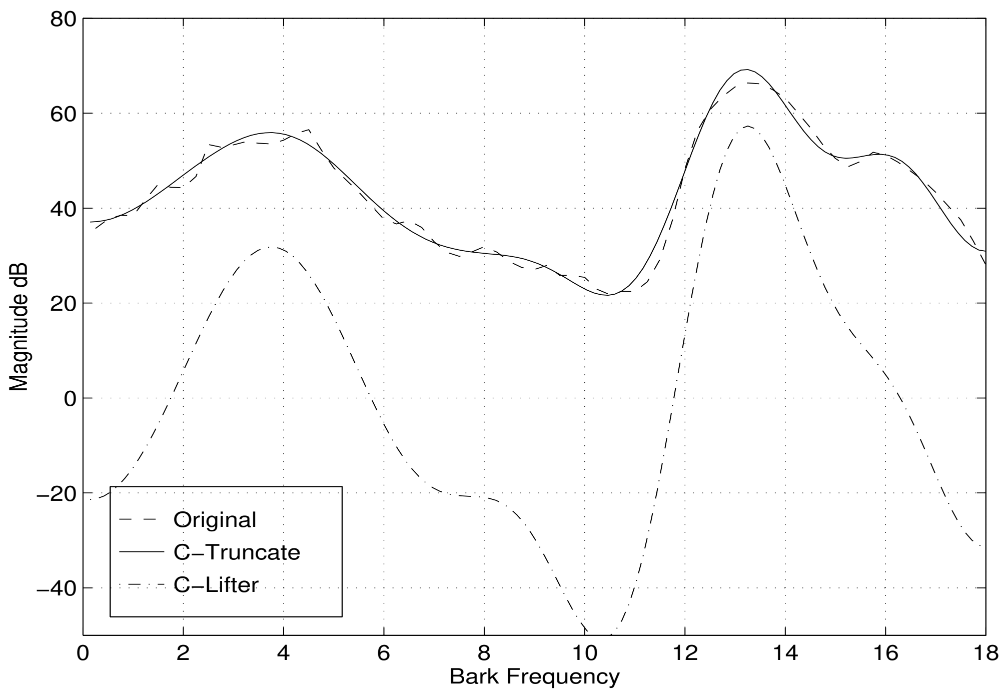


Figure 6.2 A "perceptual" spectral estimation of "ee" in "three," and the implied spectral estimations after cepstral truncation, and after raised-sin cepstral liftering.

The spectral estimation implied by cepstral truncation represents a smoothed version of the original spectral estimation. Raised-sin cepstral liftering, on the other hand, shifts the “DC level” across Bark frequency to zero, and dramatically emphasizes spectral peaks and valleys. Therefore, raised-sin cepstral liftering provides a technique to “pull-out” the local spectral peaks, significant for perceptual discrimination. Unfortunately, this technique alone also emphasizes local valleys, which have relatively less perceptual significance. Our solution to this problem is to clip the raised-sin-liftered spectral estimation at zero, when it drops below zero. Obviously, after zero-clipping in the spectral domain, transforming back to the cepstral domain provides an equivalent, but more efficient (better energy compaction) vector for spectral comparison.

### 6.2.3 Our Local Distance Metric Implementation

The dynamic filter bank in our system provides a perceptually-derived spectral estimation. This provides a first-order model of the front-end signal processing of the human auditory system, with limited frequency resolution, and a natural emphasis of onsets and transitions. However, when discriminating speech sounds, humans are sensitive to the frequency location of local spectral peaks. Therefore after this front-end model, we include processing to emulate the sensitivity to the frequency location of spectral peaks. Presumably, higher-levels of the auditory system provide this functionality.

An overview of the algorithm we use to define the local distance metric is described in Figure 6.3. Raised-sin cepstral liftering followed by spectral zero-clipping provides the location of local spectral peaks. Our local distance metric is then the Euclidean distance between original spectral vectors, each weighted by the local peak position estimations from the cepstral processing.

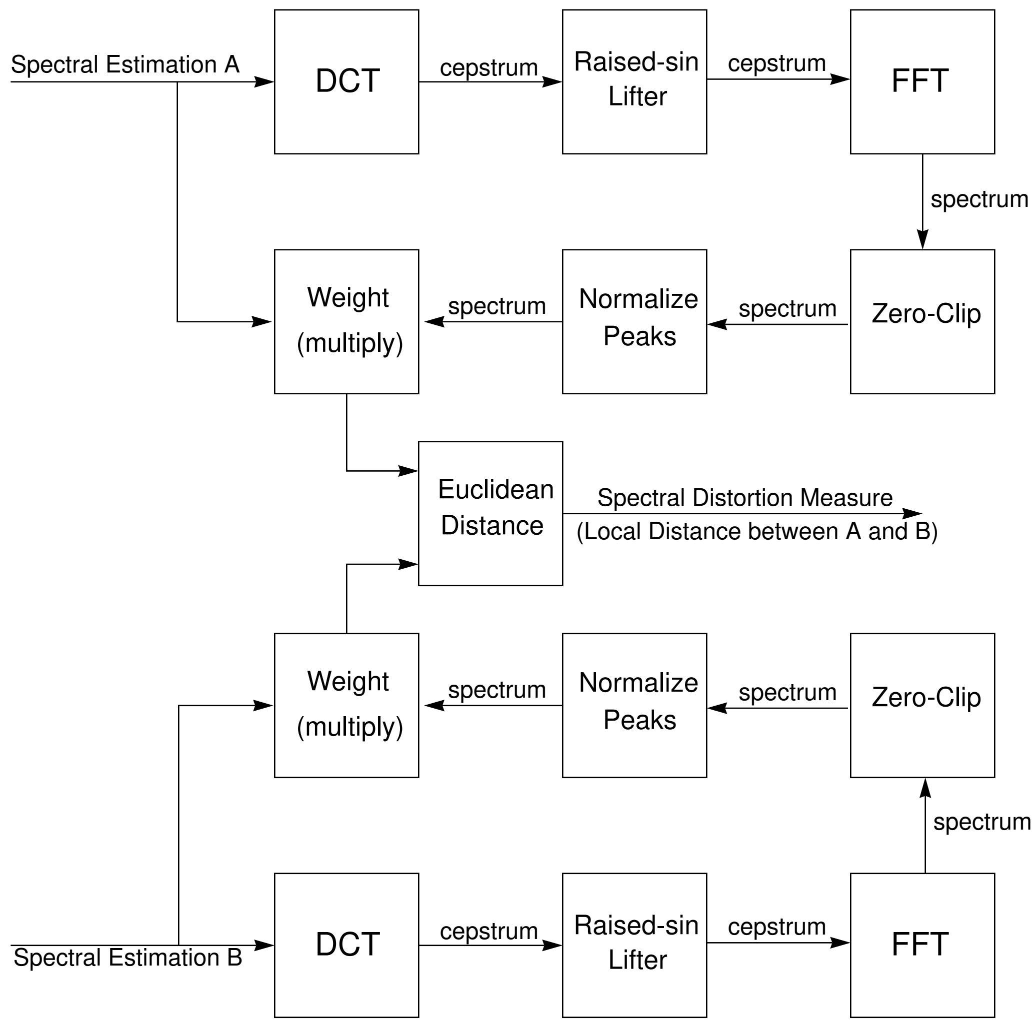


Figure 6.3 Processing to derive the local distance metric. Raised-sin cepstral liftering followed by spectral zero-clipping isolates local peaks, and provides a suitable weighting function for estimating perceptual spectral distances.

Finally, the Euclidean distance of the weighted spectral estimations is adjusted slightly to further emphasize the frequency position of the spectral peaks, and not their absolute levels. If the spectral estimation for both vectors at a particular frequency is greater than a high internal threshold, the contribution of this frequency to the local distance is decreased proportional to the amount above the threshold. Therefore, onsets which typically cause high-magnitude responses, are measured more by the frequency position of the onset, and less by the absolute magnitude of the onset response (which, of course, changes with the level of the preceding ambient noise).

There are obvious variations to our proposed distance metric technique. As a final step, the spectral estimation could be transformed back to the cepstral domain to reduce the dimension through improved energy compaction. In such an implementation, this processing alone could also be applied to any cepstrally-based speech recognition system. Given a cepstral vector (which implies a spectral estimation) this technique returns a new cepstral vector which reflects the perceptually-significant local spectral peaks. This is a novel algorithm to improve the perceptual significance of cepstral representations of speech. Future experimentation should evaluate the potential of this technique to other speech processing (recognition/coding) systems.

## 6.3 Examples of the Model's Representations

The following figures show examples of the model's representation of speech. Each figure includes the time domain waveform, the DFT-based spectrogram, the “perceptual spectrogram” from our adapting perceptual model’s spectral estimations, and the “concentrated perceptual spectrogram” obtained after the pre-processing for the local distance metric. All gray-scale spectrograms are normalized so that the peak level is set to the darkest value, and a consistent minimum threshold is set to the lightest value. The spectrograms that include additive background noise use Gaussian noise shaped to match the long-term average speech spectrum, as described in Section 6.4.

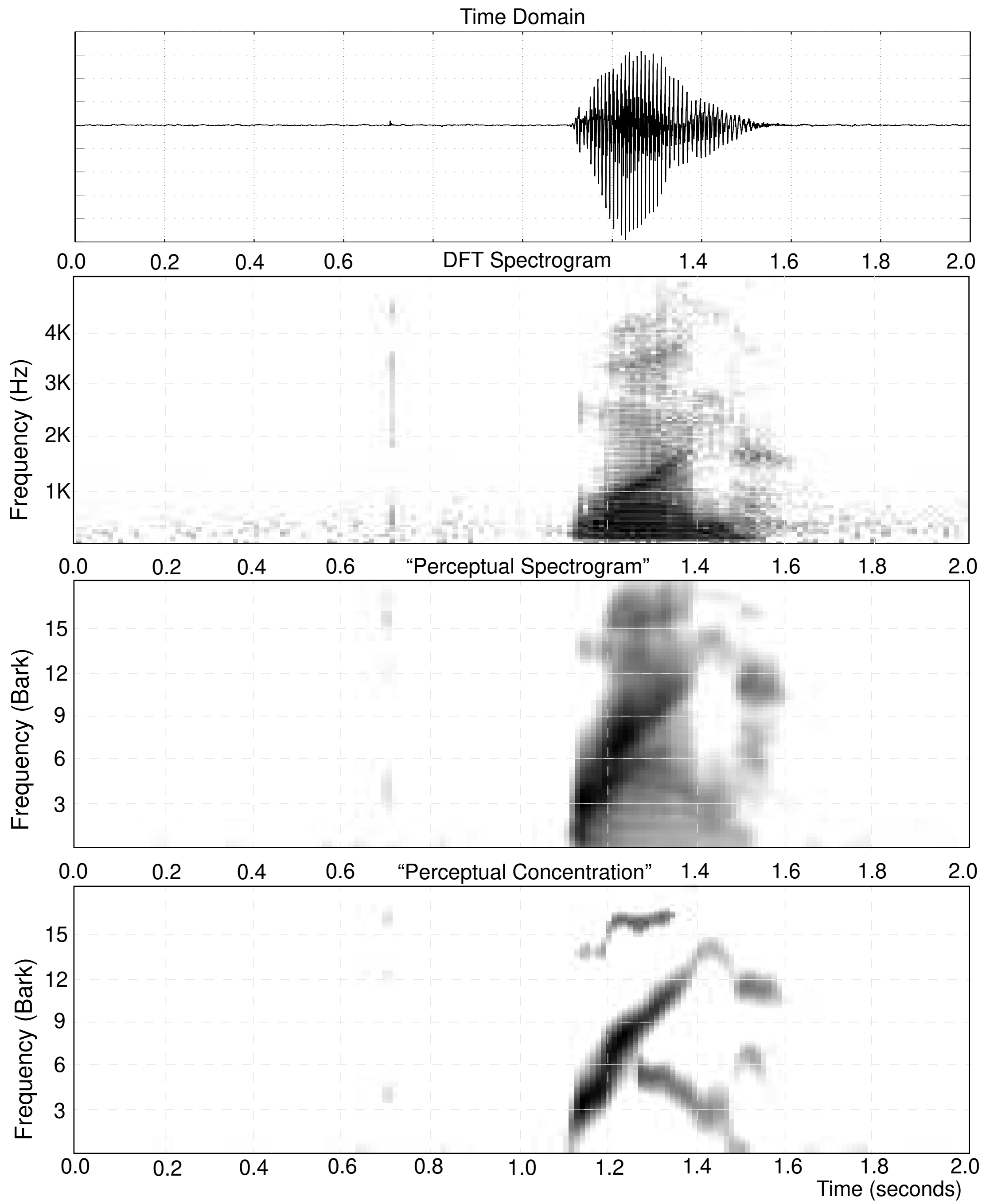


Figure 6.4  The perceptual model's representations of the word "one." The "perceptual spectrogram" is the output of the adapting filter bank, and the "perceptual concentration" shows the spectral estimation after the local distance pre-processing.

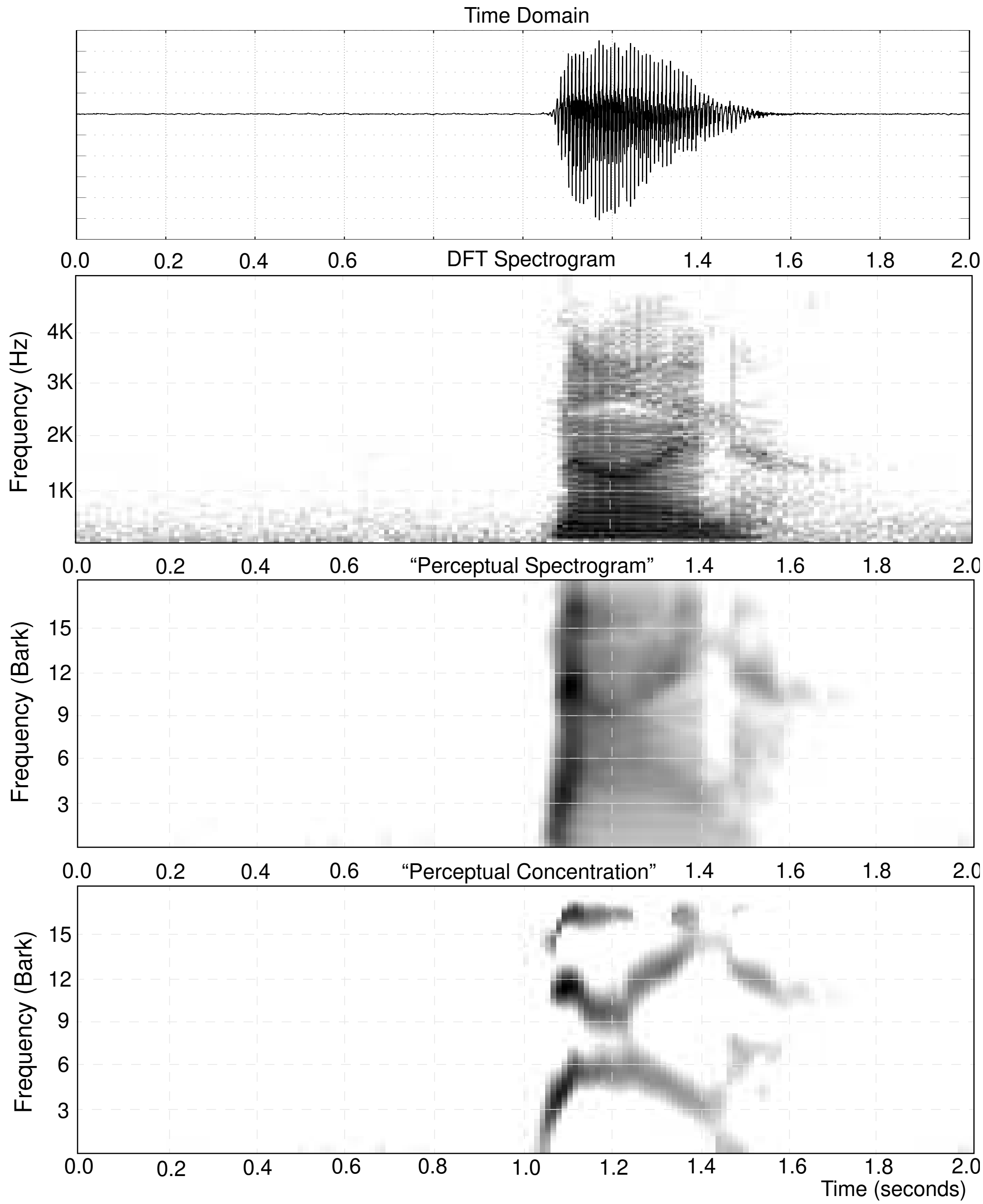


Figure 6.5 The model's representations of the word "nine." Emphasis of onsets highlights perceptually-relevant distinctions from the previous word "one."

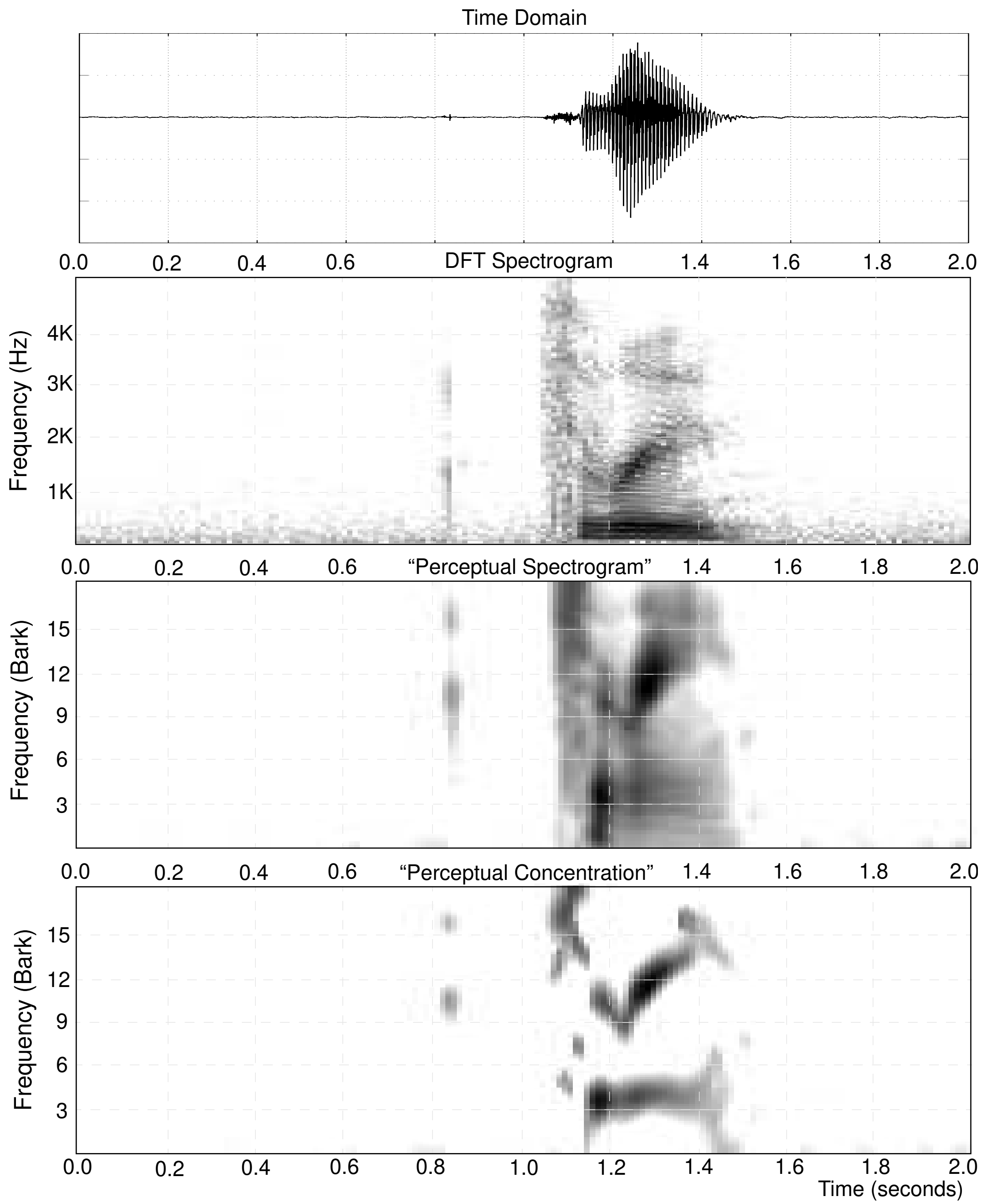


Figure 6.6 The perceptual model's representations of "three." The dynamic model emphasizes onsets and transitions, and the local distance pre-processing emphasizes local spectral peaks.

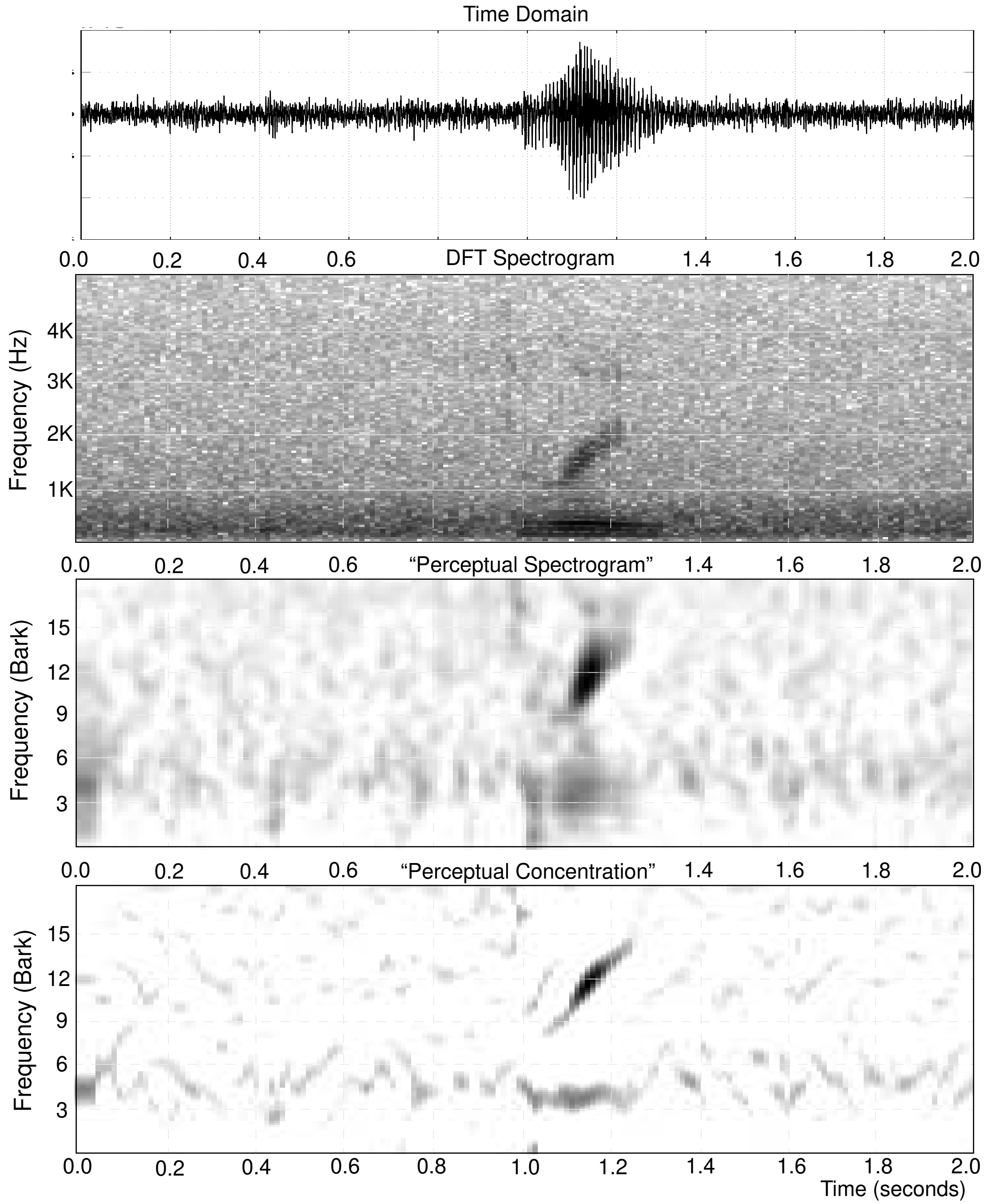


Figure 6.7 The model's representation of "three" with 15 dB Peak SNR additive average speech spectrum noise. Emphasizing onsets, transitions, and then indentifying local spectral peaks provides a more robust speech representation.

## 6.4 Performance Analysis

Figure 6.8 compares the degradation in background noise of four speech recognition systems. The first three use the DTW recognizer described in Chapter 5; first with truncated LP-cepstral coefficients, and a local distance which does not include the lowest (DC-level) cepstral coefficient; second with LP-cepstral coefficients, and a local distance weighted by a raised-sin cepstral lifter; and third with the dynamic model and the perceptually-motivated local distance pre-processing. The fourth is a human listener.

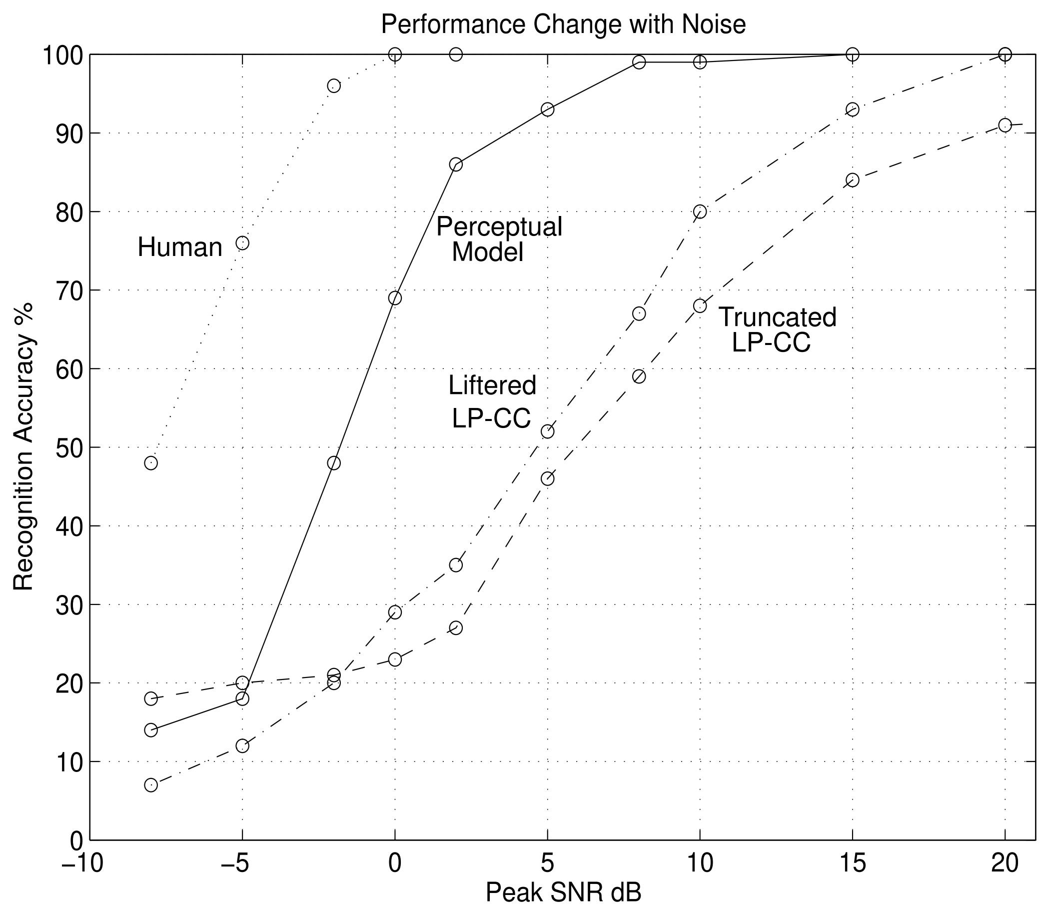


Figure 6.8 Recognition performance in noise with the perceptual model.

The vocabulary for these test were the ten digits, each spoken 10 times, for a total of 100 tokens. An eleventh set provided the templates. All tokens are from the same male speaker. Recordings were made in a double-walled sound booth through a 16-bit A/D at a sampling rate of 11025 samples/second. Before the noise is added, the recordings have a peak SNR, defined as the maximum variance over a 15 ms window in the entire 2 second recording divided by the minimum variance over a 15 ms window, of greater than 50 dB.

The templates for all tests are clean data (these systems are not ‘trained’ in noise). We add progressively higher amounts of background noise to the same set of 100 test tokens to measure the curves in Figure 6.8. The background noise is Gaussian noise with the spectral shape of long term average speech, as defined in Byrne and Dillon [1986]. Figure 6.9 shows the long-term average spectrum of the additive noise used.

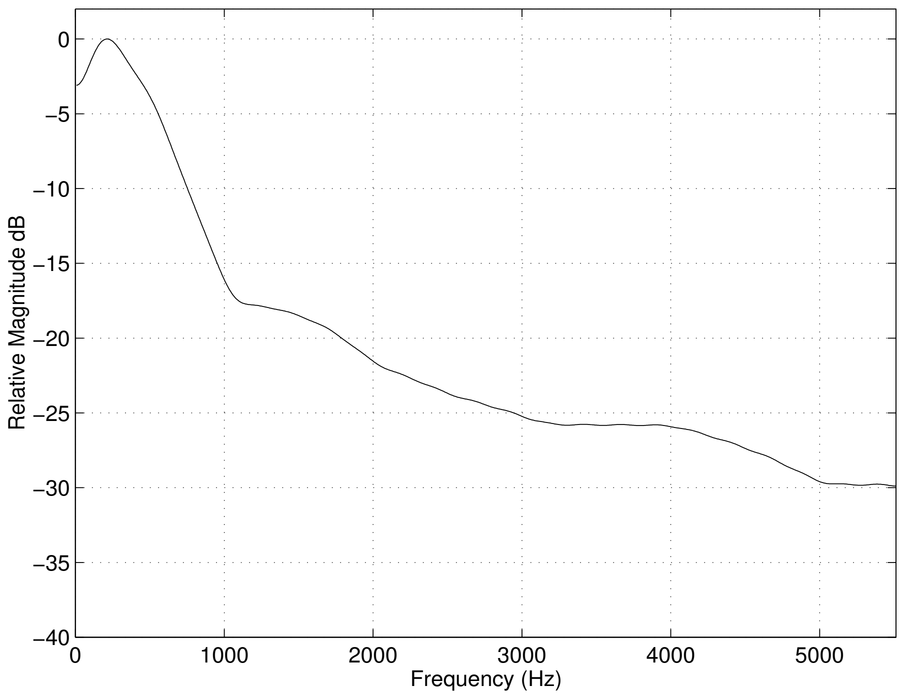


Figure 6.9 The average spectrum of the additive noise. Frequency response from Long-Term Average Speech Spectrum in Byrne and Dillon [1986].

The peak signal variance is the maximum variance measured in a 15 ms window sliding one point at a time through the clean token. The variance of the noise set to achieve the desired "Peak SNR" listed in Figure 6.8. Often the SNR for speech is calculated as the average energy of the "speech" in the input divided by energy of the stationary noise. However this approach has three drawbacks. First, it requires a segmentation of the input into speech and non-speech which is not a well-defined task, and therefore makes reproducing or comparing experiments difficult. Second, variations in the kind of speech (ratio of strong voicing to weak fricatives, and the amount of stop-preceding silences as in "six") have substantial impact on total energy averaged across non-stationary speech. Third, by including extremely low-energy pieces of the speech signal in the energy estimation, average SNR reduces the estimate of the signal energy, decreases the SNR measurement, and therefore artificially inflates system robustness results. Pragmatically, as speech falls below the noise level, the quiet segments drop first, and eventually, the listener is left with just the peaks. Nonetheless, to allow for comparisons with results that quote average SNR values, Table 6.1 lists the difference between average and peak energy for ten digit utterances.

Table 6.1 Difference Between Peak and Average Energy

| Utterance | Peak/ Average |
|---|---|
| one | 5.4 dB |
| two | 6.0 dB |
| three | 6.3 dB |
| four | 6.4 dB |
| five | 8.4 dB |
| six | 9.9 dB |
| seven | 7.9 dB |
| eight | 7.1 dB |
| nine | 5.2 dB |
| zero | 5.2 dB |

Average SNR values are typically about 6 dB lower than Peak SNR, so system performance at “5 dB Peak SNR” is roughly equivalent to system performance at “-1 dB Average SNR.”

Figure 6.10 shows the time-waveform, DFT-based spectrogram, and model representations for an utterance of “three” at 5 dB peak SNR. At this amount of noise, the perceptually-based representation still leads to recognition performance above 90%.

## 6.5 A Brief Summary

The filter bank in our model imposes frequency selectivity and a frequency scale consistent with perceptual experiments. After the filter bank, carefully-parameterized adaptation matches the perceptual forward-masking results presented in Chapter 3. In addition, we define an original processing technique (similar to raised-sin cepstral-liftering), to emphasize the perceptually-relevant local peaks in the dynamic spectral representation. The total model significantly improves the robustness of a simple speech recognition system to additive noise.

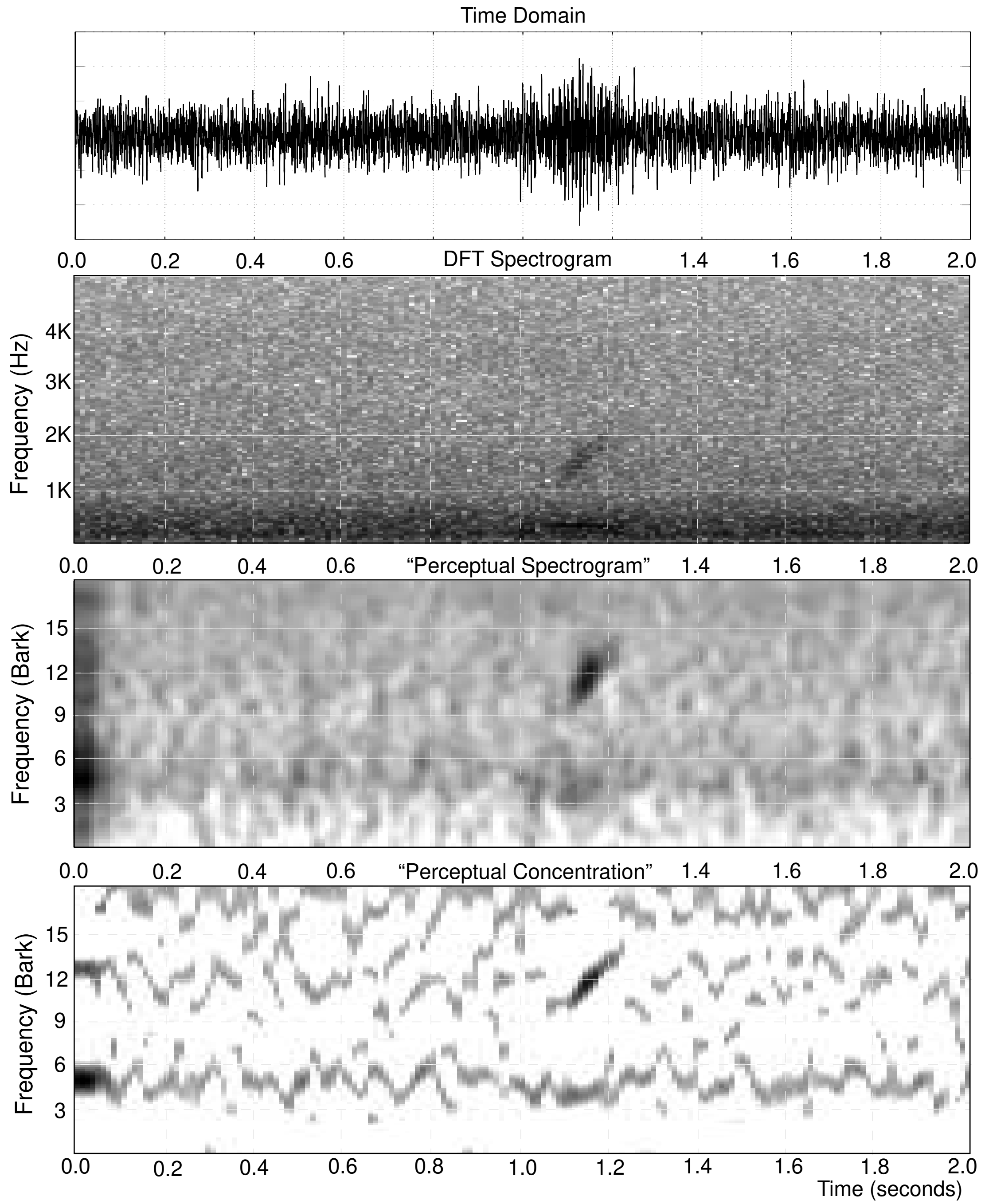


Figure 6.10 The model's representations of the utterance "three" at a peak SNR of 5 dB. At this noise level, the perceptual representation leads to over 90% accuracy, while the LPC-cepstral representation leads to just over 50%.

# Chapter 7

# Discussion and Future Direction

## 7.1 Summary

This thesis derives a first-order model of dynamic auditory perception from original pyschoacoustic forward-masking experiments. The model consists of a filter bank followed by logarithmic AGC, and includes a novel cepstral processing technique to isolate local spectral peaks. Further, the model is shown to improve the robustness of a simple speech recognition system to additive noise, by 5 to 10 dB.

Chapter 2 summarizes physiological and perceptual evidence supporting a dynamic model of auditory perception, and also provides a qualitative overview of our first-order dynamic auditory model. Chapter 3 details the perceptual forward masking experiments and results which imply parameters for the dynamic model. Chapter 4 translates the forward-masking results into model parameters. Chapter 5 provides an overview of the DTW-based speech recognition system used to evaluate the model, and discusses common spectral representations used for speech recognition. Finally, Chapter 6 explains the implementation of the dynamic auditory model with the recognition system. It includes the specifications of the filter bank, and describes a novel cepstral processing scheme (similar to raised-sin cepstral liftering) which isolates local spectral peaks for the spectral distance measure. Chapter 6 also summarizes recognition accuracy across a wide range of noise levels, comparing the proposed dynamic model to more common approaches.

## 7.2 Other Applications of the Model

As a well-quantified model of dynamic auditory perception, this work may also have applications to hearing-aid design, speech enhancement, and speech coding. The model quantitatively predicts the perceptual relevance of the acoustic cues of (non-stationary) speech. Other applications may also benefit by emphasizing the quantified, perceptually-relevant, acoustic cues.

### 7.2.1 Hearing Aid Design and Sound Enhancement

Sensorineural hearing loss is characterized by increased thresholds, reduced dynamic range, and decreased frequency selectivity [Moore, 1989]. The model's carefully-parameterized adaptation could compensate increased thresholds and reduced-dynamic range. Also, the cepstral processing described in Chapter 6, which isolates local spectral peaks, may provide a reasonable technique to compensate reduced frequency selectivity.

### 7.2.2 Speech/Audio Coding

This model would apply well to two aspects of speech, or more generally audio, coding. First, as a quantified-prediction of the perceptual-salience of dynamic acoustic cues, this model suggests which parts of speech should be most accurately encoded, or perhaps more importantly, which parts could be de-emphasized or discarded. This approach would augment current methods which typically view speech as a sequence of (un-related) static segments and exploit predominant static masking effects [Shen, 1994]. Second, regardless of the particular coding scheme used, this model provides a quantified perceptual model to evaluate the ‘dynamic spectral distortion’ introduced by the coding algorithm. If spectral changes in the original speech are extremely perceptually-salient, then the coding algorithm should be careful not to alter existing spectral changes, and also not to introduce excessive dynamic spectral artifacts [Knagenhejlm and Kleijn, 1995]. This model provides a technique to quantify the tolerability of these distortions. Current frame-based speech coding techniques are likely to have this problem, and the author knows of no well-quantified algorithm to measure dynamic spectral distortions.

## 7.3 Model Improvements

This thesis presents a first-order auditory model incorporating adaptation with more classically-analyzed energy per critical band. There are several, perhaps glaring, phenomena suggesting model improvements.

### 7.3.1 Separation of Adaptation

There is no reason to expect that auditory adaptation can be precisely modeled with a single adaptive block per frequency band. Physiologically, we expect many levels of adaptation throughout the auditory system. Outer hair cells may provide near instantaneous feedback to waves travelling along the basilar membrane [Ashmore, 1987], they may also respond to neural cues from the brain, and the brain may adapt its perceptive neural response and feedback cues. These suggest at least three levels of adaptation each with its characteristic time constant and I/O curve. Such a separation of adaptation provides the opportunity to fine-tune the model to match perceptual results more precisely.

### 7.3.2 Inner-Hair Cell Modeling and Higher-Level Processing

Explicit inner-hair cell models including phase-synchronization, realistic mechanical transduction to neural spikes, latencies, refractory-periods, and lateral-inhibition effects may prove essential for improved auditory modeling leading to robust speech recognition. Perception of sound is more than an identification of the amount of energy in each critical band as a function of time. (Pure tones "sound" different than narrow band noise, but a critical band model-- and the corresponding cepstral vector-- can not make this distinction.)

Delgutte and Kiang [1984] report detailed measurements of inner-hair cell responses to speech-like stimuli. Instead of a simple random firing that increases in rate with the amount of stimulation at a particular frequency (consistent with a Short-Time Fourier Analysis model of perception), inner-hair cells synchronize their firing to dominant time-domain periodicities of the input, even when those periodicities are more than an octave from the specific hair cell's center, or best, frequency. The resulting cross-band redundant synchronization provides important insight into the mechanisms which are most probably responsible for the noise-robustness of the human auditory system.

Seneff [1990] and Ghitza [1991] derive auditory models used for speech recognition which try to exploit the time-domain detail of inner-hair cell responses. Unfortunately, without a quantified model of the next level of auditory processing in the brain, it is not at all clear *how* to process the huge amount of resulting data [Patterson *et al.*, 1994] perceptually, physiologically, or even optimally. Instead, models must revert to over-simplifying ad hoc solutions with only modest success. This problem defines an opportunity and an interesting challenge for current auditory modeling. A solution should prove invaluable for robust speech recognition. (In the model presented here, a solution would replace the cepstral pre-processing to isolate "perceptually-relevant" local spectral peaks.)

## 7.4 The Speech Recognition System

This thesis evaluates the proposed model with a DTW-based speech recognition system. Most current systems [Lee *et al.*, 1991, Rabiner and Juang, 1993] use stochastic Hidden Markov Models. This approach provides a rich mathematical framework to "train" a system based on large amount of data, and then to find the most probable sequence of underlying states given the current speech or observation sequence. As a next step, we should evaluate our proposed model with an HMM-based recognition system.

Unfortunately, the basic HMM framework imposes the assumption that speech is a sequence of stationary, uncorrelated, segments. Our perceptual model, on the other hand, responds to speech more as a sequence of onsets and spectral transitions. Therefore, to take full advantage of a dynamic perceptual model, modifications to the basic HMM structure may be necessary [Morgan *et al.*, 1995].

## 7.5 An Underlying Theme

This work reflects the judgement that automatic speech recognition, and hopefully our understanding of speech perception, will improve by incorporating well-quantified, non-linear mechanisms reflecting the next most obvious physiological and perceptual phenomena (after critical bandwidth analysis), and the choice to pursue doing it.

# Bibliography


Ashmore, J. (1987). “A fast motile response in guinea-pig outer hair cells: the cellular basis of the cochlear amplifier,” J. Physiol. **388**, 323-347.

von Bekesy, G. (1953). “Description of Some Mechanical Properties of the Organ of Corti,” J. Acoust. Soc. Am. **25**, 770-785.

von Bekesy, G. (1960). *Experiments in Hearing*. McGraw-Hill, New York.

Byrne, D., Dillon, H. (1986). “The National Acoustic Laboratories’ (NAL) New Procedure for Selecting the Gain and Frequency Response of a Hearing Aid,” Ear and Hearing 7, 257-265.

Delgutte, B. and Kiang, N. Y. S. (1984). “Speech coding in the auditory nerve: I. Vowel-like sounds,” J. Acoust. Soc. of Am. 75, 866-878.

Dillon, H. and Walker, G. (1982). *Compression in Hearing Aids: An Analysis, a Review, and Some Recommendations*. NAL Report No. 90. Australian Government Publishing Service, Canberra.

Duifhuis, H. (1973). “Consequences of peripheral frequency selectivity for nonsimultaneous masking,” J. Acoust. Soc. Am. **54**, 1471-1488.

Evans, E. F., and Harrison, R. V. (1976). “Correlation between outer hair cell damage and deterioration of cochlear nerve tuning properties in the guinea pig," J. Physiol. **256**, 43-44P.

Fant, G., (1960). *Acoustic Theory of Speech Production.* Mouton, The Hague.

Fletcher, H. (1940). "Auditory Patterns," Rev. Mod. Physics **12**, 47-65.

Furui, S. (1986). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**, 1016-1025.

Ghitza, O. (1991) "Auditory Nerve Representations as a Basis for Speech Processing," Advances in Speech Processing (Eds. S. Furui, M. Sondhi), Marcel Dekker, NY, 453-485.

Goldhor, R. S. (1985). "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features," RLE Technical Report No. 505, MIT, Cambridge MA.

Hermansky, H., Morgan, N., Aruna, B., and Kohn, P. (1992). "RASTA-PLP speech analysis technique," Proceedings, 1992 IEEE ICASSP, San Fransisco, 121-124.

Houtgast, T. (1977). "Auditory-filter characteristics derived from direct-masking data and pulsation-threshold data with a rippled-noise masker," J. Acoust. Soc. Am. **62**, 409-415.

Jesteadt, W., Bacon, S., and Lehman, J. (1982). "Forward Masking as a function of frequency, masker level, and signal delay," J. Acoust. Soc. Am. **71**, 950-962.

Johnstone, B., Patuzzi, R., and Yates, G. K. (1986). "Basilar membrane measurements and the travelling wave," Hearing Res. **22**, 147-153.

Kates, J. (1991). "An Adaptive Digital Cochlear Model," Proceedings, 1991 IEEE ICASSP, Toronto, 3621-3624.

Kemp, D. (1978). "Stimulated acoustic emissions from within the human auditory system," J. Acost. Soc. Am. **64**, 1386-1391.

Klatt, D. (1979). "Perceptual comparisons among a set of vowels similar to [ae]: Some differences between psychophysical distance and phonetic distance," J. Acoust. Soc. Am. **66**, Suppl. 1, S86.

Klatt, D. and McManus, T. (1980). "Perceived phonetic distance among a set of synthetic whispered vowels and fricative consonants," J. Acoust. Soc. Am. **68**, Suppl. 1, S49.

Klatt, D. (1981). "Prediction of perceived phonetic distance from short-term spectra--a first step," J. Acoust. Soc. Am. **70**, Suppl. 1, S59.

Klatt, D. (1982). "Prediction of perceived phonetic distance from critical-band spectra: a first step," Proceedings, 1982 IEEE ICASSP, Paris, 1278-1281.

Klatt, D. (1986). "The Problem of Variability In Speech Recognition and In Models of Speech Perception," *Invariance and Variability in Speech Processes*, (Eds. Perkell, J., Klatt, D.) Lawrence Erlbaum Associates, New Jersey, 300-319.

Knagenhjelm, H. P., Kleijn, W. B. (1995). "Spectral Dynamics is More Important than Spectral Distortion," Proceedings, 1995 IEEE ICASSP, Detroit, 732-735.

Lee, K. F., Hon, H. W., and Huang, X. (1991). "Speech recognition using Hidden Markov Models: a CMU perspective," Speech Communication, **9**, 497-508.

Levitt, H. (1971). "Transformed Up-Down Methods in Psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

Levitt, H. (1992). "Adaptive Procedures for Hearing Aid Prescription and Other Audiologic Applications," J. Am. Acad. Audiol. **3**, 119-131.

Liberman, M. C. (1978). "Auditory-nerve responses from cats raised in a low-noise chamber," J. Acoust. Soc. Am. **63**, 442-455.

Lyon, R. F. (1982). "A Computational Model of Filtering, Detection, and Compression in the Cochlea," Proceedings, 1982 IEEE ICASSP, Paris, 1282-1285.

Lyon, R. F., and Mead, C. (1988). "An Analog Electronic Cochlea," IEEE Trans. on Acoust., Speech, and Sig. Proc. **36**, 1119-1133.

Moore, B. C. J. (1978). "Psychophysical tuning curves measured in simultaneous and forward masking," J. Acoust. Soc. Am. **63**, 524-532.

Moore, B. C. J., and Glasberg, B. R. (1983). "Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise," J. Acoust. Soc. Am. **73**, 1249-1259.

Moore, B. C. J., Glasberg, B. R., and Roberts, B. (1984). "Refining the measurement of psychophysical tuning curves," J. Acoust. Soc. Am. **76**, 1057-1066.

Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing*. Third edition, Academic Press, London.

Morgan, N., Bourlard, H., Greenberg, S., Hermansky, H., and Wu, S. L.(1995). "Stochastic Perceptual Models of Speech" Proceedings, 1995 IEEE ICASSP, Detroit, 397-400.

Neely, S. and Kim, D. (1986). "A model for active elements in cochlear biomechanics," J. Acoust. Soc. Am. **79**, 1472-1480.

Patterson, R., Anderson, T., Allerhand, M. (1994). "The Auditory Image Model as a Preprocessor for Spoken Language," Proceedings Acoust. Soc. of Japan ICSLP, 1395-1398.

Pickles, J. (1988). *An Introduction to the Physiology of Hearing.* Second edition, Academic Press, London.

Plomb, R. (1964). "Rate of Decay of Auditory Sensation," J. Acoust. Soc. Am. **36**, 277-282.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Second Edition, Cambridge University Press, Cambridge.

Rabiner, L., and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey.

Shen, A. (1994). "Perceptually-Based Subband Coding of Speech Signals," Master's Thesis, Department of Electrical Engineering, UCLA.

Sellick, P., Patuzzi, R., Johnstone, B. (1982). "Measurement of basilar membrane motion in the guinea pig using the Moessbauer technique," J. Acoust. Soc. Am. **72**, 131-141.

Seneff, S. (1990). “A joint synchrony/mean-rate model of auditory processing,” *Readings in Speech Recognition*, (Eds. A. Waibel, K. Lee), Morgan Kaufman Publishers, San Mateo, CA, 101-111.

Viergever, M, and Diependaal, R. (1986). “Quantitative validation of cochlear models using the Liouville-Green approximation,” Hearing Res. **21**, 1-15.

Wilson, J. P. (1980). “Evidence for a cochlear origin for acoustic re-emissions, threshold fine structure and tonal tinnitus,” Hearing Res. **2**, 233-252.

Zwicker, E., Flottorp, G., and Stevens, S. (1957). “Critical Band Width in Loudness Summation,” J. Acoust. Soc. Am. 29, 548-557.

Zwicker, E. (1974). “On a psychoacoustical equivalent of tuning curves,” *Facts and Models in Hearing* (Eds. Zwicker, E., Terhardt, E.), Springer, Berlin, 132-141.

Zwicker, E. and Schorn, K. (1978). “Psychoacoustical tuning curves in audiology,” Audiology **17**, 120-140.

Zwicker, E., Terhardt, E. (1980). “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” J. Acoust. Soc. Am. **68**, 1523-1525.

Zwislocki, J., Pirodda, E., and Rubin, H. (1959). “On Some Poststimulatory Effects at the Threshold of Audibility,” J. Acoust. Soc. Am. **31**, 9-14.