UNIVERSITY OF CALIFORNIA

Los Angeles

Modeling Auditory Perception for Robust Speech Recognition

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Brian P. Strope

1998

The dissertation of Brian P. Strope is approved.

Peter Narins

John Villasenor

Alan Willson

Abeer Alwan, Committee Chair

University of California, Los Angeles

1998

to Darren and Kelly

Table of Contents

1. Introduction 1

1.1 From Better Models to Robust Applications 2

1.2 Speech Recognition Overview 4

1.2.1 Front End Signal Processing 7

1.2.2 Pattern Comparison using Stochastic Models 9

1.3 Auditory Modeling Overview 18

1.4 Dissertation Overview 27

2. Adaptation and Local Peak Isolation 29

2.1 Introduction 29

2.2 Forward Masking 33

2.2.1 Experiments 34

2.2.2 Results 38

2.2.3 Modeling Implications 39

2.3 From Experimental Results to Model Parameters 40

2.3.1 AGC: I/O Curves, Attack and Release Times 41

2.3.2 Derivation of Model Parameters 43

2.3.3 Predicting Other Data 51

2.3.4 Other Models Predicting Forward Masking 54

2.4 Peak Isolation 56

2.5 Conclusion 59

3. Parameterizing the Position and Motion of Local Spectral Peaks 61

3.1 Background and Motivation 61

3.2 The Algorithm 63

3.2.1 Filtering and Adaptation 65

3.2.2 Peak Isolation 65

3.2.3 Threading Peaks 67

3.2.4 Choosing Three Peaks 67

3.3 Discussion 69

3.4 Conclusions 70

4. Modeling the Perception of Amplitude Modulation 71

4.1 Motivation 72

4.1.1 Pitch Perception 73

4.1.2 Perception of Amplitude Modulation 75

4.2 Strident Fricative Case Study: [s] and [z] 76

4.2.1 Perceptual Measurements 79

4.3 3 AM-Detection Mechanisms 80

4.3.1 Envelope Statistic 81

4.3.2 Modulation Filtering 82

4.3.3 Correlational Analysis 85

4.4 Comparing Predictions 89

4.5 Modeling Implications 93

4.6 Conclusions 95

5. Recognition Evaluations 96

5.1 Recognition Task 96

5.2 Stochastic Modeling Structure 98

5.3 Baseline Signal Processing 100

5.4 Implementation of the Model Signal Processing 101

5.4.1 Adaptation 101

5.4.2 Peak Isolation 101

5.4.3 Peak Position and Motion 102

5.4.4 Temporal Processing of Voicing Information 102

5.5 Other Techniques Targeting Noise 104

5.6 Evaluations 109

5.7 Interpreting the Results 112

6. Summary and Extensions 116

6.1 Threads 116

6.2 Correlation Inconsistencies 120

6.3 Looking Forward 123

7. Bibliography 126

List of Figures

Simplified overview of maximum likelihood speech recognition. 5

Automatic speech recognition overview. 7

Front end signal processing: windowing, spectral estimation, logarithmic compression, and decorrelation. 8

Mel-frequency spectrogram at 10-dB SNR. 9

A schematic representation of a hidden Markov model for ‘six.’ 11

An example of using a Viterbi search to solve the alignment problem........................ 13

Mechanical frequency selectivity of the basilar membrane. 20

Classifying auditory frequency ranges for speech... 23

Speech production overview. 25

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.. 36

Average forward masking data (circles), and std. dev. (error bars), together with the model fit (lines) as a function of masker level across 5 octaves, with probe delays of 15, 30, 60, and 120 ms as a parameter.. 38

Average forward masking data at 1kHz: (a) as a function of the log delay with masker level as a parameter; and (b) as the dynamic range below masker as a function of the masker level with probe delay as a parameter. The dotted line reflects the probe threshold in quiet. 40

(A) A prototypical I/O curve for a single channel in the dynamic model; and, schematic output trajectories corresponding to a level change at three different rates for (B) decreasing inputs, and (C) increasing inputs. 42

Geometry to derive recovery parameters from forward masking thresholds. 44

Geometry to derive attack (downward adaptation) parameters from forward masking thresholds as a function of masker duration. 46

The model’s prediction of the decay of forward masking as a function of masker level at 1 kHz. 49

Adaptation to, and recovery after, a pulse: (A) The response to the second pulse is diminished; and (B) Impulses, corresponding to onsets, are initially masked (similar to figures in [Goldhor 1985]). 50

Using the model to predict other forward masking data: (A) wide-band masker and probe [Plomp 1964]; (B) wide-band masker, sinusoidal probe at 1kHz [Moore and Glasberg 1983]; (C) sinusoidal masker and probe at 1kHz [Jesteadt et al. 1982]. (D) The equation provided in [Jesteadt et al. 1982] predicting the present data. 52

Peak isolation processing: log spectrum of the vowel [i] after (A) cepstral truncation; (B) raised-sine cepstral liftering; and (C) half-wave rectification. 58

Spectrogram representations for three processing strategies: left column is at 30 dB SNR, right is 5 dB SNR; top spectrograms use MFCC, middle includes adaptation, and bottom also includes peak isolation. 60

Overview of processing stages. 64

Peak positions and motion: A) Initial peaks identified after cepstral liftering; B) Neighboring peaks grouped to threads; C) Tracking three frequency positions; and D) Three frequency derivatives. 66

Comparing average spectral estimates for [s] and [z]. 77

Examples of temporal waveforms after high-pass filtering. 79

Envelope detection. 81

A modulation filtering scheme. 82

Responses of the modulation filterbank. 83

Comparisons of the amplitude modulation detection models. Dashed lines indicate standard deviations. The modulation filtering plots show the outputs of six auditory channels, each filtered by a modulation filter centered at 100 Hz. 84

Overview of the correlational processing. 86

Samples of the correlogram output and super-imposed examples of the summary correlogram decision statistic. Input signals are the same as in Figure 4.6. 88

Three predictions of TMTF data: m is the modulation depth; perceptual data are an average of [Strickland and Viemeister 1997, Dau et al. 1997b]............................. 92

Discriminating the high-pass filtered [s] and [z]: perceptual data is an average across four subjects.. 93

Spectrum of the additive noise used in the recognition evaluations was set to match the long-term average speech spectrum [Byrne and Dilllon 1986]............................... 98

Voicing features in three spectral regions, and their temporal derivatives. 104

Spectral subtraction input/output function. 107

Baseline recognition performance... 109

Recognition comparisons with other signal processing techniques. 110

Evaluations with threading and voicing information. AP is adaptation and peak isolation, and Dcep refers to cepstral derivatives. 111

Threading in a two-stage model of auditory perception.. 118

Overview of the auditory neural pathway. 124

ACKNOWLEDGMENTS

My sincere thanks go to the subjects who gave their time to help me and this work. I thankfully acknowledge my advisor, Professor Abeer Alwan, for stressing the importance of exacting, meticulous research, for critically challenging each step of this effort, and for developing an environment that encourages the exploration of new ideas and approaches. I am grateful to Professor Peter Narins for ingraining in me the significance of evolutionary biology, for sharing a world of auditory expertise, and for including me in his enlightening seminar series. I also acknowledge the UCLA Electrical Engineering faculty, and specifically Professors John Villasenor, Henry Samueli, and Nicholas Bambos for their exemplary teaching and outstanding courses which provided concepts that were useful throughout this dissertation. I am grateful to all my friends at SPAPL, and I specifically thank Jim Hant for our technical discussions, mental gymnastics, and for his patience and friendship. Finally, I acknowledge Richard Lyon and Malcolm Slaney, whose presentations and publications have provided considerable motivation for this effort. This work was supported in part by NIH-NIDCD Grant No. 1 R29 DC 02033-01A1, and the NSF.

Without Jennifer, I would have neither started nor finished.

VITA

May 4, 1967	Born, Bad Kreuznach, Germany
1989	Sc.B. Engineering Brown University Providence, Rhode Island
1989-1993	Workstation Hardware Design Hewlett-Packard Ft. Collins, Colorado
1995	M.S., Electrical Engineering University of California Los Angeles, California
1997	Teaching Assistant University of California Los Angeles, California
1994-1998	Research Assistant University of California Los Angeles, California

PUBLICATIONS AND PRESENTATIONS

A. Alwan, S. Narayanan, B. Strope, and A. Shen, “Speech production and perception models and their applications to synthesis, recognition, and coding,” Proc. of the Int. Symp. Sig. Sys. and Elec., 367-372, October 1995 (Invited).

J. Hant, B. Strope, and A. Alwan, “Durational effects on masked thresholds in noise as a function of signal frequency, bandwidth, and type,” Proc. of the Acoust. Soc. of Amer., Vol. 98, No. 5, 2908, Nov. 1995.

J. Hant, B. Strope, and A. Alwan, “A psychoacoustic model for the noise masking of voiceless plosive bursts,” Proc. of the Int. Conf. Spoken Lang. Processing, Philadelphia, 570-573, Oct. 1996.

J. Hant, B. Strope, and A. Alwan “A psychoacoustic model for the noise masking of plosive bursts," J. Acoust. Soc. Am., Vol. 101, No. 5, 2789-2802, May 1997.

J. Hant, B. Strope, A. Alwan, "Variable-duration notched-noise experiments in a broadband noise context," J. Acoust. Soc. Am., in press.

B. Strope and A. Alwan, "Mapping of constant loudness contours with filter mixtures in digital hearing aids," Lake Arrowhead Conference on Hearing Aid Research, June, 1994.

B. Strope and A. Alwan, "A novel structure to compensate for frequency-dependent loudness recruitment of sensorineural hearing loss," Proc. of the IEEE Int. Conf. Acoust. Speech Sig. Proc., Vol. V, 3539-3542, Detroit, May 1995.

B. Strope and A. Alwan, "A first-order model of dynamic auditory perception," Proc. NIH Hearing Aid Research and Development Workshop, 1995.

B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust speech recognition," Proc. of the IEEE Int. Conf. Acoust. Speech Sig. Proc., Vol. I, 37-40, Atlanta, May 1996.

B. Strope and A. Alwan, "Dynamic auditory representations and statistical speech recognition," Proc. of the Acoust. Soc. of Amer., Vol. 100, No. 4, 2788, Oct. 1996.

B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 5, 451-464, Sept. 1997.

B. Strope and A. Alwan, "Modeling auditory perception to improve robust speech recognition," Proc. of the 31st Asilomar Conf. on Sig. Sys. and Comp., Nov. 1997.

B. Strope and A. Alwan, "Robust word recognition using threaded spectral peaks," Proc. of the IEEE Int. Conf. Acoust. Speech Sig. Proc., Vol. 2, 625-628, May 1998.

B. Strope and A. Alwan, "Amplitude modulation cues for perceptual voicing distinctions in noise," to appear in Proc. of the Acoust. Soc. of Amer., June 1998.

B. Strope and A. Alwan, "Modeling the perception of pitch-rate amplitude modulation in noise," Proc. of the NATO ASI on Computational Hearing, 117-122, July 1998.

ABSTRACT OF THE DISSERTATION

Modeling Auditory Perception for Robust Speech Recognition

by

Brian P. Strope

Doctor of Philosophy in Electrical Engineering
University of California, Los Angeles, 1998
Professor Abeer A. Alwan, Chair

While non-stationary stochastic modeling techniques and the exponential growth of computational resources have led to substantial improvements in vocabulary size and speaker independence, most automatic speech recognition (ASR) systems remain overly sensitive to the acoustic environment, precluding widespread applications. The human auditory system, speech production mechanisms, and languages, on the other hand, are extremely well-tuned to facilitate speech communication in noise. Better modeling of these systems and mechanisms should illuminate robust strategies for speech processing applications. In this work, models of temporal adaptation, spectral peak isolation, an explicit parameterization of the position and motion of local spectral peaks, and the perception of pitch-rate amplitude modulation cues are shown to reduce the error rate of a word recognition system in noise by more than a factor of 4 over the typical current processing.

Chapter 1

Introduction

For the last half century, the development of speech recognition by machine has been marked by disappointment and a repeating failure to meet expectations. A few years ago, a colleague of mine in a very different field attempted to stay a step ahead of my description of this research and commented: “So, with a computer, you will be able to do better than a human.”

As scientists and teachers, we need to do a better job describing the complexity of the problems we address, the realistic capabilities of the tools we use, and the current state of our technologies. Quite plainly, fifty years of speech recognition research still pales when compared to the experiences, the drive and the necessity, a child has when learning to recognize and understand speech. But perhaps more significantly, even given the relative explosion of technology over the last two centuries, the tools available to researchers today are no match to those of the child. Countless generations of selective pressures have evolved signal and information processing machinery that we, as scientists, are unlikely to achieve in the next century.

A recent review [Lippmann 1997] compares human speech recognition to machine recognition across a wide range of tasks (10-65,000 words). Across all tasks, machine error rates are typically found to be between one and two orders of magnitude higher than those for humans. The relative performance of machines is especially poor in noisy situations and in situations where grammatical structure can not be used to constrain the task.

Understanding the failures of speech recognition as a technology leads to a profound respect for natural auditory systems and to the optimistic realization that such a potential wealth of information exists, well, right between our ears.

1.1 From Better Models to Robust Applications

While non-stationary stochastic modeling techniques and the exponential growth of computational resources have led to substantial improvements in vocabulary size and speaker independence, most automatic speech recognition (ASR) systems remain overly sensitive to the acoustic environment, precluding widespread applications. The human auditory system, speech production mechanisms, and languages, on the other hand, are extremely well-tuned to facilitate speech communication in noise. Better modeling of these systems will illuminate robust strategies for speech processing applications.

Perceiving speech in an acoustically noisy environment requires intelligent use of redundant multi-dimensional cues spread over wide-ranging time scales. A majority of psychoacoustic and speech perception research has focused on spectral cues (from roughly 400-8000 Hz) that are available through an auditory critical-band filtering mechanism [Fletcher 1940, Zwicker et al. 1957, Patterson 1976, Zwicker and Terhardt 1980, Glasberg and Moore 1990], and on nearly logarithmic loudness growth [Stevens 1956, Delgutte 1996]. This type of processing is reflected in the first stage of most ASR systems [Rabiner and Juang 1993] which obtains sequences of perceptually-warped and logarithmically-compressed short-time spectral estimations. Subsequent stages form a hierarchy of non-stationary stochastic models operating at the progressively slower rates of the speech-frame, phoneme, word, phrase and even sentence.

Despite these first-order similarities, the current ASR approach differs significantly from human perception. The ASR front end rigidly locks spectral estimates together across a frame, while human perception allows for nearly independent processing, focusing on the position and motion of vocal tract resonances in spectral regions with good signal to noise ratios [Allen 1994]. The ASR front end is time-invariant, while human perception has a context-dependence that can last for hundreds of milliseconds. The typical ASR front end also removes perceptually-salient pitch-rate information from 80-300 Hz, and relies solely on stochastic modeling of syllabic- or articulator-rate information from 2-20 Hz. We should expect synergy from modeling improvements across all these areas.

This dissertation focuses on four aspects of auditory processing: short-term temporal adaptation, an isolation of local spectral peaks, the parameterization of the position and motion of spectral peaks, and the sensitivity to pitch-rate amplitude modulation. Front end processing that incorporates adaptation, peak isolation, and peak motion parameterization is shown to reduce the error rate of a speech recognition system in noise by roughly a factor of 4 when compared to the current common approach. In addition, a quantified model of pitch-perception is shown to predict both amplitude modulation detection thresholds and the perceptual detection of voicing for fricatives in noise.

The relative failure of speech recognition applications (and perhaps of hearing aids) in realistically noisy environments may be directly linked to the lack of successful computational models of the perception of dynamic sounds in noise. The work here on auditory adaptation, the motion of local spectral peaks, and the perception of amplitude modulation in noise represent initial steps toward improved models and robust applications.

1.2 Speech Recognition Overview

In essence, typical speech recognition systems are maximum likelihood detection mechanisms. For clarity, imagine a recognition vocabulary that includes only the words ‘yes’ and ‘no.’ Further assume that a single (1-dimensional) measurement has been developed which usually provides a larger value for ‘yes’ tokens and a smaller value in response to ‘no.’ One possibility might be a measure of the amount of high-frequency energy associated with the [s] in ‘yes.’ To *train* the recognition system, we characterize the distributions of our measurement across a training set of exemplars. We might construct a 2-parameter Guassian *model* of ‘yes’ by estimating the mean and variance of our measurement given all the examples of ‘yes’ in our training data, and then repeat the process for ‘no.’ To use these models for recognition, we make the same measurement for an unknown word, estimate the probability for each of our models, and then choose (or recognize) the word corresponding to the model that provides the highest probability. Figure 1.1 shows an overview for this approach.

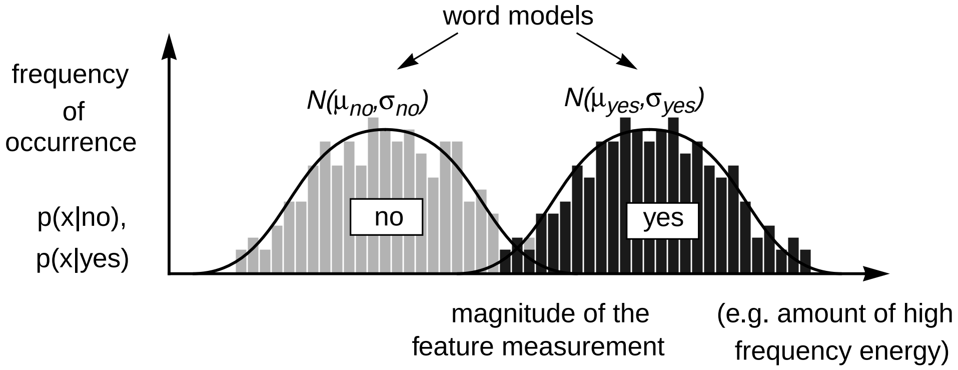


Figure 1.1 Simplified overview of maximum likelihood speech recognition.

At least two issues complicate this process. While any sound can be characterized as a 1-D pressure wave, the information in speech is clearly encoded in a high-dimensional space; a single measurement will not provide sufficient discrimination. For statistical recognition, the measurements become vectors and the estimated distributions are multi-variate.

But the more significant complication is that speech is non-stationary. The statistics that characterize the sound ‘yes’ change considerably across the word. Within the word, there may be temporal segments where the statistics are nearly stationary (e.g. during [s]), but the durations of these segments will also change with different speaking styles and rates. In the current example, consistent differences in these segments, compared across the words ‘yes’ and ‘no,’ provide a redundant and robust encoding of the binary speech information in this task. To exploit this redundancy, the recognition system uses measurements of sounds other than [s] to discriminate the words. In other words, the probability of observing each segment, given each word is estimated.

This motivates the *alignment* problem in speech recognition. In order to use the potentially redundant information encoded in each segment of the non-stationary sound, each segment must be aligned with an appropriate statistical parameterization for that segment in that word. As discussed in more detail below, the current approach to this problem is to model each segment as a state in a first-order Markov process, to associate the parameters of the multi-variate distribution for each segment with each state, and to find the best state alignment using a Viterbi search.

Therefore, most modern speech recognition systems include an initial signal processing front end which converts the (1-D) speech waveform into a sequence of time-varying feature vectors, and a statistical pattern-comparison stage which chooses the most probable phoneme, syllable, word, phrase, or even sentence, given that sequence of feature vectors. Figure 1.2 shows a simplified block diagram for this overview.

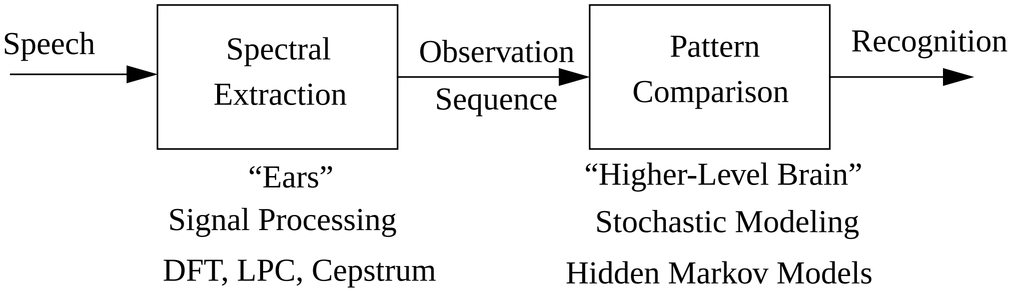


Figure 1.2 Automatic speech recognition overview.

1.2.1 Front End Signal Processing

In the front end, the speech signal is typically divided in time into nearly-stationary overlapping (10-30 ms) frames. Short-time spectral estimations of each consecutive frame form the sequences of time-varying feature vectors analyzed by the pattern matching stage. One common form of spectral estimation [Davis and Mermelstein 1980] involves integrating an initial power spectrum estimate which is weighted by bandpass-filter functions whose bandwidths approximate estimations of auditory frequency selectivity. The magnitude of the power estimates from each filter are then compressed using a logarithmic function. The resulting spectral estimates reflect two of the most studied aspects of auditory signal processing: frequency selectivity, and magnitude compression.

Because the spectral estimates are somewhat smooth across filter number, or highly correlated, each frame can be roughly decorrelated using a discrete cosine transform (DCT) as an approximation of the Karunen-Loeve (KL) transform. After the DCT, the resulting cepstral vectors, called Mel-frequency cepstral coefficients (MFCC), are compact representations of the Mel-warped log-magnitude power spectrum. (The frequency scale is called a Mel-scale after the pitch perception scale that has a similar warping.) Figure 1.3 is an overview of the typical front-end processing.

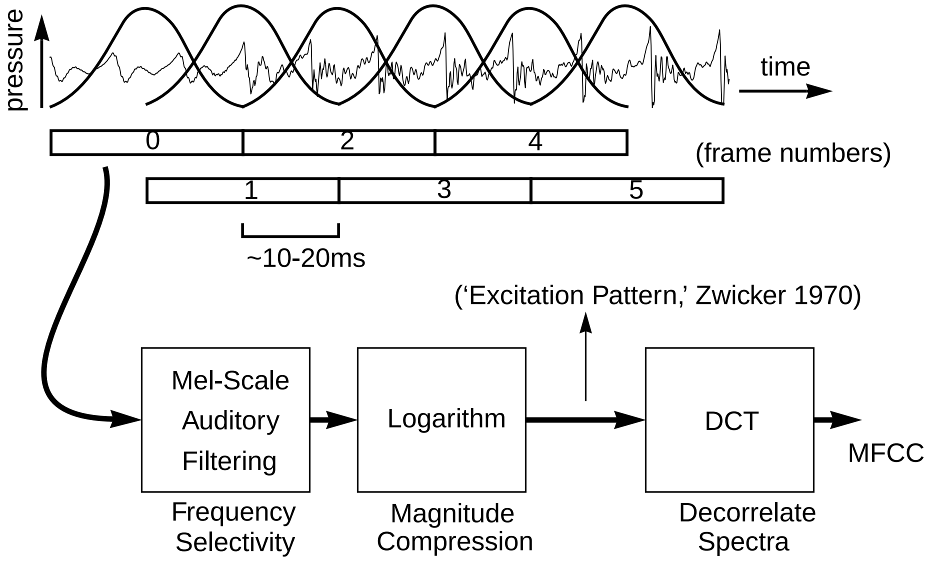


Figure 1.3 Front end signal processing: windowing, spectral estimation, logarithmic compression, and decorrelation.

In Figure 1.4, sequences of spectral estimates for the digit string “nine six one three” at a signal to noise ratio (SNR) of 10 dB are displayed as a spectrogram. The spectrogram is a graphical representation of the sequence of spectral estimates (observation features) provided by the front end as the input to the pattern matching stage. The horizontal axis is frame number (time), the vertical axis is filter number (warped frequency), and the intensity of the feature vector is mapped to darkness.

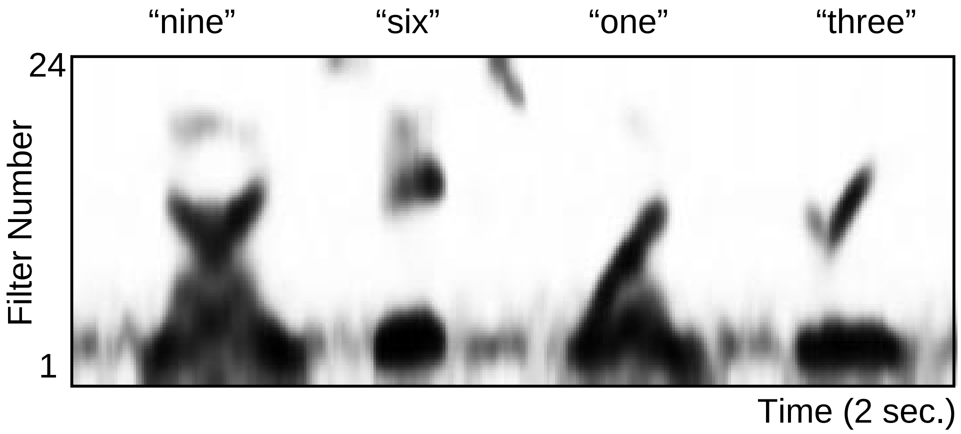


Figure 1.4 Mel-frequency spectrogram at 10-dB SNR.

1.2.2 Pattern Comparison using Stochastic Models

Hidden Markov models (HMM) are used to provide a generalized statistical characterization of the non-stationary stochastic process represented by the sequences of feature vectors. Each element of the vocabulary (word, syllable, or phone) is modeled as a Markov process with a small number of states. During recognition the current sound is compared to each of these models. The model with the highest probability of observing the current sequence of feature vectors determines which vocabulary element is recognized.

The model is hidden in the sense that the observed sequence of feature vectors does not directly correspond to the current model state. Instead, the model state specifies the *statistics* of the observed feature vectors for a specific temporal segment of the sound. State transitions are often limited so that the model can either stay in its current state or move forward to the next. In this way, each state is used to characterize statistics for a nearly stationary temporal segment of the vocabulary element. In word-based recognition, the first state might characterize the beginning of the word, and the last state might characterize the end. Figure 1.5 shows a schematic representation of a four-state model for the word “six.”

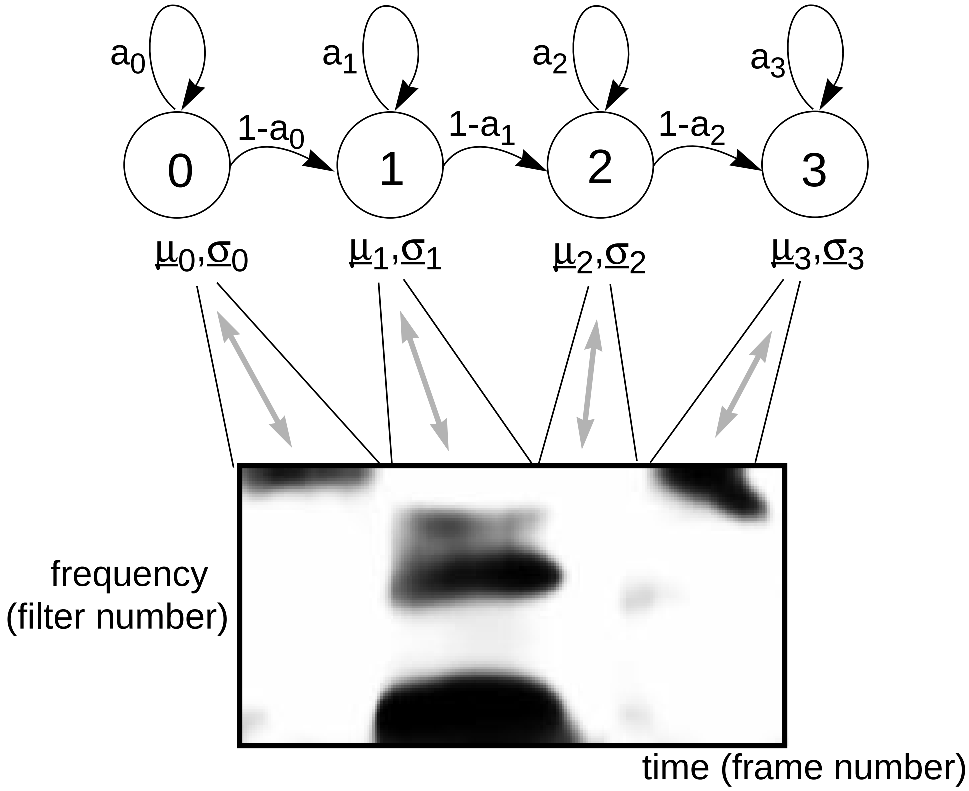

a$_0$
a$_1$
a$_2$
a$_3$
1-a$_0$
1-a$_1$
1-a$_2$
0
1
2
3
$\underline{\mu}_0,\underline{\sigma}_0$
$\underline{\mu}_1,\underline{\sigma}_1$
$\underline{\mu}_2,\underline{\sigma}_2$
$\underline{\mu}_3,\underline{\sigma}_3$
frequency
(filter number)
time (frame number)

Figure 1.5 A schematic representation of a hidden Markov model for ‘six.’

When training an HMM, a set of exemplars corresponding to a particular model are used to provide iterative improvements for both the estimates of the multi-variate distributions of the feature vectors, and the state-transition probabilities. During recognition evaluations, each trained model is compared to the current input, and the most probable model determines the word recognized. Therefore, both training and recognition require solving the temporal alignment problem of matching particular observation frames with particular model states.

Once frames are aligned with specific states during training, new model parameters (observation distributions, and state transition probabilities) are estimated from the statistics of the associated observation frames and state transitions, leading to iterative model improvements.

In general, there are two related approaches used to solve the temporal alignment problem with HMM speech recognition. The first is an application of dynamic programming or Viterbi decoding, and the second is the more general forward/backward algorithm. Both methods can be used for iterative model training and recognition evaluations.

Consider the 4-state model and the sequence of observation vectors (or frames) shown in Figure 1.5. At any point in time, if the current state is known, the probability of the model making a transition to the next state at the next frame is the joint probability of making the state transition and observing the next frame in the next state. For example, if at frame 0 we know the model is in state 0, then the probability that the model transitions to state 1 for frame 1 is the product of ($1-a_0$) times the probability density function (pdf) for state 1 ($N(\underline{\mu}_1,\underline{\sigma}_1)$) evaluated at the feature vector measured in frame 1. In this same instance, the probability of staying in state 0 is the product of a_0 times the pdf for state 0 ($N(\underline{\mu}_0,\underline{\sigma}_0)$) evaluated at the feature vector measured in frame 1.

Figure 1.6 shows a grid of points (or trellis) with the observation frame number on the x-axis and the state number on the y-axis. All possible model alignments for the current sequence of frames are described by paths along the trellis. Evaluating the probability of a particular path requires computing the joint probability of making each state transition while observing each feature vector in the associated state. Because the number of possible paths grows exponentially with the number of observation frames, evaluating each complete path explicitly is not computationally tractable.

However, while the number of possible complete paths grows exponentially with the number of observation frames, all paths must merge into the small number of model states for each input frame. Furthermore, because of the assumed first-order Markov structure, observation probabilities and state transitions are only a function of the current state, and not the path taken to get there.

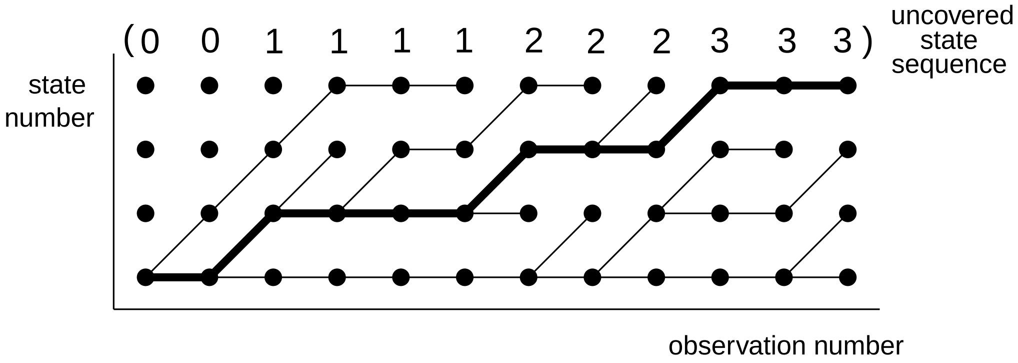


Figure 1.6 An example of using a Viterbi search to solve the alignment problem.

A Viterbi search reduces the computational dependence on the number of observations from exponential to linear, by exploiting the fact that the most probable complete path will necessarily include the most probable partial paths. In other words, if the most probable complete path passes through the {observation number, state number} point (4, 2), then the partial path to (4, 2) is also the most probable path to that point.

To identify the most probable path, it is therefore sufficient, with each new frame, only to keep track of the most probable paths to each state. In the model described in Figure 1.5-6, there are four such paths for each observation frame. Given the first-order Markov structure of our model, the probabilities for these paths are a function the probabilities of the accumulated partial paths at the previous frame, the model state transition probabilities, and the probability of observing the current frame for each state. More specifically for the current frame, the most probable path to each state is the maximum joint probability of the partial paths to the previous frame and the state transition probabilities to that state. As described above, the probability of the partial path to the current frame is then multiplied by the pdf for that state evaluated at the current observation frame.

By keeping track of the most probable state transitions for each observation, the complete "best path" can be identified by back-tracking, as shown in Figure 1.6. In this case, the most probable state sequence was (0,0,1,1,1,1,2,2,2,3,3,3). If this were the only exemplar used for training a model of the word "six," then an updated pdf for state 0 would be parameterized by the mean and covariance of the first two observation frames, and the pdf for state 1 would be determined by the statistics of the next four frames, etc. Similarly, the state transition probabilities (a_i in Figure 1.5-6) can be updated by the frequency of state transitions in the alignment. For example, a_0 would be 0.5, and a_1 would be 0.75, etc. Given the updated model, the observation frames are re-aligned, and the model parameters are re-estimated, and the process is repeated until there is little reduction in the complete path probability. (Models are, of course, trained using many exemplars, so that model parameters at each iteration are a function of the alignment of the current model to several tokens.)

A Viterbi-search solves the temporal alignment problem with an efficient iterative path-building strategy. The probabilities of the partial paths to the current frame are computed from the partial paths to the previous frame. For each transition, the (single) most probable path to each state is added to the growing partial path probabilities. This maximization (or winner takes all) strategy in essence imposes the constraint of identifying a single optimal state sequence during alignment. Because the underlying states are hidden (meaningless), this constraint is not necessary. Instead of choosing the most probable transition when building partial paths, all possible transitions from the previous partial paths that end at that particular {state number, observation number} point can be summed, with each weighted by the associated state transition probability. Notice that because of this summation, there is no way to back-track and associate particular observations with particular frames to update the model estimates.

We therefore consider the forward/backward algorithm. By summing all possible previous partial paths, the partial *forward probabilities* of observing the first N frames of the exemplar and ending at a specific state can be inductively computed from the N-1 forward probabilities. A similar iterative process is used to obtain *backward probabilities* of observing the last M frames. Combining the forward and backward probabilities provides an estimate for the probability of making each state transition while observing each frame, given the entire exemplar.

New model parameters (state transition probabilities and observation distributions) are again obtained by averaging across all exemplars in the training set for each model. However unlike with Viterbi-training, the contribution of each state transition and each observed feature vector are weighted by the probabilities of having been at that state during the time of that feature vector. If we set the most probable transition into each {observation number, state number} point to 1 and the rest to 0, the forward/backward training reduces to Viterbi training. Given the new set of re-estimated models, the algorithm iterates, re-aligning the original data to the updated models. As with training using Viterbi alignment, this iterative process converges to a local maximum for the likelihood of the model given the training set of exemplars.

Once models for the vocabulary elements are trained (their parameters have converged sufficiently after successive forward/backward iterations), new data are recognized by evaluating the probability of observing that data given each of the trained models. For alignment, the same iterative propagation of forward probabilities used in training can be used. However, a Viterbi search is often used as a simplification. That is, instead of adding the probabilities of reaching a particular state at a particular vector in the feature vector sequence, only the transition that reaches that particular state with maximum probability is considered. In addition to keeping highly unlikely state sequences from influencing the final result, this simplification leads to the identification of the single best path which can be helpful for analyzing errors.

Recognition performance is, obviously, largely dependent on a good statistical match between the test and training feature-vector sequences. Because most systems use short-time spectral estimates, distortions introduced by additive noise, or by a mis-match between the training and testing channels, considerably degrade recognition performance. One general approach to this problem is to find a parametric adjustment of the multi-variate distributions given the current acoustic environment [Gales and Young 1996]. A more pragmatic approach is simply to train the models in an environment that is a reasonable match to the expected testing environment.

This dissertation describes several methods to obtain a more perceptually relevant characterization of speech and to improve recognition robustness. First, the front-end signal processing is augmented to include short-term adaptation and a sensitivity to local spectral peaks. Second, the frequency position and motion of the local spectral peaks are explicitly tracked and then parameterized by the HMM.

Third, the presence of pitch-rate temporal information that is typically ignored for ASR is parameterized. Fourth, two sets of models are used in parallel: one characterizing clean training data and the other characterizing noisy training data. The first three approaches attempt to focus the recognition task on phonetically relevant aspects of the sequences of short-time spectral estimates, while the fourth technique provides some adaptation of the statistical characterization for the expected acoustic environment.

In many ways, the current speech recognition paradigm is a direct application of the maximum likelihood models used in psychoacoustics [Green and Swets 1966]: the signal processing front end generates a sequence of spectral estimates (or a sequence of auditory excitation patterns [Zwicker 1970]), and an optimal decision device chooses the recognized words using a maximum likelihood criterion. For general speech recognition, unlike with many simplified psychoacoustic models, ‘multiple-looks’ [Viemeister and Wakefield 1991], or multiple measurements, are used in both time and frequency.

1.3 Auditory Modeling Overview

Recent texts [Pickles 1988, Geisler 1998] review the current understanding of the (human) auditory periphery. A brief review is included here to motivate some of the modeling discussed later in this dissertation.

Acoustic pressure waves pass through the nearly (passive and) linear outer and middle ears and vibrate the basilar membrane within the snail-shaped cochlea of the inner ear. Vibrations along the basilar membrane modulate the release of neurotransmitter by hair cells, which in turn, leads to action potentials (or spikes) in the auditory nerve. In this transduction process, information about the mechanical vibrations of sound are transformed into electrical signals processed by more central neural regions.

Vibrations of the middle ear are coupled to the beginning, or base, of the basilar membrane. This membrane, which spirals in a helix up through the cochlea, functions as a non-uniform waveguide. Its stiffness decreases along its length so that wavespeed drops nearly geometrically from base to apex (ca. 35 mm in humans). If we consider a sinusoidal input starting at the base, as the traveling wave moves apically, the wavefront slows, decreasing the wavelength of the membrane disturbance and concentrating the energy per unit length over an increasingly smaller region. Finally losses due to the deformation of the membrane dominate, and the traveling wave dissipates abruptly. The location along the membrane where this occurs varies with the initial wavelength (or acoustic frequency) of the input. High frequency sinusoids concentrate and dissipate close to the base, while lower frequency sinusoids travel further toward the apex. If we consider each point along the basilar membrane as an output, the nonuniform waveguide is an efficient mechanical implementation of a filter bank, providing a physiological substrate for auditory frequency selectivity. Figure 1.7 [after von Bekesy 1953] shows a schematic overview for this mechanism. (Interestingly, while this idea had been previously applied extensively to explain psychoacoustic masking experiments [Fletcher 1940], direct physiological observations of the traveling wave [von Bekesy 1953] led to winning the Nobel prize.)

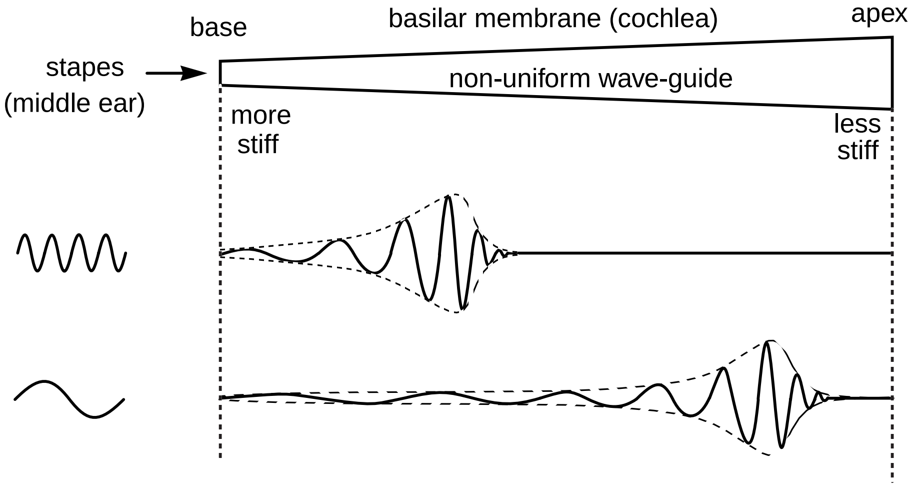


Figure 1.7 Mechanical frequency selectivity of the basilar membrane.

The deflections of tiny stereocilia (actually microvilli and not true cilia) in hair cells, located throughout the length of the basilar member, modulate the release of neurotransmitter by the hair cells. This transduction is directional: increasing positive deflection leads to increasing neurotransmitter release, while negative deflections cause little, if any, neurotransmitter release regardless of the deflection magnitude. Hair cells, therefore, provide half-wave rectification.

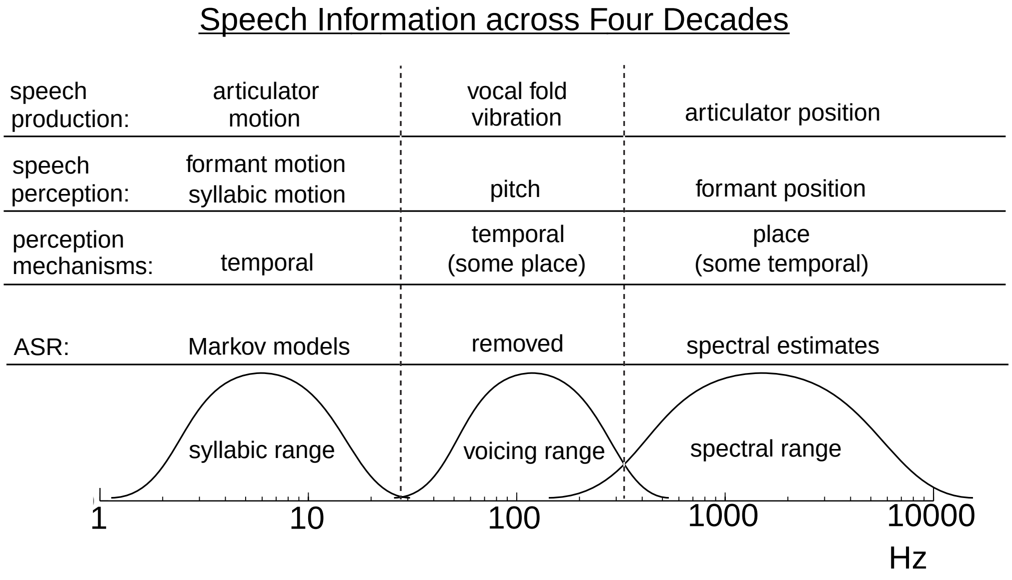
In humans, roughly 30,000 auditory neurons connect to these hair cells, and generate action potentials in response to the hair cells' neurotransmitter release. The generation of individual action potentials is stochastic, but when averaged across the ensemble of parallel auditory neurons passing information to more central regions, the action potentials can encode both the AC component of the original vibration as well as a demodulated DC component (which results from the hair-cell half-wave rectification) [Palmer and Russel 1986]. Intrinsic capacitances limit the upper frequency of the AC encoding to roughly 2-5 kHz [Joris et al. 1994]. Above 5 kHz, there is little evidence of an AC component in the temporal patterns. While the AC component generally provides a remarkably consistent linear representation of (a filtered version of) the original acoustic vibration, the DC component is considerably compressed by mechanical non-linearities in the cochlea and by adaptation and saturation in the hair-cell transduction process.

Throughout modern auditory research, there are at least two interpretations of the neural representation of sound. Considering the mechanical frequency selectivity of the basilar membrane, and the DC response encoded in the short-term average firing rate of auditory nerve action potentials, *place* theories assume short-time spectral information is encoded in the changes of average rate with the position of neural innervation along the basilar membrane. In an auditory model, the logarithm of the average intensity of the output of each filter is used to form a spectral excitation pattern [Zwicker 1970]. Differences in spectral excitation patterns have been shown to correlate with many psychoacoustic measurements including pure-tone frequency jnds (just noticeable differences) and intensity discrimination [Siebert 1968, Zwicker 1970, Delgutte 1996], spectral masking

[Patterson 1976, Glasberg and Moore 1990], and vowel discrimination [Kewley-Port and Watson 1994]. *Temporal* theories assume that the filtering of the basilar membrane provides some rejection of out-of-band noises (a filter processing gain), and that the AC components are then analyzed by more central neural centers which are sensitive to redundant *fine-time* structure available across multiple auditory nerves. Temporal theories have perhaps been most successful for describing aspects of pitch perception [Licklider 1951, Goldstein and Srulovicz 1977, Meddis and O'Mard 1997].

While there are many interesting exceptions [e.g. Lyon 1984, Deng and Geisler 1987, Seneff 1988, Ghitza 1991, Patterson et al. 1992, Potamianos and Maragos 1996], the majority of engineering applications that analyze speech use place models. As described in the ASR overview above, most use short-term log-magnitude power spectrum estimates. The output of these systems approximate those of spectral excitation patterns used in place models. Differences include: the implementation of the initial filtering (Fourier transforms, or linear prediction is often used instead of explicit filters); the non-linearity (often demodulation is achieved through squaring instead of half-wave rectification); and the low-pass filtering (usually temporal integration over a short-time analysis window, instead of an explicit low-pass filter). Most current discussions of auditory modeling assume place and temporal processing are used simultaneously [e.g. Moore 1973, Delgutte 1996].

When discussing temporal processing it is helpful to be specific about the rates considered and the distinction between modulator and carrier. The spectral analysis of speech in place models assumes that fluctuations in the range of roughly 400-8000 Hz are either mechanically analyzed in the cochlea, or that, together with any contribution of more central neural analyses of the temporal fine structure in auditory nerve firings, the complete system response to fluctuations at these rates is well modeled by log-magnitude excitation patterns. Here, this range (about 400-8000 Hz) will be called the *spectral* range (see Figure 1.8). The temporal models cited above include mechanisms which analyze temporal detail in the spectral range.



speech production:	articulator motion	vocal fold vibration	articulator position
speech perception:	formant motion syllabic motion	pitch	formant position
perception mechanisms:	temporal	temporal (some place)	place (some temporal)
ASR:	Markov models	removed	spectral estimates

Figure 1.8 Classifying auditory frequency ranges for speech.

A linear model of speech production [Fant 1960] is helpful when considering Figure 1.8. Briefly, when we speak, air from the lungs is forced through the vocal folds which open and then slam shut with nearly regular periodicity, typically in the range of 100-200 Hz. (The vocal fold dynamics have similarities to those that occur at the lips when either a child makes a ‘raspberry’ sound, or a trumpet player blows into the mouthpiece.) Because the vocal folds close abruptly, harmonics of the fundamental are also produced. The geometric configuration of the vocal tract, as controlled by the position of the articulators (tongue, lips, teeth, jaw, etc.), determines how the harmonics are resonated in different spectral regions. For speech, these resonances are called formants. During unvoiced sounds (e.g. [s]), turbulence is usually generated by forcing air through a sufficiently narrow constriction somewhere in the vocal tract. Again, vocal tract resonances influence the spectral shape of these noise-like speech sounds. A simplified linear speech production model, therefore, includes either a noise source and/or a periodic driving function (often approximated by an impulse train), and a linear filter to model the resonances of the vocal tract. Figure 1.9 [after Geisler 1998] shows an overview for the speech production process.

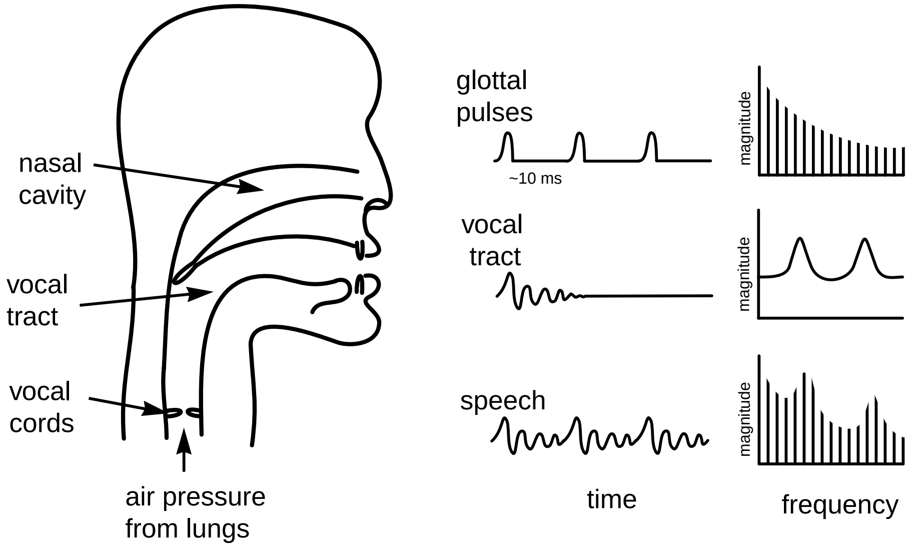


Figure 1.9 Speech production overview.

The right side of Figure 1.8 shows that formants, as determined by the position of the articulators, occur in the spectral range. Signals in the spectral range are first processed by the auditory system as carriers. Non-linearities in the auditory system (and in ASR front ends) demodulate the carrier signal to DC. Slow fluctuations in the amount of energy in a particular spectral region cause slow fluctuations in the "DC" response, which tracks the envelope of the carrier. The left side of Figure 1.8 shows that changes in the configuration of the articulators, alter the vocal tract resonances, modulating the spectral range in specific regions. Perceptually, these changes typically cue syllabic and word-level information as the speech progresses from one sound to the next. Here, we will refer to this range (roughly 2-20 Hz) as the *syllabic* range. Quite remarkably, and as one of the best recent examples of the redundancies in speech, syllabic range modulations, together with extremely limited frequency resolution (3 or 4 noise carriers), are sufficient for human speech recognition [Shannon et al. 1995]. After the demodulation of the noise carriers at specific places along the basilar membrane, the perception of these sounds must depend entirely on temporal processing.

The fundamental frequency of vocal fold vibrations occurs between the spectral and syllabic regions in the *voicing* range (center column in Figure 1.8). When the spacing of harmonics are significantly greater than the bandwidths of auditory filters, low-frequency harmonics can be resolved in different auditory filters, providing a potential place cue for voicing information (especially for high-pitch voices). On the other hand, temporal representations of information in the voicing range exist for all harmonics. High frequency harmonics (above 1-2 kHz) are not resolved by the auditory system: each auditory filter, or channel, contains multiple harmonics. The response in each channel is, therefore, modulated at the fundamental frequency of vocal fold vibration. After half-wave rectification and low-pass filtering, this modulation is well represented in the temporal firing patterns of the auditory nerve [Cariani and Delgutte 1996a-b]. Below 1-2 kHz, there are increasingly fewer harmonics in each auditory channel, reducing the modulation at the fundamental. However, as the number of harmonics per channel is decreasing, the ability of the auditory system to represent temporal information is increasing.

That is, the harmonics below 1-2 kHz can be represented directly and individually in the temporal firing patterns of the auditory nerve. Temporal pitch perception models assume more central neural processing pools the timing information across channels providing a composite response dominated by the common periodicity (i.e. the fundamental) [Licklider 1951]. Therefore, regardless of whether the harmonics are grouped mechanically in the initial filter, or subsequently in the temporal comparison of the neural representations of individual harmonics, the auditory system provides a strong temporal coding of pitch information.

As shown in Figure 1.8, the typical ASR system characterizes the spectral region using compact representations of log-magnitude spectral estimates, and characterizes the syllabic region with non-stationary stochastic modeling. The voicing region is usually ignored. In this dissertation, the first three areas considered: adaptation, peak isolation, and the explicit parameterization of the position and motion of spectral peaks, influence both spectral and syllabic representations. More precisely, the adaptation mechanism most directly influences the syllabic range, peak isolation modifies representations in the spectral range, and the explicit parameterization of the position (spectral) and motion (syllabic) influences representations in both. Finally, this dissertation includes modeling of the temporal processing in the voicing range.

1.4 Dissertation Overview

Chapter 2 describes the adaptation and spectral peak isolation mechanisms.

Chapter 3 describes an algorithm developed to parameterize of the position and motion of local spectral peaks. Chapter 4 focuses on the temporal aspects of the voicing distinction between strident fricatives, and compares predictions from three modeling approaches. Chapter 5 describes a series of recognition evaluations. Finally, the summary in Chapter 6 includes an outline of other research directions motivated by this work.

Chapter 2

Adaptation and Local Peak Isolation

This chapter describes two mechanisms which augment the common ASR front end and provide temporal adaptation and isolation of local spectral peaks. A dynamic model consisting of a linear filter bank with a novel additive logarithmic adaptation stage after each filter output is proposed. An extensive series of perceptual forward masking experiments, together with previously reported forward masking data, determine the model's dynamic parameters. Once parameterized, the simple exponential dynamic mechanism predicts the nature of forward masking data from several studies across wide ranging frequencies, input levels, and probe delay times.

2.1 Introduction

Most ASR systems model speech as a non-stationary stochastic process, by statistically characterizing a sequence of spectral estimations. The common technique for spectral estimation includes an approximation of auditory filtering, a compressive non-linearity (usually the logarithm) and decorrelation of the spectral estimation through an approximate KL transform (the DCT). These steps represent only rough approximations of the most fundamental aspects of auditory modeling: frequency selectivity and magnitude compression. In the last 5-10 years the frequency selectivity for ASR front-ends has slowly migrated from a linear to a perceptually-based frequency scale [Davis and Mermelstein 1980]. This progress, toward a better auditory model for ASR, has improved robustness [Jankowski et al. 1995].

A large discrepancy remains between current auditory models and the approximations used in ASR front ends. Recent efforts to incorporate more sophisticated auditory models with ASR systems, however, have shown little to no improvement over the common front end, typically at a severe increase in computational costs [Jankowski et al. 1995]. The challenges are to determine what auditory functionality missing from the current front end would be useful for improving recognition robustness and to design effective simple mechanisms which reproduce that functionality.

This chapter focuses on two aspects of audition not included in current representations: short-term adaptation and sensitivity to the frequency position of local spectral peaks. For each, a mechanism with low computational complexity is described which adds to the common front end and provides a representation that is more robust to background noise. The dynamic mechanism is parameterized by psychophysical data described here and in the literature [Kidd and Feth 1982]. The peak isolation mechanism is a simple modification of a previous cepstral liftering technique [Juang et al. 1987].

To incorporate a dynamic mechanism within a front end, a method of quantifying auditory adaptation must first be identified. There is considerable physiological and psychophysical evidence of dynamic audition. Short-term adaptation, usually defined as a decreasing response after the onset of a constant stimulus, has been measured in individual auditory nerve firings [Smith and Zwislocki 1975]. The neural response to a stimulus is also reduced during the recovery period following adaptation to a prior stimulus [Harris and Dallos 1979]. Here the general term *adaptation* is used for both dynamic processes (short-term adaptation and post-adaptation recovery), and its direction is explicitly specified when significant. *Attack* refers to the decreasing response following stimulus onset, while *release* and *recovery* both refer to the increasing response following stimulus offset. Motility of outer hair cells, the likely source of an active cochlear response, also adapts with time constants which may be significant when quantifying short-term adaptation [Ashmore 1987]. Finally, neural responses to onsets and abrupt spectral changes are substantial [Delgutte and Kiang 1984], providing a physiological substrate for the sensitivity of human speech perception to onsets and dynamic spectral cues [Furui 1986]. Although recognition systems typically statistically characterize the evolution of relatively static spectral segments, the auditory system responds most strongly to dynamic segments. This response strength can be viewed as a consequence of adaptation. What remains is to quantify the adaptation, and to design a mechanism which reproduces the functionality.

The task is similar to observing evidence of frequency selectivity and requiring a specification (critical bandwidths) and a mechanism for its realization (a filter bank). Following the example of using static masking data to quantify frequency selectivity [Fletcher 1940], adaptation was quantified from a series of dynamic, forward-masking experiments. The adaptation mechanism designed is a modified form of automatic gain control (AGC) which adds an exponentially adapting linear offset to logarithmic energy. Just as the current triangular filters used in the common ASR front end are first-order approximations of auditory frequency selectivity, the simple dynamic mechanism provides only a first-order approximation of auditory adaptation. The strategy is to parameterize simple dynamic mechanisms from forward masking thresholds to provide a better approximation of the auditory response to dynamic stimuli.

Dynamic auditory models [e.g. Cohen 1989, Goldhor 1985, Kates 1991, Lyon 1982, Lyon 1984, Lyon and Mead 1988, Seneff 1988] are often physiologically-based computational models which characterize a relatively low level of the complete auditory system, or resort to some speculation either about higher-level processing and/or about appropriate dynamic parameters. Because these systems often require processing time-domain signals for each auditory filter (~100 filters) at the full sampling rate, they imply a large computational burden, making them difficult to use in engineering applications [Jankowski et al. 1995]. Also, successfully separating and quantifying measurable functionality (e.g. frequency selectivity, or short-term adaptation), which may be distributed across several related physiological processes, is not a simple task. Other researchers [Aikaiwa and Saito 1994, Hermansky and Morgan 1994] have proposed novel computationally efficient techniques, targeted at automatic speech recognition, which emphasize spectral dynamics with varying perceptual accuracy and recognition improvements. The approach here differs from most detailed physiological models in that it 'closes the loop' with observations of top-level functionality. Because the relatively simple model of frequency selectivity followed by additive adaptation is consistent with underlying physiological processes, the resulting quantified non-linear model provides useful approximations of the perception of non-stationary speech.

2.2 Forward Masking

Forward masking reveals that over short durations the usable dynamic range of the auditory system is relatively small, and largely dependent on the intensity and spectral characteristics of previous stimuli. A probe following a masker is less audible than the probe following silence. As the duration between the masker and probe decreases, the probe threshold is increasingly a function of the intensity of the preceding masker, and decreasingly a function of the absolute probe threshold in silence. Forward masking can be viewed as a consequence of auditory adaptation. After adaptation to the masker, recovery time is necessary before the relatively less intense probe becomes audible. The amount of forward masking is also a function of the duration of the masker, reflecting the time required for the auditory system to adapt completely to the masker. Forward-masking, therefore, provides an opportunity to measure the rate and magnitude of effective auditory adaptation and recovery.

To build the dynamic model, data describing sinusoidal forward masking, were desirable. The most complete data of pure-tone forward masking experiments is from [Jesteadt et al. 1982]. Although this data includes a wide range of frequencies and masker levels, the longest probe delay measured is only 40 ms, short of the duration necessary for complete adaptation. To obtain recovery parameters, a set of pure-tone forward-masking experiments which included probe delays from 15 to 120 ms across wide ranging frequencies and masker levels was performed. Short-delay pure-tone forward-masking data from the literature [Kidd and Feth 1982] as a function of masker duration were used to quantify attack parameters.

2.2.1 Experiments

The forward-masking experiments used long-tone maskers followed by short tone-like probes of the same frequency and phase. The masker was long enough to ensure complete auditory adaptation before masker offset, while the probe was short enough to measure the response of the auditory system at a relatively specific time. A two alternative forced choice (2AFC) experimental paradigm was used.

2.2.1.1 Stimuli

Figure 2.1 shows an example of the stimuli. A decaying 60 ms probe tone followed one of two 300 ms maskers, separated by 500 ms (Fig. 1.a). The subjects chose which masker the probe followed. Masker and probe frequencies ranged from 250-4000 Hz in octave intervals, probe delays were 15, 30, 60, and 120 ms, and masker levels spanned roughly 50 dB with three points. All signals were ramped on and off in 5 ms with the appropriate half period of a raised-cosine. Probe-delay times are specified between the peaks of the envelopes of the masker offset and probe onset.

In forward masking, it is often difficult to determine what cue subjects are using, or when the subject detects the probe. The solution here is similar to that in [Plomp 1964]. Both the probe and the masker in the non-probe interval decay with the same 20 ms time constant, and both end at the same time relative to the masker onset. Detecting the probe onset was a sufficient cue to determine the probe interval, but detecting a decaying sinusoid (the tail of the probe) was not. Subjects were not given feedback.

To reduce the spectral splatter of transitions, the entire stimulus was filtered through a linear-phase, FIR filter, with a bandwidth of one critical band [Zwicker and Terhardt 1980]. In the Figure 2.1 example, the frequency is 1 kHz (Figure 2.1.B), the delay from masker to probe is 15 ms (Figure 2.1.C), and (measured at the envelope peak) the probe is 8 dB less intense than the masker. The stimulus is shown after the critical band filter.

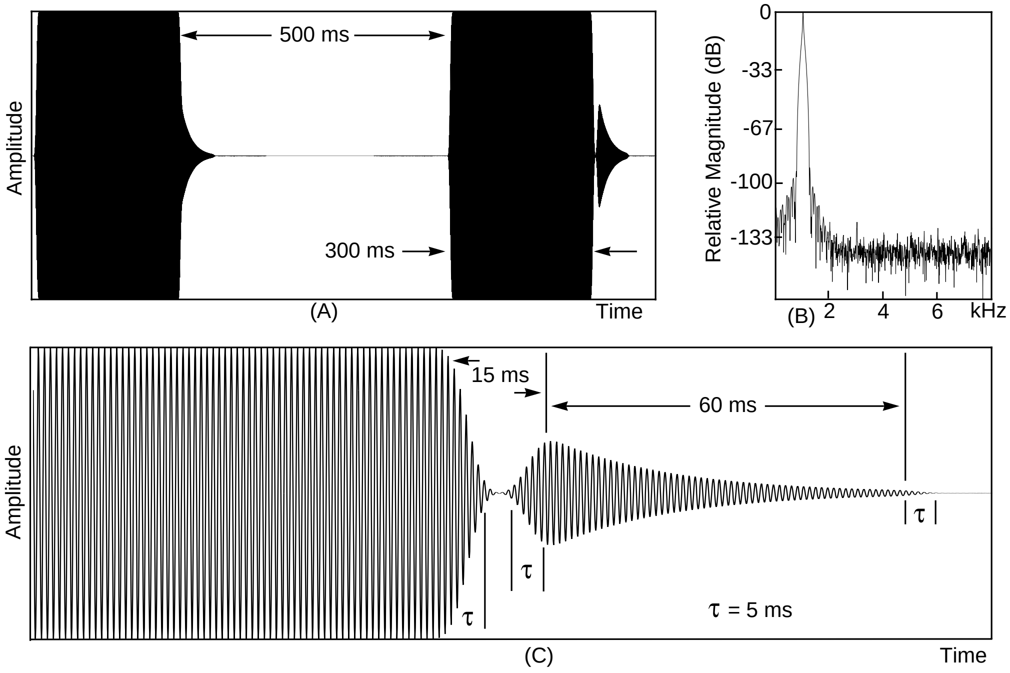


Figure 2.1 Forward masking stimuli: (A) Large time-scale view of a single 2AFC trial; (B) Fourier Transform of the probe signal (128 ms rectangular window); (C) Smaller time-scale view of the probe following the masker by 15 ms.

2.2.1.2 Subjects

Five subjects, including the first author, participated in the experiments. All are native speakers of American English. One subject is female, and the others are male. Their ages ranged from 23 to 28 years. Hearing thresholds for each were at, or below, 20 dB HL at frequencies used in this study.

2.2.1.3 Methods

For each condition, the level of the probe was adaptively varied to find its threshold. An adaptive “transformed up-down” procedure [Levitt 1971] determined the 79% correct point, defined as the threshold for the 2AFC task. The initial adaptation step size of 4 dB was reduced to 2 dB and 1 dB after the first and third reversals. The initial probe was clearly audible. The experiment continued for nine reversals. The probe levels at the last six reversals were averaged to determine a threshold. Thresholds were averaged across the five subjects to obtain the values used for parameterizing the model.

2.2.1.4 Equipment and Calibration

Computer software generated digital stimuli on-line. The sampling rate was 16 kHz, and the quantization was 16-bit linear. An Ariel Pro Port 656 converted the digital samples into an analog waveform, and the pre-amp of a Sony 59ES DAT recorder drove TDH-49P earphones. Tests were performed in a double-walled sound-isolated chamber. Stimuli were presented binaurally with identical waveforms to each ear. The system was calibrated by measuring the response to digitally synthesized sine-waves using a 6-cc coupler and a Larson-Davis 800B Sound Level Meter. Pre-amp levels and digital internal level offsets were set to place an 80 dB SPL 1kHz tone within 0.2 dB. A linear-phase FIR equalization filter was adjusted until pure tones from 125-7500 Hz measured within 0.5 dB.

2.2.2 Results

Figure 2.2 summarizes the average threshold increase (circles) across the five subjects as a function of masker level with probe delay as a parameter. The solid lines in Figure 2.2 indicate the model's fit to the forward masking data. The derivation of the model is described in the following sections.

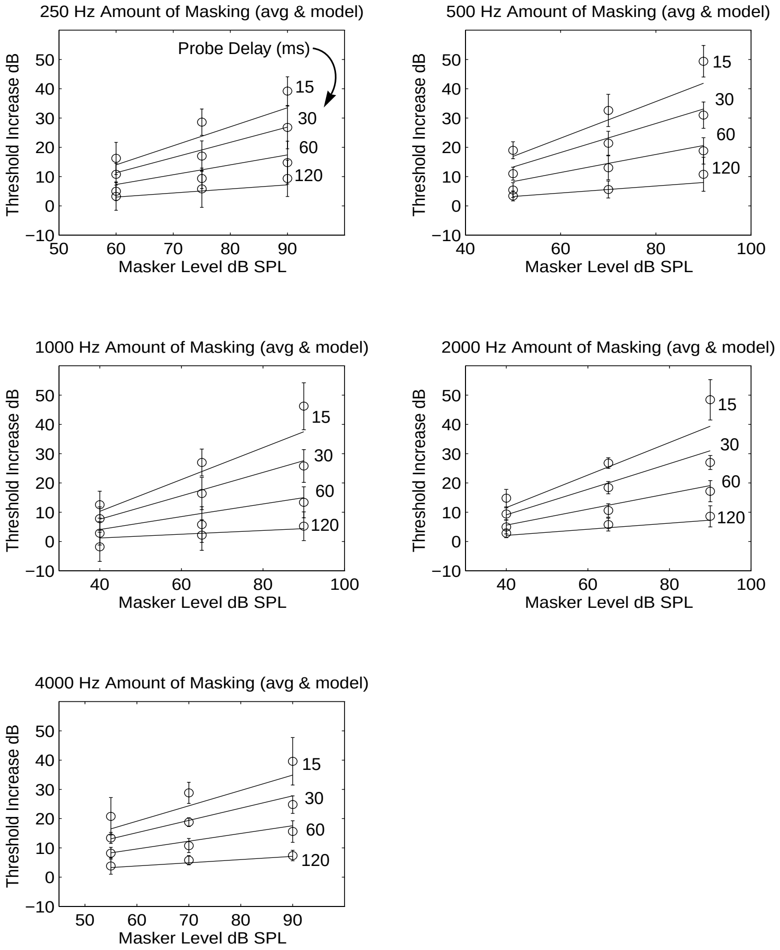


Figure 2.2 Average forward masking data (circles), and std. dev. (error bars),

together with the model fit (lines) as a function of masker level across 5 octaves, with probe delays of 15, 30, 60, and 120 ms as a parameter.

2.2.3 Modeling Implications

The amount of forward masking (in dB) decays as a straight line as a function of the logarithm of the probe delay (first described in [Plomp 1964]). A straight line with respect to logarithmic probe delay can be approximated by an exponential with respect to linear probe delay. This suggests additive exponential adaptation in dB.

Figure 2.3.A plots the threshold increase as a function of probe delay, and Figure 2.3.B shows the effective dynamic range below masker, defined as the difference between the masker and probe threshold levels, as a function of masker level. Figure 2.3.A shows that the rate of decay of the forward masking (shown on a log time scale) increases with an increasing amount of masking. These data may suggest different adaptation rates for different masker intensities, or complexity beyond a simple exponential adaptation of dB level. Such complexity is not necessary. The adapting mechanism derived below has a greater initial distance to target after a more intense masker offset. Exponential processes decay more quickly over the same amount of time when the output is further from the final static target. Therefore, a simple exponential dynamic mechanism can predict a faster rate of decay of forward masking with more intense maskers.

Figure 3.B shows that even at short delays the dynamic range below masker depends on the level of the masker. At short delays there is little to no time for adaptation. Without time for adaptation, the static characteristics of the dynamic mechanism determine the forward masking threshold.

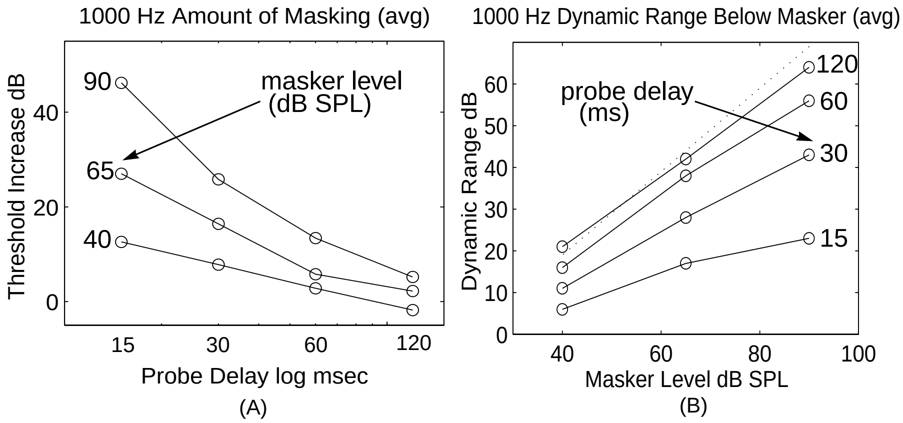


Figure 2.3 Average forward masking data at 1kHz: (a) as a function of the log delay with masker level as a parameter; and (b) as the dynamic range below masker as a function of the masker level with probe delay as a parameter. The dotted line reflects the probe threshold in quiet.

2.3 From Experimental Results to Model Parameters

In the perceptual model, a dynamic adaptation stage follows each output of a linear filter bank. At each point in time, each adaptation stage slowly adjusts an internal offset to move its output incrementally closer to an I/O target.

The dynamic adaptation stages are referred to as automatic gain control (AGC). However, it is significant that the AGC is implemented as an adapting additive offset to the log energy of the signal, and not as an adapting multiplicative gain. There are at least two points that appear to require additive, and not multiplicative, adaptation. First, the measured incremental neural response to a second onset after partial adaptation to a first is not proportional to an adapted amount of multiplicative gain [Smith and Zwislocki 1975]. Second, AGC that adjusts a multiplicative gain proportional to the linear distance to the I/O target does not predict a higher rate of decay of forward masking for greater amounts of masking.

2.3.1 AGC: I/O Curves, Attack and Release Times

Time constants describing the rate of adaptation for the dynamic mechanisms are defined as the time required for the logarithmic distance to target to reduce by a factor of *1/e*. Different time constants are used for attack (decreasing offset), and release (increasing offset). Over short durations, the AGC stage has little time to adapt, and is therefore nearly linear. On an I/O curve, when the input changes abruptly, the output initially tracks the input, moving in nearly a 45 degree line. Over long durations with static inputs, the output approaches the I/O target.

Figure 2.4.A shows a prototypical I/O curve for a single channel in the dynamic model. At low levels, the I/O function is nearly linear, over normal levels it is compressive, and at extremely high levels it is again linear. The general shape of the prototypical I/O curve was motivated by the saturating response of the basilar membrane [Johnstone et al. 1986]. For each adaptation stage, a fixed internal threshold, corresponding to the static audibility threshold, is imposed at the compression threshold. Similarly, the compression region ends, and the model again becomes linear, at a high level of equal loudness (near 90 dB SPL) as a function of the center frequencies of each adaptation stage. By carefully choosing the threshold and I/O curve for each adaptation stage, the AGC sections map a specified static input range as a function of center-frequency into a normalized internal level consistent with constant loudness contours.

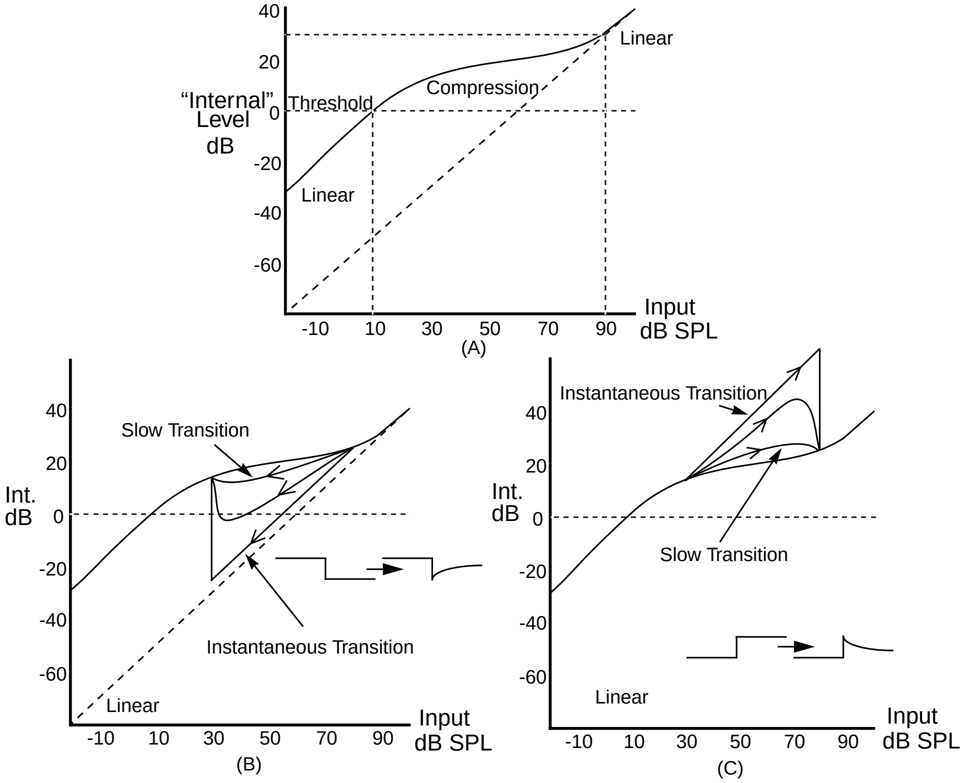


Figure 2.4 (A) A prototypical I/O curve for a single channel in the dynamic model; and, schematic output trajectories corresponding to a level change at three different rates for (B) decreasing inputs from 80 to 30 dB SPL, and (C) increasing inputs from 30 to 80 dB SPL.

Figure 2.4.B-C schematically show the response of the model to decreasing and increasing inputs, respectively. When the input changes abruptly, the trajectory on the I/O curve moves nearly in a 45 degree angle, and then eventually settles to the target on the I/O curve. When the input changes slowly, the output trajectory follows the I/O curve more closely. The model predicts forward masking when output trajectories momentarily fall below the internal threshold, as in Figure 2.4.B.

2.3.2 Derivation of Model Parameters

The model's forward-masking prediction is derived from the response of the dynamic mechanism to forward-masking stimuli. When the output of the adapting (dynamic) mechanism is just at threshold during the onset of the probe, the model predicts a forward-masking threshold.

To simplify the model and this derivation, a constant I/O slope is imposed across the compressive region. Figure 2.5 describes the geometries necessary to measure the model's prediction of the forward masking threshold with long maskers as a function of masker level and probe delay. Before the masker offset, the output trajectory reaches the target on the I/O curve (point A in Figure 2.5). As the masker shuts off abruptly, the output trajectory instantly falls along the diagonal (from A to B). Once the trajectory is below the compressive region, the distance to target is constant, and the model adapts by slowly increasing toward maximum additive offset (from B toward C). At some point during this adaptation (point C), the onset of the probe causes an abrupt transition from below threshold back up along a new diagonal (from C to D). If the probe level is intense enough to place the trajectory above threshold (at the instant of the probe onset) the probe is audible. If the internal level just reaches threshold, the model predicts a forward masking threshold (at point D).

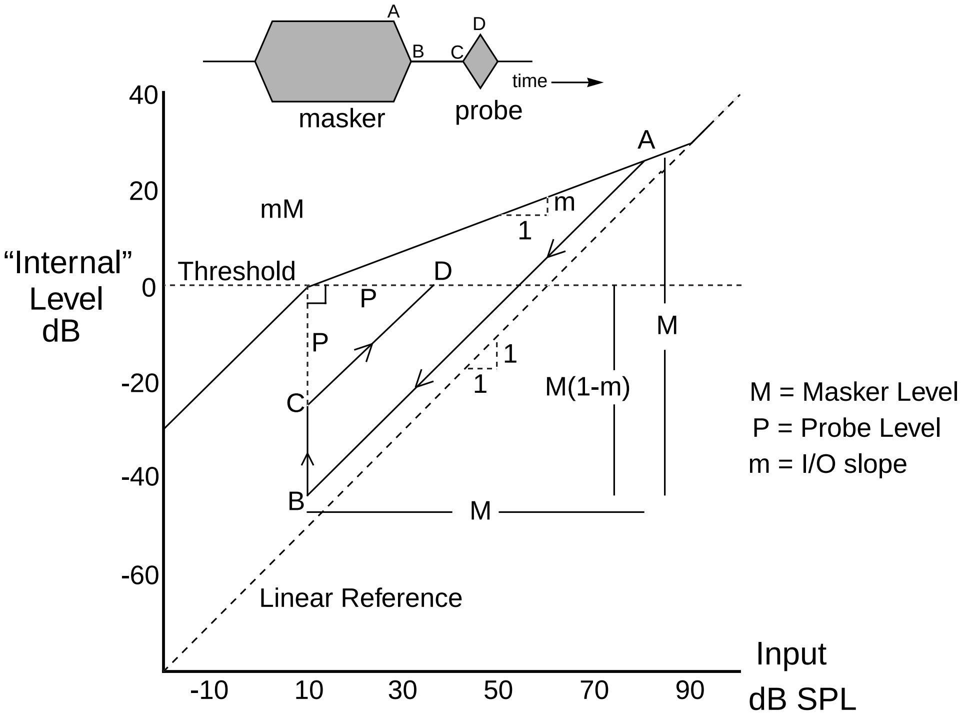


Figure 2.5 Geometry to derive recovery (upward adaptation) parameters from forward masking thresholds.

Incremental adaptation of the model is implemented using a (non-constant coefficient) first-order difference equation leading to an exponential decay of the logarithmic distance to target. From the geometry in Figure 2.5, the probe level at threshold *P* as a function of masker level *M*, discrete-time probe delay *n*, I/O slope

m, and incremental adaptation a, is:

$$P = M(1-m)a^n \quad ,$$

where P and M are both referenced to the static threshold. Instantaneously, or with no delay (n ~ 0), the model predicts a short-term dynamic range below masker (M - P_0) equal to the vertical distance between the static I/O curve and threshold:

$$M - P_0 = M - M(1-m) = Mm \quad .$$

Therefore, the data points at the shortest delay (Figure 2.3.B) provide an approximation for the I/O slope parameter m. An iterative procedure was used to minimize the total MSE between the model predictions of the probe thresholds and the average forward masking data for all data points at each center frequency, as a function of the two model parameters m and a. The total MSE is relatively insensitive to the I/O slope, m, compared to the adaptation parameter, a. Therefore, the initial estimate of m from the short-delay conditions was averaged with the value that minimizes total MSE, to determine a final m estimate. A second MSE minimization as a function of only a, determined the final a estimate.

Geometries necessary to derive attack (downward adaptation) parameters are described in Figure 2.6. Before the onset of the masker, the model reaches the static threshold (at point A in Figure 2.6). At the abrupt masker onset, the output trajectory translates diagonally upward (from A to B) and then slowly drops toward the I/O target as the model adapts (from B to C to D). If the duration of the masker is short relative to the downward time constant, the trajectory will not reach the I/O target by the time of the abrupt masker offset (point C). In response to the masker offset, the output trajectory corresponding to the short masker moves diagonally (from point C), crossing the internal threshold at a lower point than the trajectory corresponding to the longer masker (from point D). After brief recovery during a short probe delay, the model predicts less forward masking from the short-duration masker.

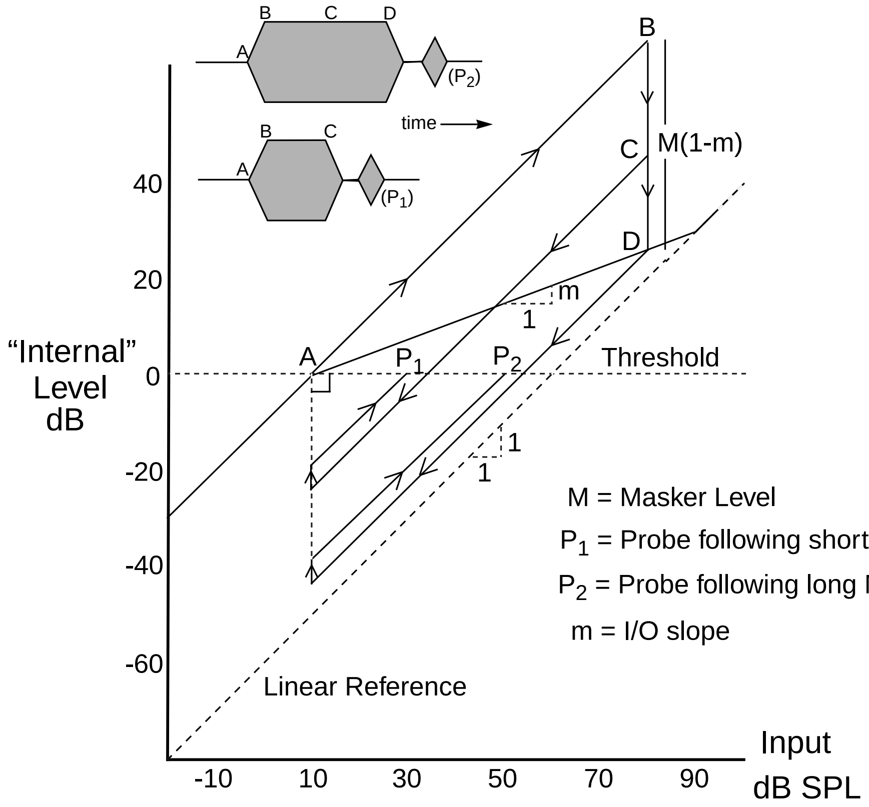


Figure 2.6 Geometry to derive attack (downward adaptation) parameters from forward masking thresholds as a function of masker duration.

Following incomplete downward adaptation (or attack), and as a function of the attack parameter *b*, discrete-time masker duration *nd* and probe delay *nu*, the model predicts a probe threshold of:

$$P = M(1-m)(1-b^{nd})a^{nu} \quad .$$

The probe threshold difference, Δ*P*, between short and long masker durations is:

$$\Delta P = M(1-m)b^{nd}a^{nu} \quad .$$

The probe threshold difference equation was solved for the model parameter *b*, and then its value was estimated from the differences reported in [Kidd and Feth 1982], using the *m* and *a* parameters derived above.

Two versions of the dynamic model were implemented: a full-rate system and a down-sampled version. The full-rate system uses rounded exponential filter shapes [Glasberg and Moore 1990], and then adapts the envelope of each filter output at the full sampling rate. The down-sampled system obtains Mel-scale power spectrum estimations every 10 ms by weighting and adding power spectrum points from an FFT, and then adapts these outputs at the down-sampled rate. On an HP715 workstation, the down-sampled system runs at 0.43 times real time, while the full-rate implementation requires 9.4 times real time. All evaluations included here use the down-sampled implementation.

Table 2.1 summarizes the model parameters and adaptation time constants across frequencies. The *a* and *b* terms are with respect to a 100 Hz spectral sampling rate (or frame rate). Adaptation stages with center frequencies between measured points use a weighted average of neighboring parameters. Attack time constants are approximately 3-4 times shorter than release time constants. These times, and more accurately their ratio, approximate those derived from physiological data [Goldhor 1985].

Table 2.1 Adaptation Parameters

Freq. Hz	Slope *m*	*a*	*b*	release (ms)	attack (ms)
250	0.19	0.864	0.474	68	13
500	0.20	0.854	0.510	63	15
1000	0.26	0.816	0.543	49	16
2000	0.29	0.851	0.525	62	16
4000	0.34	0.858	0.507	65	15

Figure 2.7 shows the model’s prediction of the decay of masking at 1 kHz. Note that the decay rate of forward masking is greater with more intense maskers,

and that the decay is nearly linear with logarithmic time.

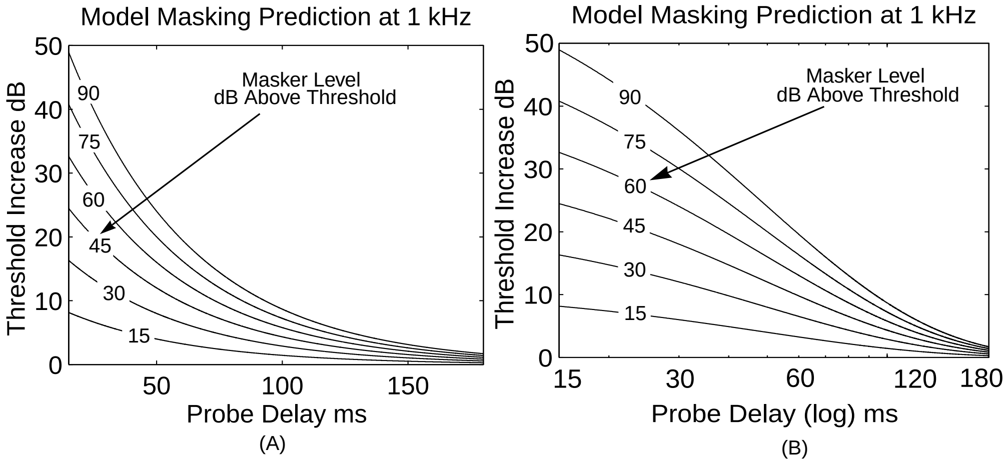


Figure 2.7 The model’s prediction of the decay of forward masking as a function of masker level at 1 kHz: A) with a linear time reference and B) with a logarithmic time reference.

Figure 2.8 shows two examples of the model’s behavior at 1 kHz. Figure 2.8.A shows the response to two consecutive pulses. The model adapts in response to the onset of the first pulse, and the response to the onset of the second pulse rides on top of the partial recovery from adaptation. Figure 2.8.B shows forward-masking examples. The model starts adapting at the onset of the long pulse, and then recovers after its offset. Lower-intensity impulses following the long pulse, corresponding to potential probe onset points, again ride on top of the model’s recovery from adaptation to the pulse. The responses to the impulses are initially below threshold (masked) and with time, rise above threshold.

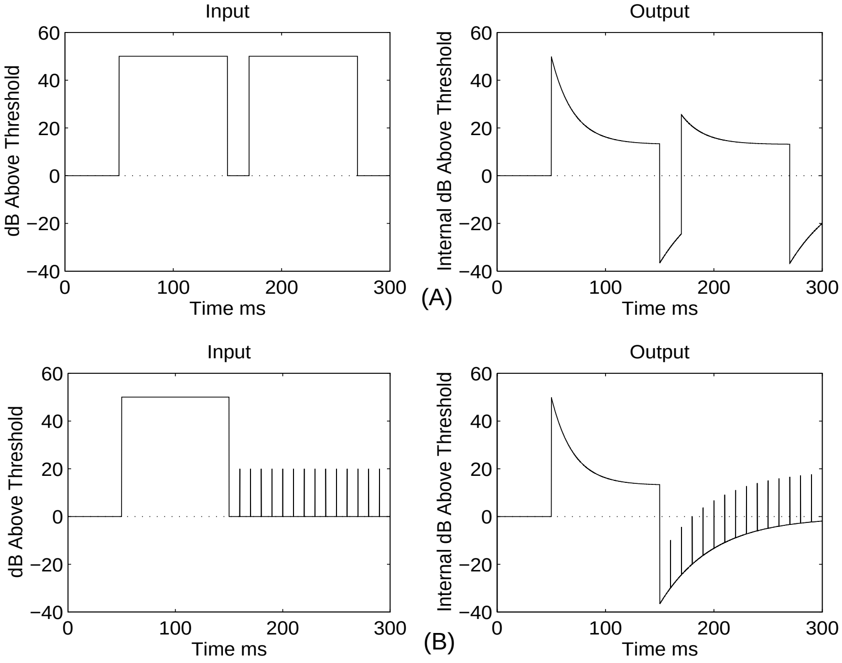


Figure 2.8 Adaptation to, and recovery after, a pulse: (A) The response to the second pulse is diminished; and (B) Impulses, corresponding to onsets, are initially masked (similar to figures in [Goldhor 1985]).

Figure 2.2 includes the model's fit to the average forward-masking data. The computational model approximates forward masking data for a wide range of masker levels and probe delays across several frequencies. The standard deviation of the error is: 2.7, 2.9, 3.2, 3.1, and 2.4 dB, at 250, 500, 1k, 2k, and 4k Hz, respectively. Most notably, however, the model consistently underestimates forward masking at the shortest probe delays. At least two factors contribute to this error.

The exponential derivation assumes the 15 ms delay between the masker and probe is silence. This assumption provides the maximum possible distance to target during the 15 ms, the maximum amount of recovery, and the lowest prediction of forward masking. In fact, the stimuli had 5 ms of offset, 5 ms of silence, and 5 ms of onset during this interval. Any non-silence during the 15 ms delay decreases the distance to target, reduces the amount of recovery, and increases the estimation of forward masking. Ignoring the finite onsets and offsets reduces the model's predictions of the amount of forward masking at short delays.

In this derivation, forward masking is assumed to occur when insufficient auditory recovery keeps the response to the probe below threshold. However, at shorter (near zero) delays, with extremely similar maskers and probes, the probe may only be audible as a change in level at the end of the masker [Moore and Glasberg 1983], and not as a separate event. Even though the response to the probe is above threshold, the subject may not distinguish the probe from the masker, and therefore not detect the probe. Because the derivation requires the model's response to the probe to be below threshold to be masked, it underestimates the amount of forward masking especially at short delays with intense maskers.

2.3.3 Predicting Other Data

Figure 2.9 shows the model's predictions of previous forward masking data. Figure 2.9.A shows the model's prediction of average data with wide-band stimuli [Plomp 1964]. These data provide relatively complete measurements of forward masking across level and delay. In the results shown in Figure 2.2, there is only slight variation of forward masking with frequency. Because the adapting response of the model to wide-band stimuli approximates the response at middle frequencies, the wide-band data were predicted using the model parameters derived from the 1kHz data. Although the model underestimates these data, the trends are consistent.

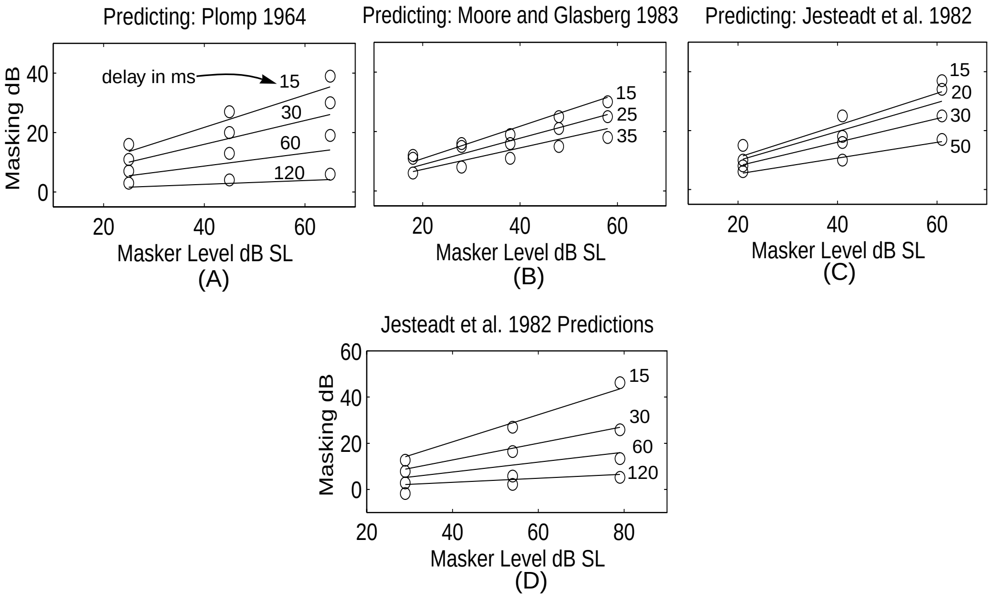


Figure 2.9 Using the model to predict other forward masking data: (A) wide-band masker and probe [Plomp 1964]; (B) wide-band masker, sinusoidal probe at 1kHz [Moore and Glasberg 1983]; (C) sinusoidal masker and probe at 1kHz [Jesteadt et al. 1982]. (D) The equation provided in [Jesteadt et al. 1982] predicting the present data.

Figure 2.9.B and C show the predictions for wide-band and pure-tone maskers of 1 kHz pure tones, respectively [Moore and Glasberg 1983, Jesteadt et al. 1982]. These measurements were made only at relatively short delays. Authors have historically disagreed on how to specify delay in a forward masking experiment

[Plomp 1964]. Here delay is measured between the envelope peaks, while [Jesteadt et al. 1982] used zero-voltage points, and [Glasberg and Moore 1983] chose half-voltage points between the masker and probe offset. The present study used 5 ms ramps, [Glasberg and Moore 1983] used 10 ms, and [Jesteadt et al. 1982] used 5 ms for the masker and 10 ms for the probe. To compensate for these differences 2.5 ms is subtracted from the delay reported in [Jesteadt et al. 1982], and 10 ms is added to the numbers in [Glasberg and Moore 1983]. The masker level in the 1kHz band for the wide-band masker is determined by the energy in the critical band [Zwicker and Terhardt 1980] centered at 1 kHz. Although comparisons are only possible at relatively short delays, the model overestimates the amount of masking by wide-band noises, and underestimates masking by pure tones. Once parameterized, however, the simple dynamic mechanism approximates dynamic psychophysical responses.

Figure 2.9.D shows the prediction of data from this study by an equation proposed in [Jesteadt et al. 1982]:

$$P = a(b - \log\Delta t)(M - c) \quad .$$

P and *M* are the levels of the probe and masker above threshold, and the constants *a*, *b*, and *c* are chosen to fit the average forward-masking data at 1kHz in [Jesteadt et al. 1982]. Even though the parameters in this equation were chosen from a data set that did not include measurements at the longer delays used in this study, it provides an excellent prediction of the present data.

2.3.4 Other Models Predicting Forward Masking

Other auditory models have been derived which, in general, provide a better fit to forward-masking data. Most, however, do not readily extend to a general processing scheme suitable for an ASR front end. For the dynamic mechanism derived in this paper, a signal is masked when the response is below threshold. To fit forward-masking data, other models typically parametrize a decision device, and thereby impose explicit interpretations of the front end's response. If the parameterized decision device is removed to use the auditory model for an ASR front end, it is less clear how the recognition system would correctly interpret a masked signal.

Forward, backward, and forward/backward masking combinations have been predicted with great precision assuming a relatively standard model of filtering, rectification, power-law compression, temporal integration and a decision device [Oxenham and Moore 1994]. In its original derivation, however, there was no mechanism to account for the level-dependence of forward masking. Either the temporal window shape [Oxenham and Moore 1994], or the power-law compression [Oxenham and Plack 1996] may vary with level. The decision device required an unusually high minimum detectable temporal amplitude variation of 6 dB, which may not extend well to a general processing scheme. Finally, if forward masking is entirely a consequence of temporal integration, physiological measurements of adaptation are ignored, and there is no mechanism which explains physiological and perceptual sensitivity to onsets and transitions.

Other researchers have proposed models using adaptation mechanisms to explain forward masking [Shannon 1990, Dau and Pueschel 1996a-b]. The first of these [Shannon 1990] uses a modified version of a previous model [Zwislocki 1969] which includes filtering, envelope detection, power-law compression, rapid and short-term adaptation, and long-term integration. The long-term integrator is bypassed in forward-masking tasks. Immediately following a stimulus, the model assumes that there is no rapid onset component in response to a probe, that this component recovers exponentially with time, and that the relative level of this component is used to determine forward masking. The model is somewhere between a complete processing mechanism and an equation summarizing psychophysical responses, and therefore, is also difficult to incorporate into ASR systems. The exponential recovery of the rapid onset component has similarities to the exponential adaptation used in the dynamic mechanism described in this paper.

Dau and Pueschel [1996a-b] have also developed a general auditory model which together with an ‘optimal decision device’ predicts well a wide variety of psychophysical data. In each channel, the model uses linear filtering, half-wave rectification, and low-pass filtering, followed by five adaptation stages and a low-pass filter. The output is correlated with templates that store the model’s response to other (masker-only) conditions to predict masking thresholds, imposing a relatively complex post-processing mechanism to fit the data. The model provides a dynamic spectral representation of speech which is likely to improve recognition robustness; potential application improvements may warrant the significant computational complexity.

2.4 Peak Isolation

Both speech perception and the response of individual auditory nerves are extremely sensitive to the frequency position of local spectral peaks. There are several mechanisms, and corresponding modeling approaches, which may explain this sensitivity. Physiologically motivated by the local fan-out of the neural connections to outer hair cells, [Lyon 1982] suggests cross-coupling AGC stages to improve static spectral contrast, providing functionality similar to the higher-level lateral inhibitory network in [Wang and Shamma 1994]. Significant effort [Lyon 1984, Seneff 1988, Ghitza 1991] also focuses on modeling how the auditory system derives, and makes use of, redundant temporal micro-structure. Auditory nerves with center frequencies as far as an octave away from a local spectral peak can synchronize their response to the frequency of the peak, providing a composite neural representation dominated by that frequency [Delgutte and Kiang 1984]. Similarly, perceptual discrimination of vowels is more sensitive to the frequency location of spectral peaks than to other aspects of the spectral shape [Klatt 1982]. These data suggest that the auditory system may derive a noise-robust representation by attending to the frequency locations of local spectral peaks.

The dynamic model was also evaluated with a novel processing technique, based on raised-sin cepstral liftering [Juang et al. 1987] together with explicit peak normalization, which isolates local spectral peaks. Raised-sine cepstral liftering is weighting the cepstral vector by the first half-period of a raised-sine function.

The cepstral vector is an expansion of the even log spectrum in terms of cosine basis functions. The c_0 term specifies the log-spectrum average, the c_1 term approximates the log-spectrum tilt, etc., and high cepstral terms represent quickly-varying ripples across the log spectrum. Weighting the cepstral vector specifies the relative emphasis of different types of log-spectrum variations. A raised-sine lifter de-emphasizes slow changes with frequency associated with overall level and vocal driving-function variations, as well as fast changes which may reflect numerical artifacts [Juang et al. 1987].

It is helpful to view the effects of cepstral liftering in the log spectral domain. Figure 2.10.a starts with the log spectrum, from a vowel [i], implied by a truncated cepstral vector. Figure 2.10.b shows the log spectrum implied after raised-sine cepstral liftering. The average level as well as slow (and fast) variations with frequency are de-emphasized, leaving components that change with frequency. This process emphasizes both spectral peaks and valleys.

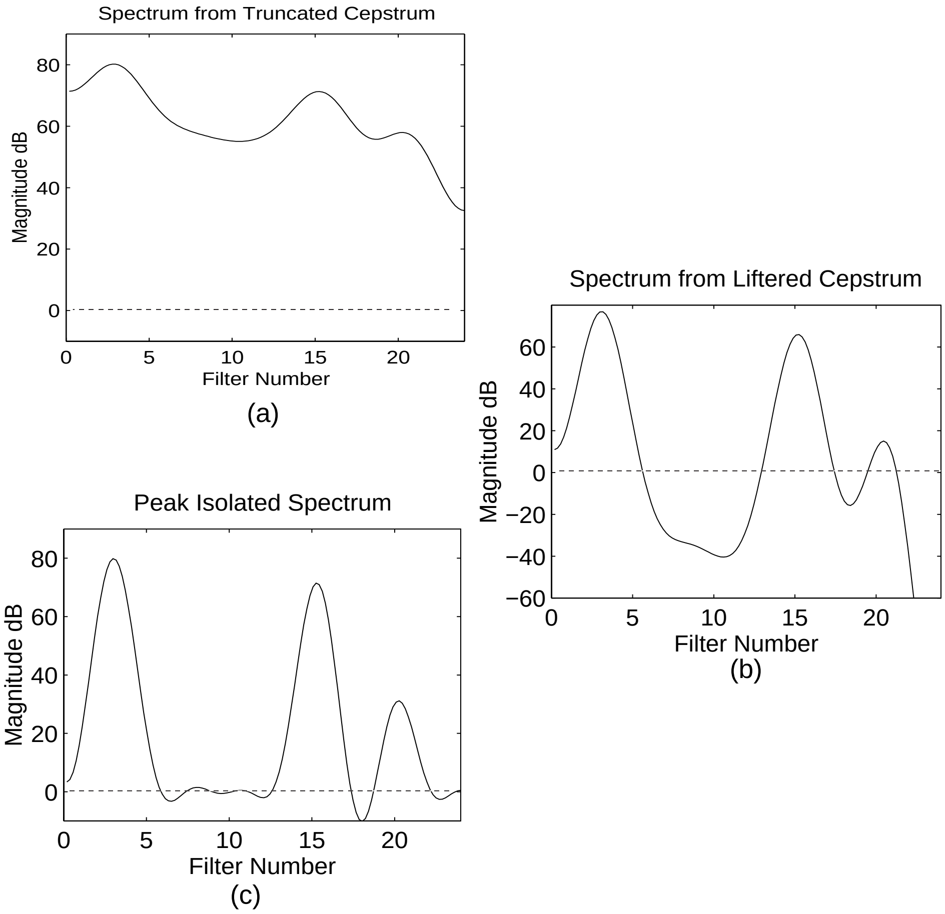


Figure 2.10 Peak isolation processing: log spectrum of the vowel [i] after (a) cepstral truncation; (b) raised-sine cepstral liftering; and (c) half-wave rectification and peak normalization.

The valleys are removed by half-wave rectifying the log spectral estimate implied after raised-sine liftering, and a final vector is obtained by transforming back to the cepstral domain. Because the half-wave rectifier is non-linear, explicit transformation from cepstrum to log spectrum (processing through the rectifier) and then transformation back to cepstrum are required. The raised-sin lifter also affects the magnitude of the peaks. Therefore, before transforming back to the cepstrum, peaks are scaled to the level measured in the original log spectrum. The final peak-isolated estimation is shown in Figure 2.10.c.

2.5 Conclusion

Two mechanisms are described to modify the sequences of spectral representation used for speech recognition. The first is a (post-logarithm) exponential adaptation mechanism which is parameterized to approximate forward masking data. Adaptation leads to relatively stronger responses to onsets and spectral transitions, improving spectrotemporal contrast across time. The second, based on cepstral liftering, isolates local spectral peaks and improves spectrotemporal contrast across frequency.

Figure 2.11 shows spectrogram representations for the digits “nine six one three” at two signal to noise ratios and for three different processing strategies. The first representation uses Mel-frequency cepstral coefficients (MFCC, described in Chapter 1), the second adds adaptation, and the third includes peak isolation. Recognition evaluations described in Chapter 5 show increased recognition performance in noise using these representations.

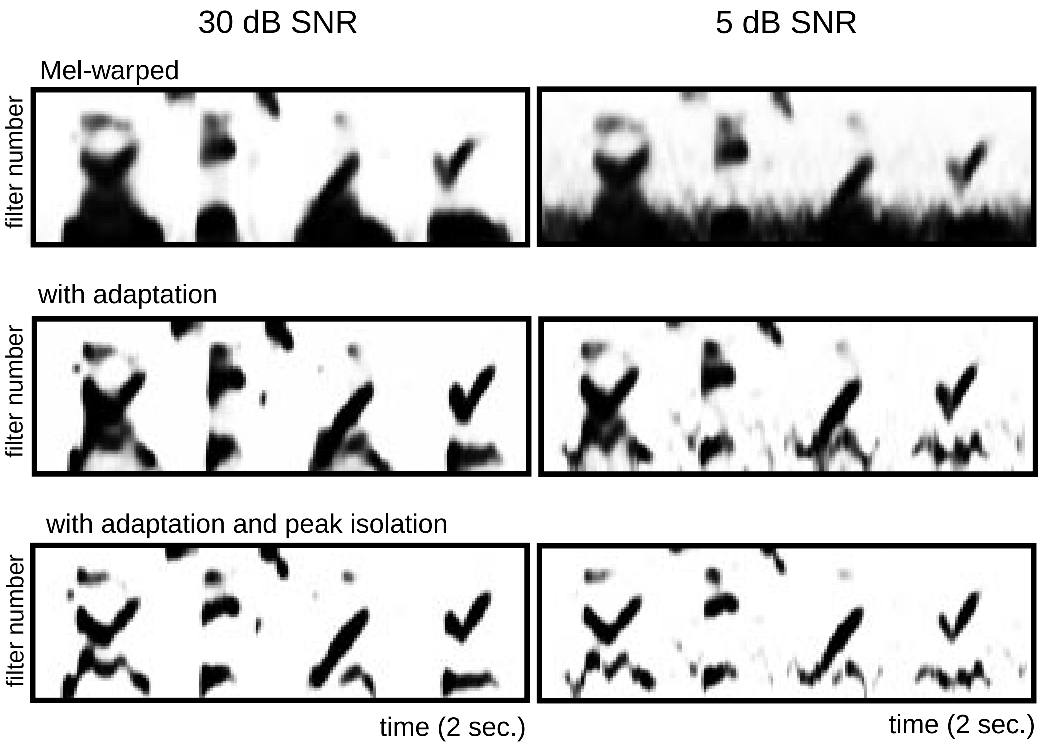


Figure 2.11 Spectrogram representations for three processing strategies: left column is at 30 dB SNR, right is 5 dB SNR; top spectrograms use MFCC, middle includes adaptation, and bottom also includes peak isolation.

Chapter 3

Parameterizing the Position and Motion of Local Spectral Peaks

3.1 Background and Motivation

The eigenfunctions of a resonating vocal tract are manifested acoustically as formants in speech. The analysis of formants has provided significant insights into speech production mechanisms, and motivation for speech coding algorithms. Referring to ASR in 1981, D. Klatt wrote [Klatt 1981]:

> "These schemes will succeed only to the extent that metrics can be found that are (1) sensitive to phonetically relevant spectral differences such as those caused by formant frequency changes, and (2) relatively insensitive to phonetically irrelevant spectral differences associated with a change of speaker identity or recording

conditions.”

Although various compensation schemes for changing acoustic environments are often used, the predominant characterization of speech for statistical speech recognition is based on sequences of short-time (10-20 ms) spectral estimations, which characterize the coarse spectral envelope of each successive frame [Rabiner and Juang 1993]. This representation is only an *implicit* characterization of the formant structure of speech, and as such, does not provide direct access to the phonetically relevant formant motion described above. Explicit characterizations of speech dynamics typically focus on the motion of the cepstral representation of the short-time spectral estimates [e.g. Deng 1994, Deng et al. 1994], and thereby parameterize changes in the ‘complete’ spectral shape and not the specific (potentially robust) formant motion.

More direct formant tracking usually involves first identifying local spectral peaks in a sequence of spectral estimations [Schafer and Rabiner 1970, McCandless 1974]. Alternatively, Teager energy operators [Maragos et al. 1993, Foote et al. 1993, Hanson et al. 1994, Potamianos and Maragos 1996], Hilbert Transforms and Wigner Distributions [Rao 1996], as well as changes in the cross-correlation of the temporal fine-structure between neighboring auditory frequency channels [Deng and Kheirallah 1993] have been used to identify formant frequencies in speech. Formant tracks are then pieced together using heuristics [Schafer and Rabiner 1970, McCandless 1974], hidden Markov models [Kopec 1986], or the minimization of a cost function [Larpie and Berger 1994]. The two-stage process has also been collapsed to one using extended Kalman filters [Rigoll 1986, Niranjan et al. 1994]. Unfortunately, formant tracking systems are often non-robust; they are only occasionally evaluated in noise, and are almost never tested in the context of an ASR task.

The processing schemes described in Chapter 2 enhance the representation of spectral dynamics and more specifically changing spectral peaks. While such sensitivity may be phonetically relevant, the characterization of the formant motion is still implicit. Formant motion is only weakly characterized by the temporal derivative of the overall spectral estimate, and by the sequence of underlying states in the statistical model. Neither of these is a direct characterization, and neither provides an obvious means to exploit the dominant frame-to-frame correlations of local spectral peaks. Finally, context dependent spectral representations may, in general, be poor matches to ASR algorithms which rely on the characterization of segmentally stationary statistics. A more direct parameterization of the motion of spectral peaks, on the other hand, may prove to be a better match.

In essence, the algorithm described here introduces a simple and robust form of formant tracking, and augments the frame-based feature vector used for ASR with an explicit parameterization of the formant position and motion.

3.2 The Algorithm

The algorithm described in this chapter builds on that of Chapter 2. A block diagram of the processing stages in the algorithm is shown in Figure 3.1. While the initial filtering and subsequent liftering are processed at the sampling rate (11025 samples/sec), the remaining processing occurs at the down-sampled frame rate (100 frames/sec) and has a lower order of computational complexity.

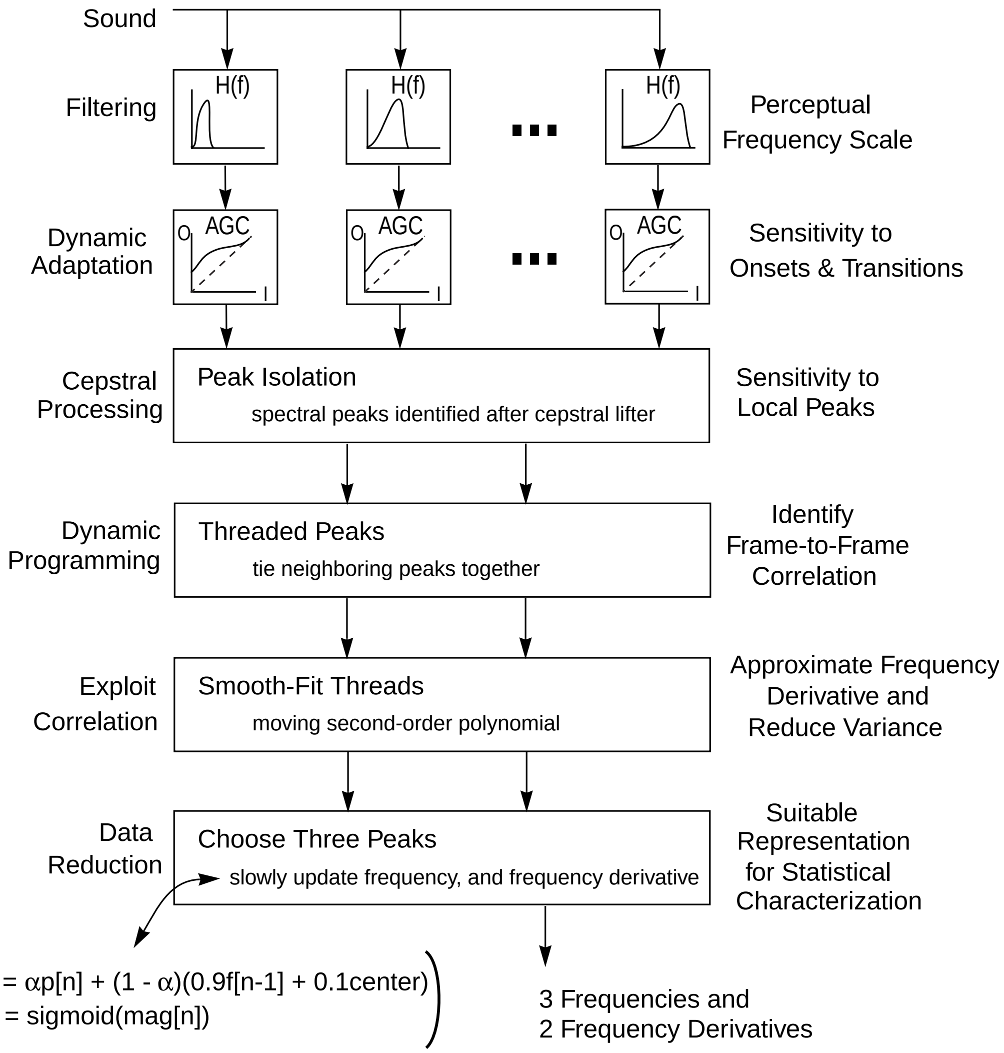


Figure 3.1 Overview of processing stages.

3.2.1 Filtering and Adaptation

The filtering stage, after [Davis and Mermelstein 1980], is implemented by integrating power spectrum estimates weighted by triangular filters that have bandwidths of 100 Hz for center frequencies below 1 kHz, and bandwidths of 0.1 times the center frequency above 1 kHz. The resulting frequency resolution is therefore linear below 1 kHz, and logarithmic above 1 kHz.

The adaptation stage for each frequency channel acts as an automatic gain control which incrementally adjusts an additive logarithmic offset to reduce the distance to a target input/output point. Adaptation emphasizes onsets and represents changing spectral peaks more strongly than static ones. Together, these two stages significantly affect how spectral peaks are identified and processed in subsequent stages. Figure 3.2.A includes a spectrogram of four digits, “nine six one three,” at 10 dB SNR, after filtering and adaptation.

3.2.2 Peak Isolation

Local spectral peaks are first identified independently in each frame by finding the local maxima in the log-spectral estimate, after raised-sin cepstral liftering [Juang et al. 1987]. For each peak, the frequency position and log magnitude are stored. Because the raised-sin cepstral lifter alters the level of the local spectral peak, the log-magnitude value is taken from the corresponding frequency position in the spectral estimate before raised-sin liftering.

In Figure 3.2.A, note the relatively strong temporal correlation between the frequency positions of the local spectral peaks through formants and formant transitions.

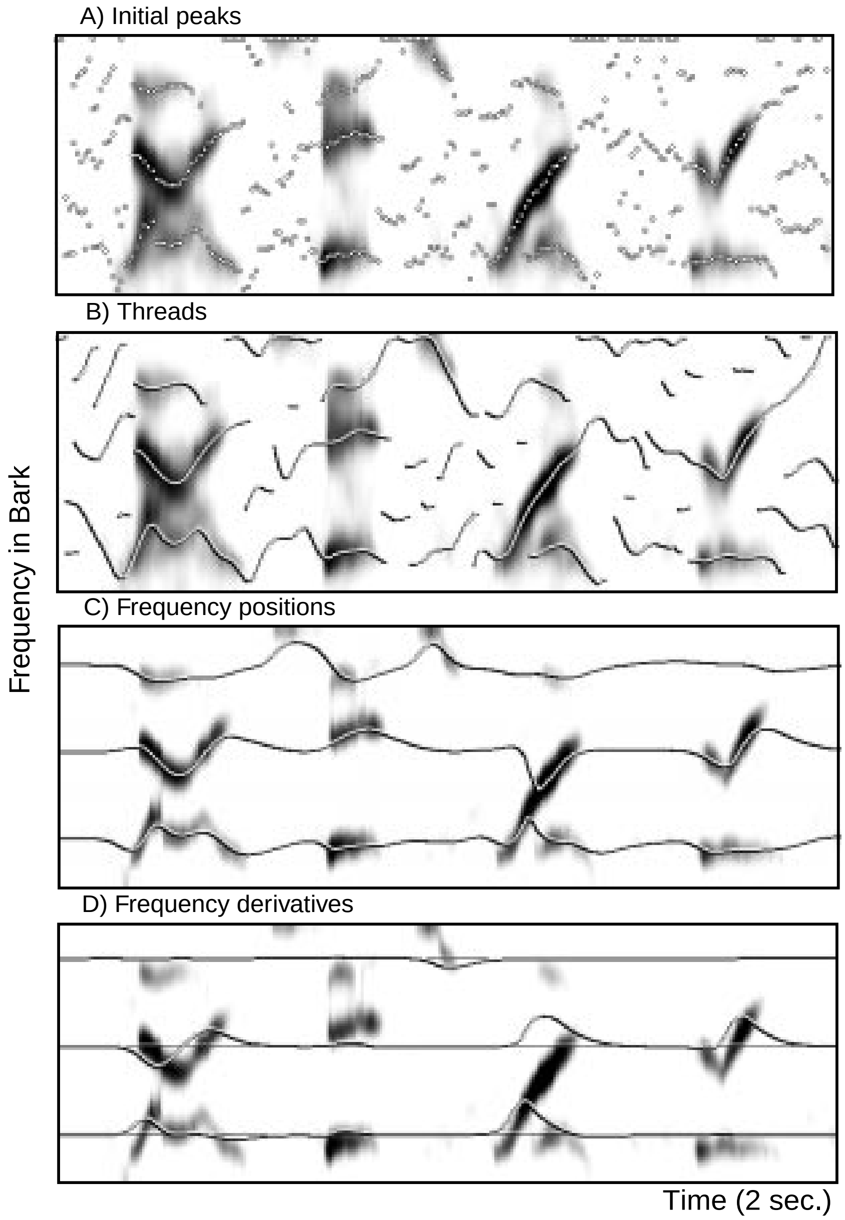


Figure 3.2 Peak positions and motion: A) Initial peaks identified after cepstral liftering; B) Neighboring peaks grouped to threads; C) Tracking three frequency positions; and D) Three frequency derivatives.

3.2.3 Threading Peaks

This is the first of two stages which group peaks based on their spectro-temporal proximity. The task is to connect the spectral peaks together in time into *threads*, and the approach used here is a form of dynamic programming. Each peak (in each frame) is connected to the closest thread that extends into at least one of the last two frames. If the frequency distance to the closest thread is greater than approximately 10% of the total (warped) frequency range, then a new thread is started. If no peak connects to the end of a given thread for two successive frames, then that thread is ended. Figure 3.2.B shows a moving seven-point (70 ms) second-order polynomial fit to each resulting thread. For each thread that includes at least four peaks, the temporal derivative as implied by the moving second-order polynomial is also stored.

3.2.4 Choosing Three Peaks

The second stage imposes a structure on the threads enabling a more systematic characterization, and also attempts to reduce their variance. Threads from the first stage start and end somewhat randomly, which makes storing them for analysis or comparison not obvious. Also, there is significant variance in the reliability of the thread measurements. That is, dominant formant transitions are tracked more reliably than small peaks in background noise.

The second stage limits the representation of the threads to three peaks in frequency for each frame. Three new *tracks*, centered at relatively low, medium, and high frequencies, are used to represent the information from the threads. The log magnitude of the original spectral peak is used when integrating frequency positions and derivatives from the corresponding thread. This introduces an inertial response that updates more quickly to information from more dominant peaks.

In the implementation, each track is assigned a center frequency, or DC offset. The three center frequencies are equally spaced on the warped frequency scale. At each frame the frequency position of the track incrementally adjusts toward the closest thread in that frame. The increment of adjustment is a sigmoidal function of the magnitude of the thread. The equation that describes this adjustment is:

$$f[n] = \alpha\, p[n] + (1-\alpha)(0.9f[n-1] + 0.1f_0),$$

where n is the frame index, $f[n]$ is the frequency of the track, $p[n]$ is the frequency of the nearest peak, f_0 is the center frequency or DC offset, and the variable α, which controls the rate of the increment, is a sigmoidal function of the log magnitude of the peak. Ignoring the DC offset, the equation describes a non-constant coefficient first-order low-pass filter. The sigmoid maps log magnitude to the appropriate (0,1) interval, so that the filter changes from low-pass to all-pass. Because the log magnitude of the peak is measured after the adaptation stages, transitions and onsets, in general, incur the most abrupt track changes.

An identical structure is used to track the frequency derivatives of the threads. For each of the three tracks, the current frequency derivative estimate is incrementally updated to the derivative measured at the closest peak. The size of the increment is a sigmoidal function of the log magnitude of the peak. A final (fixed) low-pass filter with a -3dB point of just over 15 Hz is applied both to the three frequency tracks, and to the three derivatives. Figure 3.2.C shows the final frequency positions for the three tracks, and Figure 3.2.D shows the frequency derivatives.

3.3 Discussion

This parameterization of the motion of local spectral peaks differs from more traditional formant tracking [e.g. Schafer and Rabiner 1970, McCandless 1974] in several ways. The initial filtering and adaptation greatly influence the resulting spectrotemporal representation. The frequency resolution is warped to a perceptual scale, and signal dynamics play a significant role in determining which peaks are identified. The two-stage process to identify the final tracks is aimed at identifying the robust, slowly-varying information which is likely to be highly correlated with underlying articulator motion. The tracking process also includes an inertial component dependent on the magnitude of the (adapting) response of the peak. Initial frequency derivative estimates are calculated before the imposition of explicit frequency ranges, reducing the influence of artifacts from these heuristics on the derivative estimates. Finally, by limiting the representation to three peaks with centers equally-spaced on the warped frequency scale, some of the complications introduced by the merging and splitting of higher formants are avoided. A simplified task may lead to a more robust system.

3.4 Conclusions

This chapter describes a processing scheme which attempts to parameterize the phonetically relevant information represented in formant position and motion. In essence, the threading described represents a grouping of spectrotemporal patterns as an early stage of auditory scene analysis. Instead of extracting information directly from individual spectral estimates, the current approach imposes a structure which extracts information from the spectrotemporal relationships between dominant spectral peaks. Chapter 6 includes further discussion of how this type of early scene analysis may help explain other current challenges for psychoacoustic modeling of non-stationary sounds.

Detailed descriptions of recognition evaluations using the final frequency positions and derivatives are presented in Chapter 5. When augmenting the processing in Chapter 2, the current representation eliminates roughly 30% of the previous recognition errors.

Chapter 4

Modeling the Perception of Amplitude Modulation

Currently, most ASR systems integrate spectral estimates over multiple pitch periods and remove explicit pitch and voicing information. However, amplitude modulation cues in voiced speech provide a robust and salient pitch perception which may be instrumental for recognizing speech in noise. In this chapter, three psychoacoustic models are used to predict the temporal modulation transfer function (TMTF) [Viemeister 1979] and the detection of voicing for high-pass filtered natural fricatives in noise. Models using an envelope statistic and modulation filtering predict the TMTF data, while predictions from a model using a summary autocorrelogram approximate both data sets.

4.1 Motivation

During voiced speech, vibrating vocal folds excite time-varying resonances of the vocal tract. Given a sequence of feature vectors representing log-magnitude spectral estimates of vocal-tract transfer functions, most ASR systems use a hierarchy of non-stationary stochastic models operating at the progressively slower rates of the speech analysis frame (10-30 ms), phoneme, word, phrase, and even sentence, to determine what was most likely said [Rabiner and Juang 1993]. More relevant to the current study, these systems do not use pitch or voicing information.

Instead, the signal processing for feature vector extraction, usually reflects some form of deconvolution, attempting to isolate vocal-tract transfer-function estimates from the influences of the driving function. Linear prediction, for example, is used with a predictor polynomial that is significantly shorter than the expected glottal periodicity. Similarly, when homomorphic analysis is used for ASR, the high-quefrency cepstral terms, which can represent the periodic ripple across the spectral estimate resulting from a harmonic driving function, are ignored. Finally with the currently popular Mel-frequency cepstral coefficients (MFCC), the initial spectral estimate is first averaged (in time) over multiple pitch periods and then integrated across frequency to obtain a first-order approximation of auditory frequency selectivity. The output is then compressed by a logarithm, and finally the discrete cosine transform provides some decorrelation of the log-magnitude spectral estimate across frequency. Higher-order terms in the resulting cepstral vector are again ignored. Integrating across time and frequency reduces the variance of the spectral estimate, and together with the truncated cepstral vector, nearly eliminates any periodic source information.

Deconvolution is an important step for isolating the phonetic information about "*what was said*," from aspects of the prosodical information more concerned with "*how it was said*." But as the first processing stage of current systems, it is most likely eliminating large parts of the perceptually salient information that humans use to identify and recognize speech in naturally noisy environments.

Speech communication has evolved to be robust in noise. Redundancies are, therefore, ubiquitous. Perceiving speech in noise requires an intelligent use of the potentially unreliable, but redundant, multi-dimensional cues spread over wide-ranging time scales. While deconvolution must occur somewhere in the recognition process, blindly eliminating a potential wealth of redundant cues may not be appropriate for the first stage. More plainly, rigid blind deconvolution in the first stage is unlikely to be optimal.

4.1.1 Pitch Perception

Processing voicing information in speech requires analyzing the harmonic structure associated with a quasi-periodic vocal driving function, and might therefore be considered as an aspect of pitch perception.

In 1951, Licklider proposed a duplex theory to explain many aspects of pitch perception, including the perception of the missing fundamental (or residue pitch), and the pitch of modulated noise [Licklider 1951]. Briefly, Licklider envisioned neural machinery which measured the running temporal autocorrelation in each auditory frequency channel. Pitch perception correlated to the common periodicities measured across channels.

In 1984 Lyon was able to simulate an implementation of the duplex theory, labeling the graphic output a *correlogram* [Lyon 1984]. Since then, Meddis and colleagues [Meddis and Hewitt 1991a-b, Meddis and O'Mard 1997] have formalized the simulations and included a final stage that adds the running autocorrelations across each channel generating a *summary correlogram*. Cariani and Delgutte have also shown that similar processing of measured auditory nerve impulses is sufficient to predict many classical pitch perception phenomena [Cariani and Delgutte 1996a-b]. Finally, other researchers have replaced the autocorrelation function with different mechanisms that measure temporal intervals in each auditory channel [e.g. Patterson 1992, Ghitza 1991, de Cheveigne 1998].

In general (and as shown in Licklider's original sketches achieved without the aid of computer simulation), simulations using these models provide a graphical output that correlates well to pitch perception. The time lag for the peak in the summary correlogram is usually found to be the reciprocal of the frequency of the perceived pitch, and the height of the peak may correlate to qualitative pitch-salience, or pitch strength. With few exceptions however, the models are not used to predict psychoacoustic just-noticeable-differences (jnds) with general stimuli.

Together with the lack of a clearly identified physiological substrate for the implementation of the required timing measurements, this line of research remains somewhat ‘open-loop.’

4.1.2 Perception of Amplitude Modulation

Processing voicing information in speech might also be considered an aspect of amplitude modulation perception.

In 1979, Viemeister applied a linear systems approach to the detection of acoustic envelope fluctuations [Viemeister 1979]. His model was first fit to data describing the detection of sinusoidal amplitude modulation of wideband noise, and then used to predict the detection of other harmonic envelopes. Motivated by the close relationship between standard deviation and autocorrelation, Viemeister’s model used the standard deviation of a demodulated envelope as the statistic to predict human performance. Although this measure does not characterize the perceived pitch of the amplitude modulation (the standard deviation measures the magnitude and not the rate of envelope fluctuations), a more complicated simulation involving autocorrelation was not necessary to predict the detection data. More recently, this model has been extended to predict other amplitude modulation detection data [Strickland and Viemeister 1996, Strickland and Viemeister 1997].

In 1989, Houtgast measured modulation masking that suggested explicit neural modulation filtering [Houtgast 1989]. Narrow bandwidth noise modulators were found to mask the perception of sinusoidal modulators, in a manner similar to the spectral masking of tones by narrow-band noises. Modulation tuning has also been measured physiologically [e.g. Langner 1992]. However, other modulation masking experiments using sinusoids have been less conclusive [Strickland and Viemeister 1996, Bacon and Grantham 1989]. Nonetheless, a model of modulation filtering has been implemented and shown to correlate to many aspects of amplitude modulation perception [Dau et al. 1997a-b].

In essence, modulation filtering replaces the single low-pass filter in the envelope statistic model with a second bank of filters. The modulation filtering simulations here also include a better approximation of auditory filtering than the single band-pass filter used in the envelope statistic model.

Therefore, there are at least three modeling approaches which may be helpful for analyzing the periodic envelope fluctuations in voiced speech: autocorrelation or interval-based temporal processing, the measurement of an envelope statistic, and explicit modulation filtering. To choose between them, implementations of each were first fit to predict TMTF data, and then each was used in a case study to predict the discrimination of voicing for strident fricatives in noise.

4.2 Strident Fricative Case Study: [s] and [z]

Fricatives are generated by forcing air through a sufficiently narrow constriction in the vocal tract to generate a turbulent noise-like source. With voiced fricatives, the vocal folds also vibrate adding low-frequency energy at the first few harmonics of the fundamental frequency. The relative level of the first harmonic, compared to that of the adjacent vowel, has been shown to be a good indicator for voicing distinctions with fricatives [Stevens et al. 1992, Pirello et al. 1997].

Here, the strident fricatives [s z] with the vowels [a i u] were recorded as CV syllables from four talkers. Figure 4.1 compares average log-magnitude spectral estimates for [s] and [z]. The voiced [z] has low-frequency energy not present in [s].

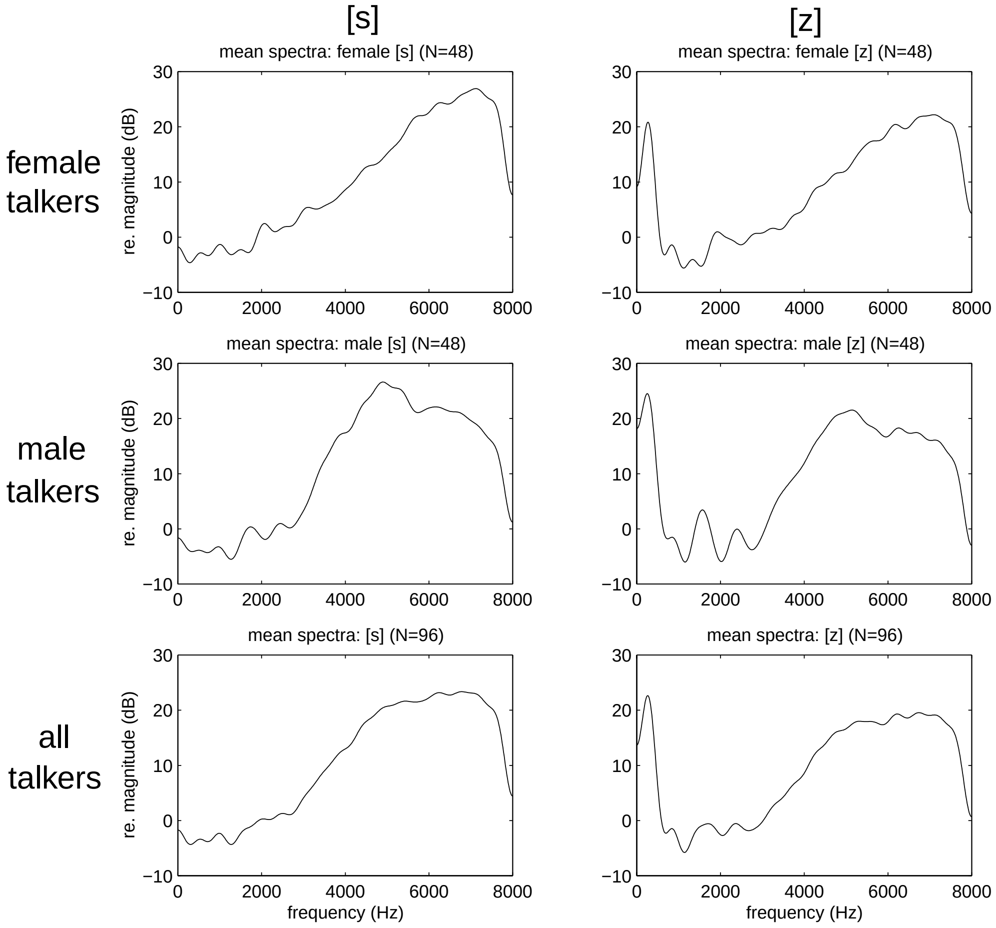


Figure 4.1 Comparing average spectral estimates for [s] and [z].

Current ASR systems would use the presence of low-frequency spectral energy to discriminate these sounds. However, there are situations where this particular difference can be obscured: e.g. a high-pass channel or a competing low-pass noise.

Figure 4.2 shows examples of the temporal waveform for [s] and [z], after each has been high-pass filtered above 3 kHz. Without low-frequency spectral components, the low-frequency pitch-rate information is represented in the envelope of the high-frequency noise-like carrier. These figures provide evidence that the vibrating vocal folds can modulate the pressure source that drives the turbulence for a voiced fricative. The modulated noise source leads to a potential redundant cue of voicing in a spectral region with significant speech energy. ASR systems that integrate spectral estimates over multiple glottal periods do not distinguish these sounds, while listeners, on the other hand, distinguish them at low signal to noise ratios (see Section 4.4).

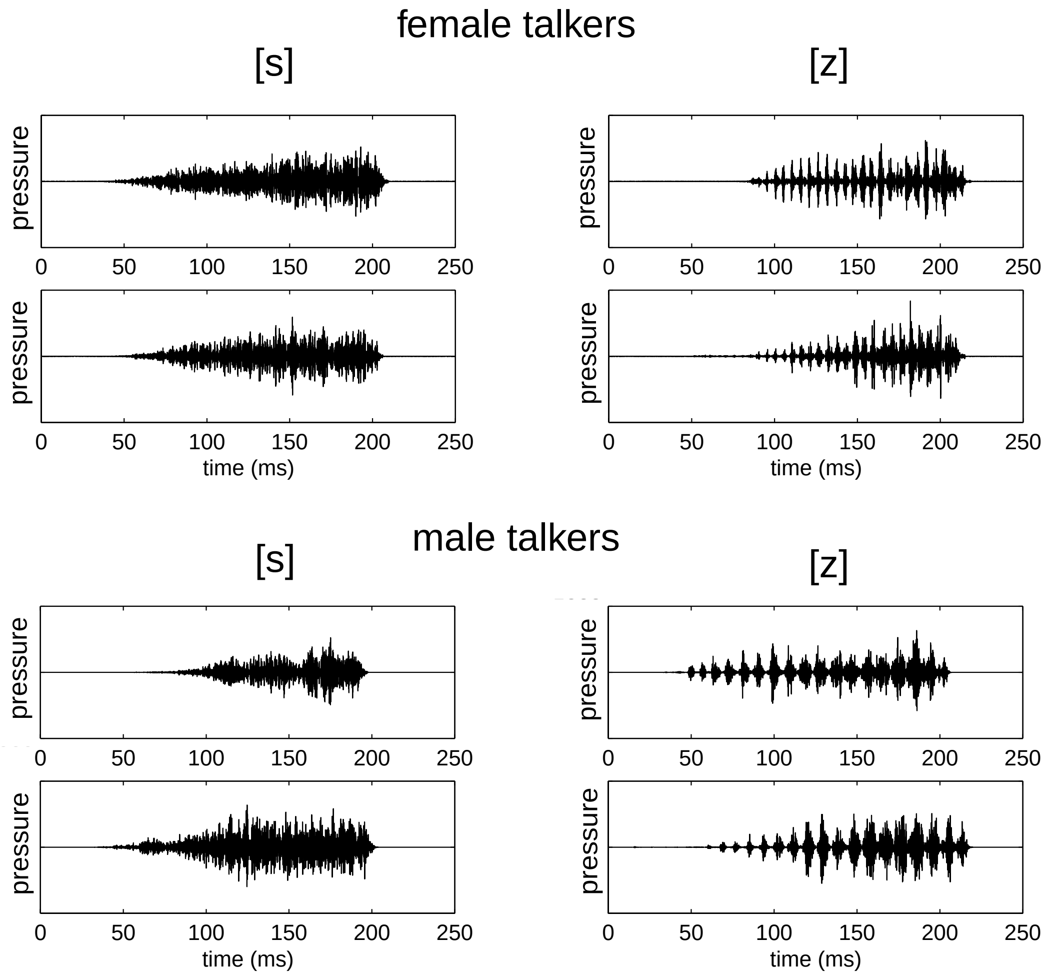


Figure 4.2 Examples of temporal waveforms after high-pass filtering.

4.2.1 Perceptual Measurements

To measure the perceptual sensitivity to this potential voicing cue, the discrimination of these sounds was measured in wide-band noise. The syllable initial fricatives were both temporally isolated from the adjacent vowel, and high-pass filtered above 3 kHz. During the perceptual tests, tokens were then centered in

1 second of spectrally flat noise.

Adaptive tests [Levitt 1971] were used to track the perceptual discrimination of the isolated fricative as a function of SNR at two *d'* levels. For each trial, the subject was required to identify a randomly chosen token as either [s] or [z]. Feedback was provided. The initial SNR was high enough that the fricatives were clearly distinguishable for all subjects. The SNR was increased after an incorrect response, and decreased after either 2, or 3, correct responses. A reversal is defined as a change in the direction of the SNR step. The SNR step size started at 4 dB, and was reduced to 2 dB after the first reversal, and to 1 dB after the third. The average of the SNR at the next 6 reversals provided an initial threshold estimate. If the variance in this estimate was less than 2 dB, the measurements stopped, otherwise the experiment continued for up to 6 more reversals. The average of three such measurements provided a final threshold estimate for each subject. When 2 (or 3) correct responses are required, the threshold estimate converges to a 70.7% (or 79.4%) correct response rate. For this experiment, these correspond to *d'* values of 1.09, and 1.64, respectively. Four audiometrically normal subjects participated in the experiment. Average thresholds across these four subjects are shown together with model predictions in Figure 4.10 below.

4.3 3 AM-Detection Mechanisms

The task in this experiment may require detecting periodic envelope fluctuations which become increasingly weak with more additive noise. Perhaps the most direct approach is to model this perception using an envelope statistic.

4.3.1 Envelope Statistic

Figure 4.3 shows a block diagram of the signal processing in an envelope statistic model. This classical approach reduces auditory processing to: auditory filtering (approximately measured along the basilar membrane), half-wave rectification (approximated in inner hair cell transduction) and low-pass filtering (measured throughout higher-levels of auditory processing). From an engineering perspective, the band-pass filter selects a channel, the half-wave rectifier is the non-linearity that modulates the carrier down to DC, and the low-pass filter tracks the envelope.

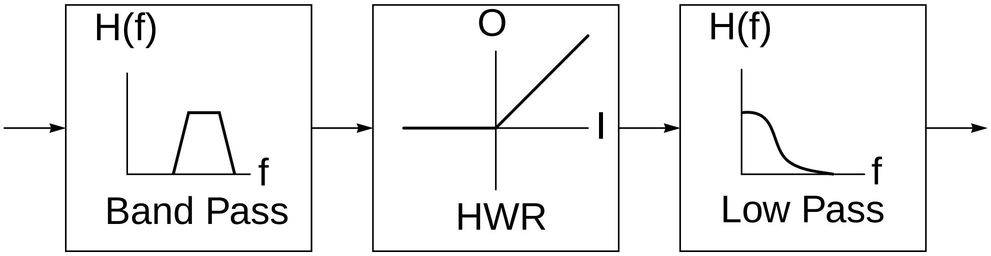


Figure 4.3 Envelope detection.

The model’s sensitivity to amplitude-modulated wideband noise increases with increasing bandwidth in the initial filter, while the reduction of sensitivity with increasing envelope frequency is mostly determined by the final low-pass filter.

4.3.2 Modulation Filtering

A schematic overview of an implementation of modulation filtering is shown in Figure 4.4. Building from the envelope detection processing above, the model includes multiple 4th-order gamma-tone filters [Patterson et al. 1992] which provide a better approximation of auditory filtering, and replaces the single low-pass filter with a second filterbank that analyzes the envelope spectrum.

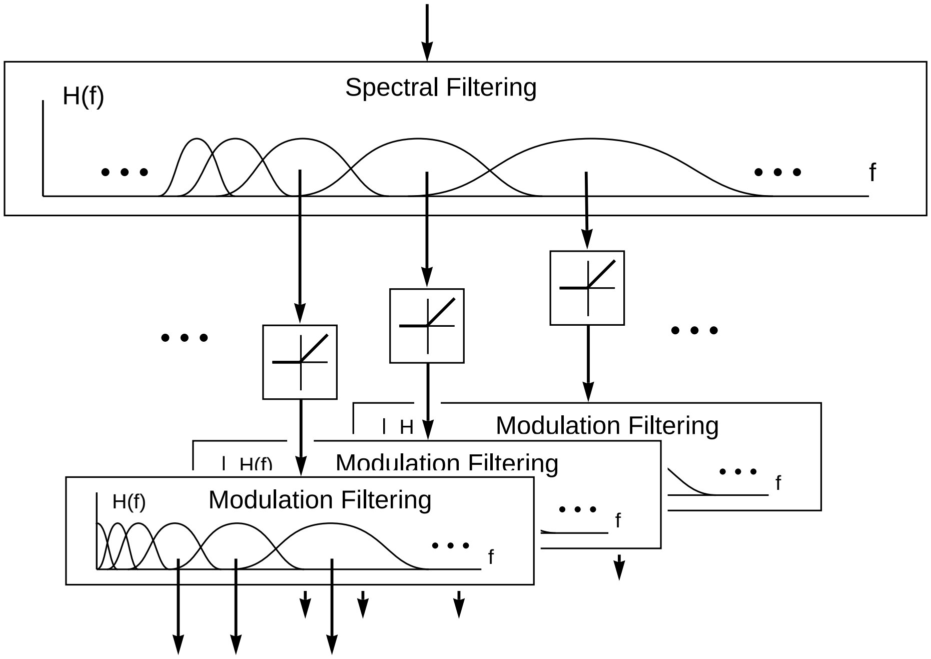


Figure 4.4   A modulation filtering scheme.

The frequency response for the modulation filters used (Q_{3dB} of 2, and -12 dB DC gain) were from [Dau et al. 1997a]. For each filter, our implementation used a second-order pole and a first-order (real) zero at DC. The distance of the zero to the unit circle was set to meet the DC specification. The resulting frequency responses are shown in Figure 4.5.

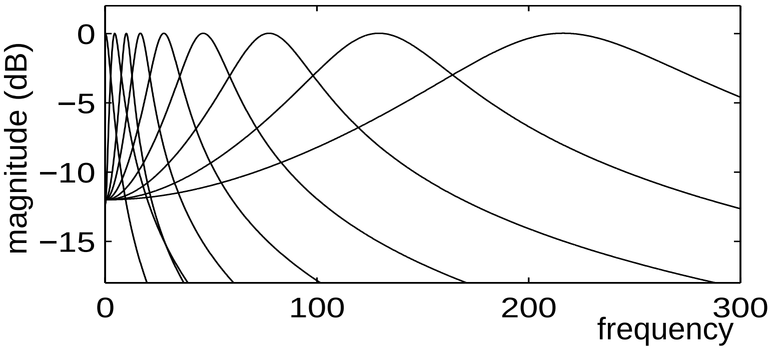


Figure 4.5 Responses of the modulation filterbank.

Both the modulation filtering and the envelope detection model analyze the magnitude of the fluctuations of the envelope of the acoustic waveform. As stated previously, the primary difference is that modulation filtering assumes a second filtering stage tuned to different envelope modulation rates. Figure 4.6 compares the processing output of these two models to a noise carrier with no modulation and with 56% (20log(m) = -5) modulation. Although the standard deviation of the input is the same for the modulated and unmodulated cases, the outputs of both models have relatively more fluctuation in the modulated case.

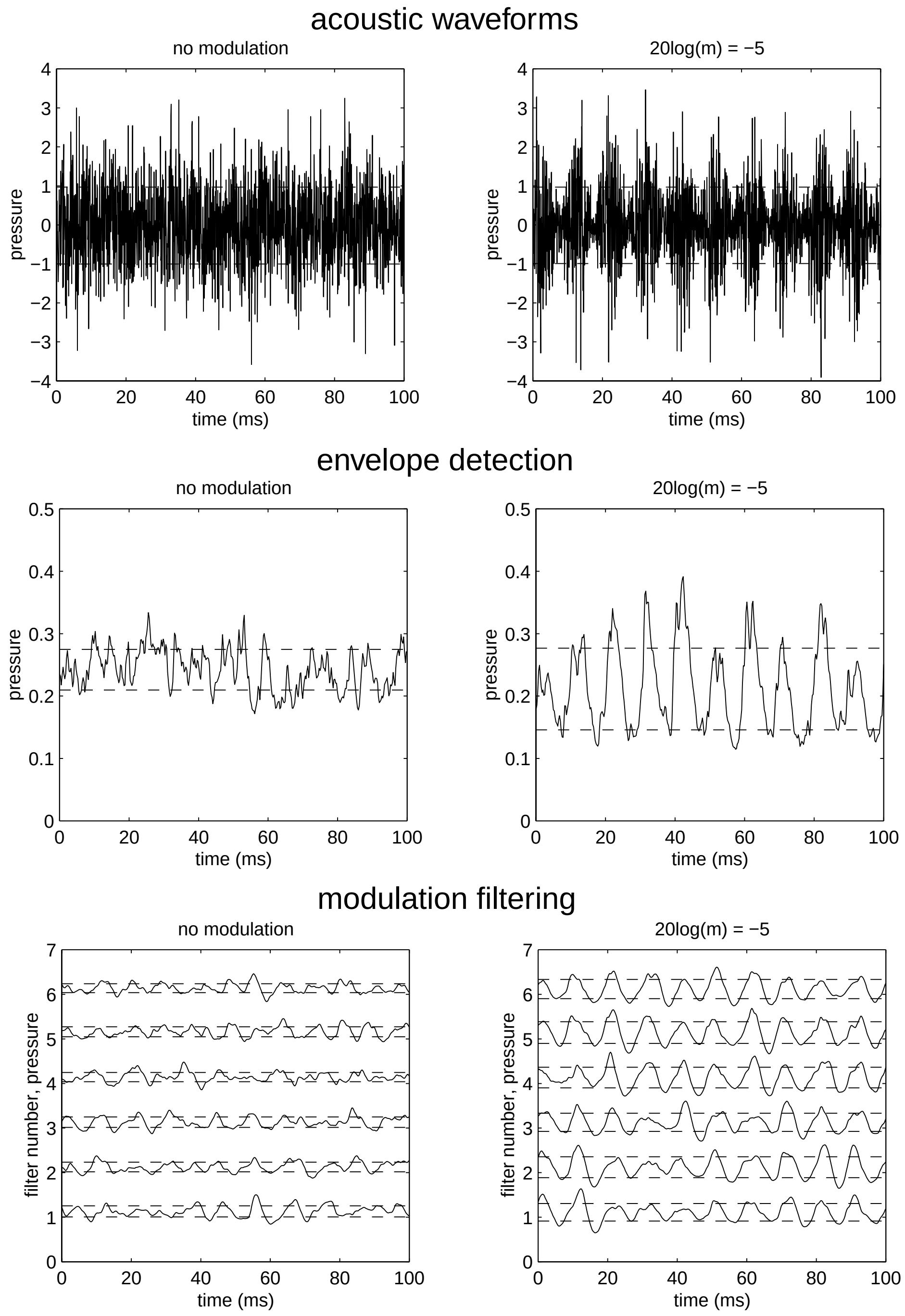


Figure 4.6 Comparisons of the amplitude modulation detection models. Dashed lines indicate standard deviations. The modulation filtering plots show the outputs of six auditory channels, each filtered by a modulation filter centered at 100 Hz.

4.3.3 Correlational Analysis

An overview of the correlational analysis is shown in Figure 4.7. This is an implementation of Licklider's model [Licklider 1951] together with a final stage that adds correlation estimates across channels [Meddis and Hewitt 1991a]. The first stage is the same gamma-tone approximation of cochlear filtering, used above. The transduction stage includes half-wave rectification, low-pass filtering, and a 2nd-order Butterworth high-pass filter with a cut-off of 4 Hz. Running autocorrelations are computed in each filter channel, and the results are added across channels.

Our implementation of running autocorrelation for each channel involves two stages. First, the instantaneous product of the current input and a version of the input delayed by some amount τ is computed for all time and all values of τ:

$$x_1(t,\tau) = x(t)\, x(t-\tau). \qquad (4.1)$$

Second, to form a running autocorrelation estimate, these sequences are low-pass filtered (in time and for each value of τ) to below one half of the final correlation sampling rate:

$$x_2(t,\tau) = x_1(t,\tau) * h_{lpf}(t). \qquad (4.2)$$

In the evaluations below, the correlation sampling rate was 25 Hz, and $h_{lpf}(t)$ was implemented as a 6th-order Butterworth filter with a -3dB point at 10 Hz. That is, after the low-pass filter, the running autocorrelations were sampled every 40 ms, and then summed across frequency channels to generate a sequence of summary correlogram estimates.

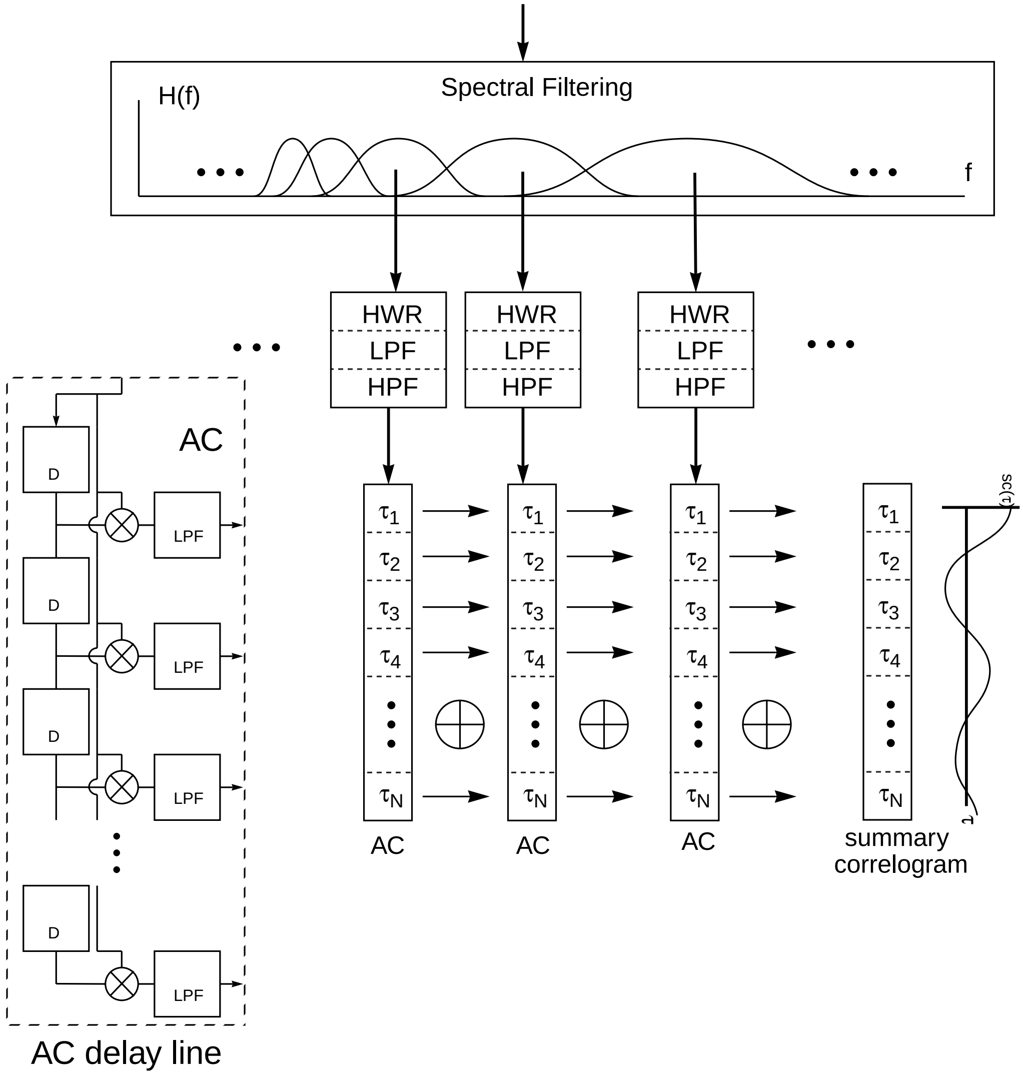


Figure 4.7 Overview of the correlational processing. Inset shows autocorrelation delay-line detail.

As described above, the position of the peak in the summary correlogram has often been shown to correlate with the reciprocal of the perceived pitch-frequency, although some authors have considered the entire waveform of the summary correlogram [Meddis and Hewitt 1991, Meddis and O'Mard 1997]. Here, a compromise between these is used. For each sample of the summary correlogram, our statistic is the maximum difference, across all delay values τ, between the summary correlogram values at delays of τ and $\tau/2$:

$$\textit{statistic} = \max [\ sc(\tau) - sc(\tau/2)\ ],\ (0 < \tau < 20\ ms). \qquad (4.3)$$

With a sinusoidal envelope, this difference peaks at a value of τ equal to the period of the sinusoid. Figure 4.8 includes examples of this decision statistic using the same noise carrier with no modulation, and with 56% modulation (20log(m) = -5) at 100 Hz. In the modulated case, the first peak (after zero delay) in the summary correlogram occurs at the period of the modulation or 10 ms. When there is no modulation, the summary correlogram approximates an impulse. Adding the individual correlation estimates across channels reduces some variance; consistent modulation shapes across channels add together, while inconsistent shapes often cancel each other. However, considerable variation remains across summary correlogram samples, shown in the lower half of Figure 4.6, due to the stochastic carrier.

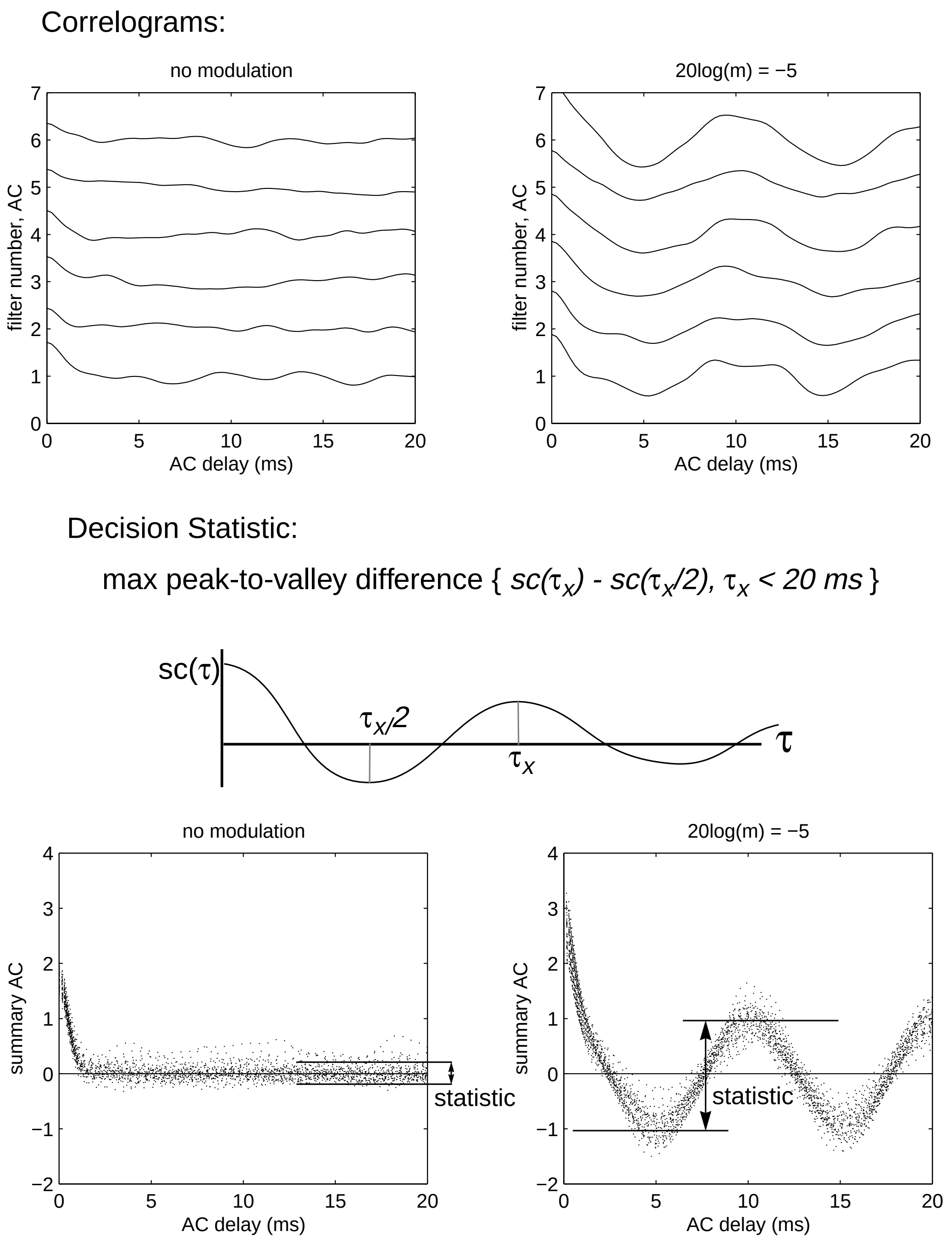


Figure 4.8 Samples of the correlogram output and super-imposed examples of the summary correlogram decision statistic. Input signals are the same as in Figure 4.6.

4.4 Comparing Predictions

The temporal modulation transfer function (TMTF) is a measure of auditory sensitivity to amplitude modulation as a function of modulation frequency. More specifically, the minimum detectable sinusoidal amplitude modulation depth is typically measured as a function of modulation frequency using wide-band noise carriers.

Each of the three models were first adjusted to predict TMTF measurements averaged from previous studies [Strickland and Viemeister 1997, Dau et al. 1997b]. The resulting models were then used to predict the discrimination thresholds for the high-pass filtered [s] and [z] tokens in noise. Because the natural fricatives were non-stationary, all three models were evaluated using multiple measurements in time, or multiple ‘looks’ [Viemeister and Wakefield 1991].

For the envelope statistic model we found the best match using an initial filter bandwidth of 3 kHz, centered at a frequency of 5.5 kHz. With these parameters, the filter was also roughly a matched-filter for the high-pass filtered [s] [z] tokens. The low-pass filter was a 1st-order Butterworth with a cut-off of 90 Hz. The normalized fourth-moment statistic [Strickland and Viemeister 1996 and 1997] was used.

To obtain multiple measurements in time, the output of the envelope detection mechanism was segmented using partially overlapping 50-ms windows that had 10-ms raised-cosine onset and offsets and a 30-ms constant center. The window increment was 40 ms so that onset and offset slopes intersected at the 0.5 level. The window length was chosen to ensure multiple periods in each window for the pitch-frequency range of interest. By modulating the DC offset in the envelope, the shape of the window can dominate measurements using the standard deviation or the fourth-moment. Therefore, the DC offset for each 50-ms segment was removed before weighting by the raised-cosine, and then added back before measuring the decision statistic.

Threshold predictions were obtained by using the difference in the decision statistic in signal and non-signal intervals over 100 simulations to estimate *d'* for each 'look.' Assuming independence of the individual measurements, a total detection *d'* was estimated as the length of a *d'* vector containing all looks [Green and Swets 1966]. With a stimulus duration of 500 ms used for the TMTF data, the vector included 12 elements, or 12 looks. A line was fit to the log of total *d'* estimates as a function of the log of the modulation depth. From this line, the modulation threshold was estimated from the point where the line crossed the *d'* threshold of 1.26 tracked in the perceptual TMTF measurements [Strickland and Viemeister 1997, Dau et al. 1997b].

With the modulation filtering and correlation models, the initial filtering stage was six 4th-order gamma-tone filters with center frequencies ranging from 4280 Hz to 6970 Hz. Filters overlapped at their half-power points, and the bandwidths were set using the equation provided in [Glasberg and Moore 1990]. To predict the TMTF data using modulation filtering, only the modulation filter tuned to the probe envelope frequency was considered. When predicting the fricative data, two modulation filters centered at 120 Hz and 200 Hz were used. The same windowing used with the envelope statistic simulations were used with the modulation filtering, and the standard deviation was the measured statistic.

As seen previously [Dau et al. 1997a-b], the modulation filtering was too sensitive to predict human performance without adding a large amount of internal noise. To obtain the best match to the TMTF data, a balance of internal noise was added both before and after modulation filtering.

Using the correlation model, the peak distance statistic described above was measured every 40-ms for the summary correlogram. To approximate the shape of the TMTF data, the first-order low-pass filter was used with a cut-off frequency at 280 Hz.

TMTF threshold predictions for all three models are shown in Figure 4.9. Each model provides a reasonable prediction across this frequency range.

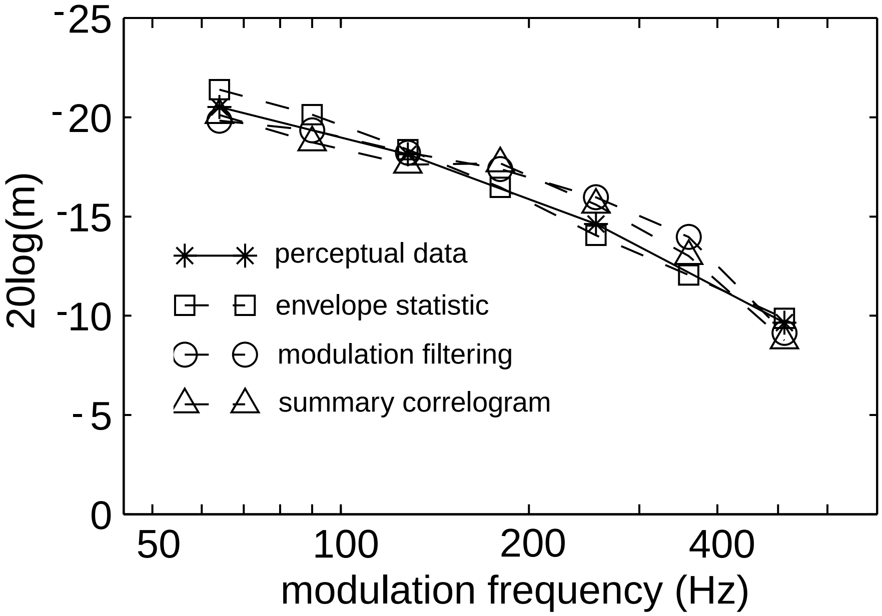


Figure 4.9 Three predictions of TMTF data: *m* is the modulation depth; perceptual data are an average of [Strickland and Viemeister 1997, Dau et al. 1997b].

Predicting the voicing detection thresholds for the natural, non-stationary, fricatives in noise required finding the fricatives (or more specifically finding the voicing in the fricative) within the 1 second of noise. For all model predictions below, only the three consecutive temporal segments that maximized the difference from the background noise were analyzed, providing three temporal looks per token. Total *d'* values were then estimated as a function of SNR.

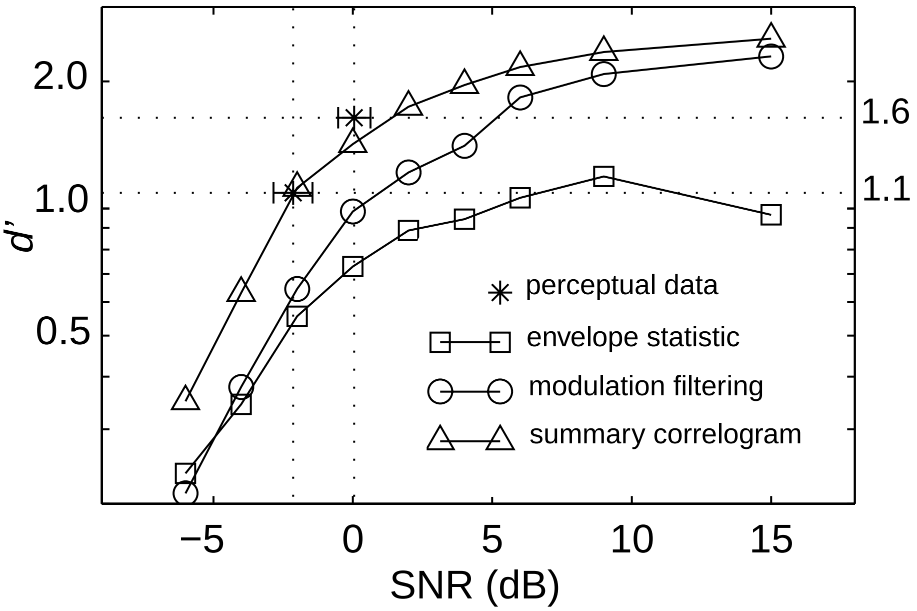


Figure 4.10 Discriminating the high-pass filtered [s] and [z]: perceptual data is an average across four subjects.

Figure 4.10 shows the *d'* estimates for each model's prediction of the discrimination of the high-pass filtered [s] and [z] tokens in noise. The model based on correlations provided the best prediction.

4.5 Modeling Implications

It appears that the envelope statistic was not sufficient to discriminate the [s] and [z] tokens (even at relatively high SNR values), because the measurement does not distinguish between the periodic voicing cues in [z] from the aperiodic fluctuations in [s]. Both the modulation filtering and the autocorrelation processing include specific modulation tuning and provide closer predictions.

Reasons for the difference in performance between these two are less clear, and could be specific to these simulations. By reducing the amount of internal noise, the modulation filtering model provides a better estimate of the [s] [z] data, but then over-estimates the TMTF sensitivity. One primary difference is that the autocorrelation mechanism integrates correlation estimates across frequency, while the modulation filtering simulations use the more general assumption that each output corresponds to an independent measurement. Integrating correlation estimates across frequency channels de-emphasizes envelope components uncorrelated across frequency in favor of correlated components. Another difference is that the correlation simulations used the low-pass filter to limit sensitivity, while the modulation simulation included internal noise.

It may be interesting to note that if the auditory system does include a cross-channel interval based representation, redundancies in this representation are likely to make it inefficient to maintain across many areas. Efficient decorrelation of the (potentially smooth and periodic) summary correlogram might approximate a cosine transform. Such periodic transformations exist in other perceptual systems [e.g. Wang and Shamma 1994]. In this case, the decorrelated representation would have many of the properties of the (demodulated) output of a modulation filterbank. The difference is that the envelope analyzed was first processed to identify common correlations across a wider frequency range.

4.6 Conclusions

This chapter identifies a secondary temporal cue which can reliably indicate voicing distinctions between [s] and [z]. This amplitude modulation cue had not been identified in previous studies of voiced fricatives [e.g. Stevens et al. 1992, Pirello et al. 1997]. Furthermore, once the cue was identified it was not clear what processing should be used to detect it. Three possibilities were investigated.

While cross-channel interval-based processing has been quite successful in predicting many aspects of pitch perception, here we show that these mechanisms can also predict TMTF thresholds and the detection of voicing for high-pass filtered strident fricatives in noise. Simulations using envelope-statistic and modulation-filtering models, fit to predict the TMTF data, did not predict the isolated speech data.

Chapter 5

Recognition Evaluations

The signal processing described in the previous three chapters was used in a series of speech recognition evaluations in noise. This processing shows improvements over other common signal processing techniques used to increase the noise-robustness of speech recognition systems. When compared to typical ASR speech representations, our processing reduces the error rate in noise by roughly a factor of 4.

5.1 Recognition Task

The evaluations here are speaker-independent word recognition tasks using the digits from the male talkers in the TI-46 database. As a collection of isolated words (digits, alphabet, and commands) recorded by Texas Instruments in 1981, this database represents perhaps the most trivial industry-standard recognition task.

Two modifications are made to the database to increase the challenge of the task: first the recognition evaluations are performed in considerable amounts of noise (SNR from 0 to 30 dB); second, the digits are placed randomly within two seconds of (noisy) silence.

Adding noise is a significant challenge for ASR, but one aspect of that challenge is that with higher levels of background noise, *finding* the word within the background noise becomes increasingly difficult. Many ASR systems, and certainly the ones evaluated here, make maximum likelihood decisions without any explicit confidence measures. The most probable word is chosen, even if that probability is extremely low. As the background noise increases, the likelihoods of all models drop to the point where the maximum likelihood response can occur in the background noise itself. At this point, the representation of the speech signal is corrupted so much that the model ‘finds’ the best match to the background noise. The TI-46 digits are hand-aligned and centered in the files in the database. Without adding surrounding silence, finding the speech is not an issue.

The additive noise used in these evaluations was shaped to match an estimate of the long-term average speech spectrum [Byrne and Dillon 1986], as shown in Figure 5.1.

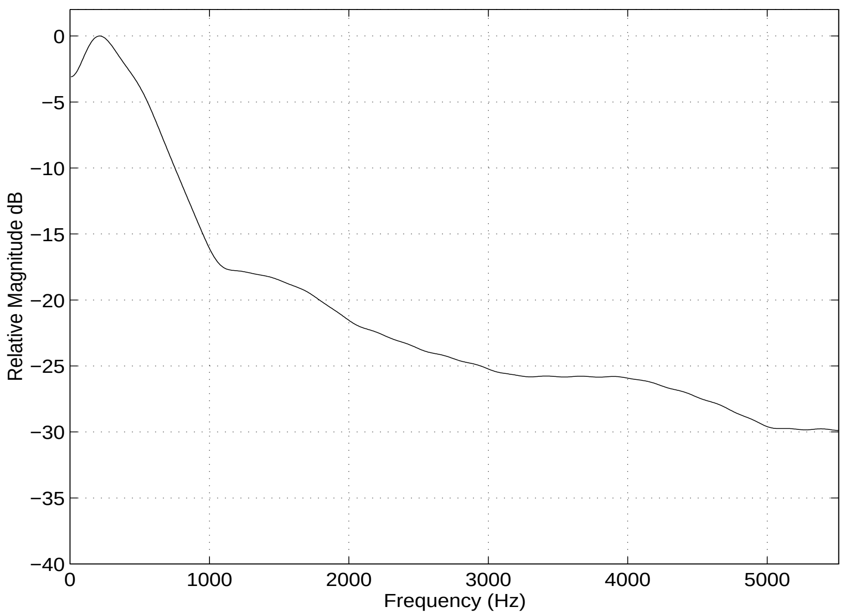


Figure 5.1 Spectrum of the additive noise used in the recognition evaluations was set to match the long-term average speech spectrum [Byrne and Dilllon 1986]

5.2 Stochastic Modeling Structure

These evaluations use the general hidden Markov model (HMM) structure outlined in Chapter 1 [Rabiner and Juang 1993]. Models were trained for each word, and the model that provided the maximum likelihood for an unknown token determined the word recognized. However, instead of using a single model for each word, two models for each word were used in parallel: one trained from clean data, and the other from data corrupted by noise. The model that provided maximum likelihood from either set determined the word recognized.

For all models, 6-states per word, simple left-to-right state transitions, continuous Gaussian densities, diagonal covariances, and fixed global variances were used. Mean feature vectors and transition probabilities for each state were trained as described below, but variances were set to the global variance estimated over all tokens in the training set. This technique is useful with limited training data and when the testing environment is significantly different from the training environment [Jankowski et al. 1995].

The clean models were trained in two stages. Training words were first isolated from the surrounding silence based on the total signal energy. The models were initialized assuming a uniform distribution of the words across the 6 states in the model. Iterative Viterbi (max-path) alignment and training was then applied until the average log probability decreased by less than a threshold. The forward-backward algorithm improved the estimate for each model using a similar convergence criterion.

When the test environment differs from the training environment, recognition performance deteriorates. A common approach to address this issue is to train models using noisy data [Rabiner and Juang 1993]. One set of clean models was built, as described above, and then a second set of ‘noisy models’ was built using training data at an SNR of 9 dB. Both sets of models were used for recognition; the model with the highest probability (from either set) determined the word recognized. To train the noisy models, stationary background noise was added to the training data, and then forced-Viterbi alignment with the corresponding clean model was used to isolate the noisy speech from the background. The same Viterbi and forward-backward training algorithms used for training clean models, were then used to train noisy models from the isolated noisy words. Recognition of the testing data was performed using Viterbi alignment with both sets of models and choosing the model with the highest probability.

5.3 Baseline Signal Processing

Two baseline front-ends are considered: linear prediction cepstral coefficients (LPCC), and Mel-frequency cepstral coefficients (MFCC). Each front end computes a spectral estimation every 10 ms using overlapping 30-ms Hamming windows. LPCC are computed in two stages [Rabiner and Juang 1993]: 12th order, autocorrelation-based linear prediction provides an all-pole vocal-tract transfer function. Real cepstral coefficients are then recursively computed for this minimum-phase estimation. MFCC are computed in three stages [Davis and Mermelstein 1980]. The power spectrum is computed using a zero-padded fast Fourier transform (FFT). To estimate the energy at the output of each approximate auditory filter, power spectrum outputs are weighted by a triangular filter shape and then summed. The filters have a half-power bandwidth of 100 Hz up to center frequencies of 1 kHz, and a bandwidth of 0.1 times the center frequency above 1 kHz. A discrete cosine transform (DCT) converts the spectral estimation obtained from the logarithmic energy across filters into a final cepstral vector. A 13 element cepstral vector, and its temporal derivative (approximated by the slope of a line fit to 7 cepstral points) are obtained for each front end, but the undifferentiated spectral level term (c_0) is ignored during recognition. Therefore, the baseline feature vectors have 25 elements.

5.4 Implementation of the Model Signal Processing

Figure 3.1, in Chapter 3, shows a block diagram for the adaptation, peak isolation, and threading processing.

5.4.1 Adaptation

The adaptation mechanisms described in Chapter 2 are implemented as a modification of the process used to obtain MFCC. Before the DCT, the logarithmic filter energies of MFCC are processed through the dynamic stages derived in Chapter 2 to obtain the adapting spectral estimation vector MFCCA. Therefore, the adaptation mechanisms alter the sequences of logarithmic energy estimates obtained for each approximate auditory channel.

5.4.2 Peak Isolation

The peak isolation mechanism was described in Chapter 2. A truncated cepstral vector is obtained for each frame. This cepstral vector is weighted by a modified raised-sin lifter [Juang et al. 1987], and the inverse DCT (IDCT) is used to transform back to a (modified) spectral estimate. This estimate is half-wave rectified, and the individual peaks were scaled to match the peak magnitudes of the original spectral estimate. The DCT is then used to obtain a final cepstral representation of the peak isolated spectral estimate.

5.4.3 Peak Position and Motion

The threading algorithm used to parameterize the position and motion of dominant spectral peaks was described in Chapter 3. During the peak isolation processing, the frequency position of local spectral peaks are stored for each 10-ms frame. A two-stage process is used to convert these peaks into the final representation used by the recognition system. In the first stage the peaks are threaded using dynamic programing, the threads are fit to moving 7-point second order polynomials, and the frequency derivatives are estimated. The second stage tracks the dominant peak frequency and associated frequency derivative for each of three equally-spaced spectral regions. Unlike the adaptation and peak isolation mechanisms which *alter* the 25-element feature vector used for recognition, the current processing *adds* six more elements. However, during training it was found that the frequency derivative in the highest frequency region had little variance across the training set. It was therefore ignored for these recognition evaluations. When the parameterization of peak position and motion is included in the recognition evaluations, there are five additional elements in the feature vector.

5.4.4 Temporal Processing of Voicing Information

Chapter 4 showed that a correlation-based representation of pitch-rate amplitude modulation was consistent with perceptual data describing the detection of amplitude modulation and the detection of voicing for high-pass filtered [s] and [z] tokens in noise. Therefore, of the three mechanisms considered in Chapter 4, only the correlation-based processing was evaluated in recognition tasks. The correlation model [after Licklider 1951], maintains running autocorrelations in each auditory channel, and then adds these together to identify the common periodicities across channels.

The model was extended to generate suitable voicing features for ASR. Instead of analyzing a single high-frequency region, three regions are used, corresponding to the three regions in the peak position and motion processing above. The voicing statistic used is the maximum peak-to-valley difference between any time-lag τ_1 and any smaller time-lag τ_2 ($\tau_2 < \tau_1$). Finally because the presence of voicing (and not the amount of voicing) is assumed to be relevant for ASR, the logarithmic magnitude of the voicing statistic is further compressed by a sigmoidal function. Figure 5.2 shows the voicing features in the three frequency regions together with their temporal derivatives, for a total of six voicing features. In general, the voicing features mark voiced speech within the uncorrelated noise background. The voicing features are low during the [s] sounds in “six.”

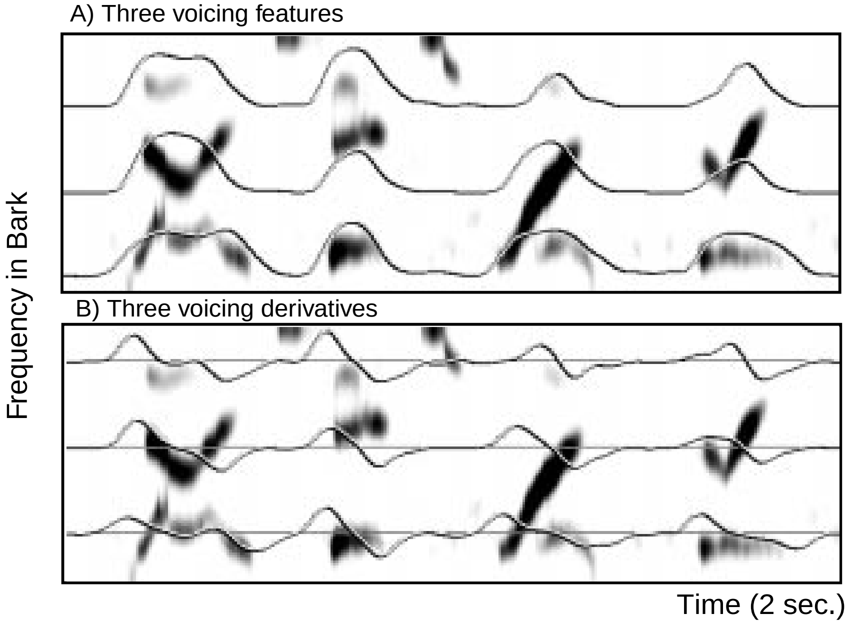


Figure 5.2 A) Voicing features in three spectral regions, and B) their temporal derivatives.

5.5 Other Techniques Targeting Noise

In addition to comparisons with the baseline LPCC and MFCC features, recognition evaluations were also performed using RelAtive-SpecTrAl (RASTA) processing [Hermansky and Morgan 1994] and a variety of common signal processing techniques that are targeted specifically at improving recognition performance in noise: spectral subtraction, spectral scaling, non-linear spectral scaling, and cepstral normalization.

RASTA involves filtering the logarithmic temporal trajectories (log energy temporal excitation patterns) with a bandpass filter that has a sharp zero at DC. By de-emphasizing slow and fast changes with time, RASTA also provides an adapting response. In the comparisons below, the RASTA technique was applied directly to the logarithmic filter energies, without the additional PLP processing used in its original optimization [Hermansky and Morgan 1994]. The ‘standard’ RASTA filter:

$$H(z) = 0.11\frac{(2 + z^{-1} - z^{-3} - 2z^{-4})}{1 - 0.94z^{-1}}$$

was used and performance was not compared with other RASTA variations which optimize the compressive and expansive non-linearities for the specific acoustic environment. Unlike the RASTA technique which can be described as a (smoothed) first-order differentiation, the adaptation mechanism proposed in Chapter 2 does not provide zero output for constant input. Instead, the adaptation stages converge to static targets on the I/O curves. Also unlike the RASTA technique, (offset) recovery is roughly three times slower than (onset) adaptation.

The power spectrum of the sum of two uncorrelated signals is the sum of the two power spectra for the individual signals. That is, uncorrelated signals are additive in power. The power spectrum for speech in noise, *X(f)*, is the sum of the power spectrum for the clean speech, *S(f)*, and the noise power spectrum, *N(f)*.

$$X(f) = S(f) + N(f)$$

Spectral subtraction assumes that given a reliable power spectrum estimate for a stationary background noise, $\overline{N}(f)$, an approximation of the original clean speech signal, $\bar{S}(f)$, can be obtained by subtracting the noise estimate from the power spectral estimate for the signal and noise.

$$\bar{S}(f) = X(f) - \bar{N}(f) = S(f) + N(f) - \bar{N}(f)$$

Unfortunately, short-time power spectral estimates of (even) stationary noise signals have considerable variance. That is, the values of the measured background noise, *N(f)*, will change considerably from frame to frame. For some frames the speech signal, *S(f)*, will be near zero, and the measured background noise, *N(f)*, will fluctuate to be less than the stationary background noise estimate, $\bar{N}(f)$. Therefore, after spectral subtraction, the final result can be negative. Because the next step for ASR is to take a logarithm, these negative values must be clipped.

And that is the beginning of the end. Choosing the clipping level sets an arbitrary floor on the log-magnitude spectral estimates (e.g. 0 dB if power estimates below 1 are clipped). Consider a background noise which averages 30 dB above the clipping point. Figure 5.3 shows the log-magnitude input/output function for spectral subtraction.

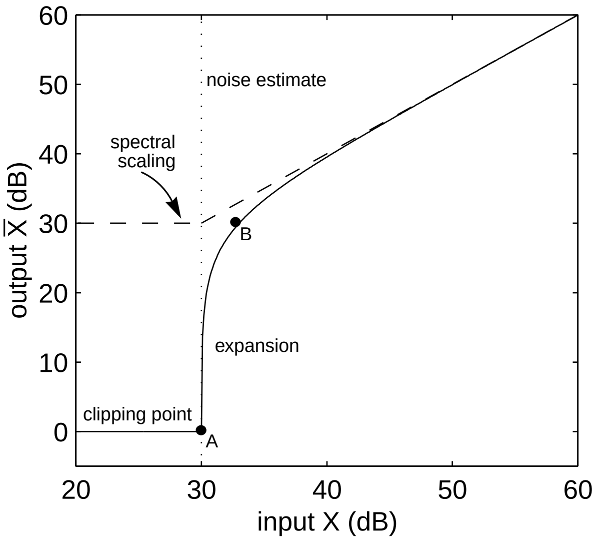


Figure 5.3 Spectral subtraction input/output function. Points A, (30, 0) and B, (33, 30) show the expansion of spectral subtraction.

When the measured noise is exactly 30 dB, the subtraction of the expected power spectrum lowers the current power estimate to 0, which is then clipped to 1 before the logarithm, leading to a final log-magnitude value of 0 dB, or point A in Figure 5.3. However, if the current log-magnitude power spectrum fluctuates up to 33 dB, or 3 dB higher, then its power spectrum is twice the expected value. After spectral subtraction and the logarithm, the final estimate is 30 dB, or point B in Figure 5.3. Spectral subtraction expands the original 3-dB noise fluctuation into a 30-dB fluctuation. Needless to say, recognition systems are extremely sensitive to random 30-dB fluctuations. One solution to this problem is to raise the clipping point to the level of the noise estimate. Then spectral subtraction approximates spectral scaling.

In spectral scaling, the reference level for the logarithm is the current noise power-spectrum estimate. Equivalently, the log-magnitude of the average background noise is subtracted from each current log-magnitude estimate. The final result is then clipped below 0 dB. The dashed line in Figure 5.3 shows the input/output function for spectral scaling. (To compare spectral scaling with spectral subtraction a fixed 30 dB offset is added to the spectral scaling function in Figure 5.3.) As the noise level rises, the dynamic range of the speech above the noise reduces. The recognition system is, of course, dependent on the diminishing fluctuations of the speech above the noise.

Non-linear spectral scaling tries to correct for this loss of dynamic range. As implemented here, two log-magnitude spectrograms are obtained: one from linear spectral scaling, and a second copy which is then scaled (after the logarithm) so that the peak dynamic range, above the noise floor, is fixed to a specific value. A weighted average of these two is used as the final sequence of log-magnitude spectral estimates. For the evaluations below, the relative weights used in the averaging were iteratively optimized to improve recognition performance for this task.

A second technique to compensate for this loss of dynamic range is cepstral normalization. As the dynamic range across a single log-magnitude spectral estimate reduces, the length of the cepstral vector also reduces. In cepstral normalization, the total length of each cepstral vector is normalized to unity. For the task evaluated here, only non-linear spectral scaling and cepstral normalization provided clear improvements.

5.6 Evaluations

Figure 5.4 shows the degradation of recognition performance for the two baseline front ends, MFCC and LPCC. Using a frequency scale that is warped to approximate auditory frequency selectivity increases recognition robustness. A similar improvement was found previously [Jankowski et al. 1995]. This trend is consistent with the ASR shift from LPCC to MFCC in the last 5-10 years.

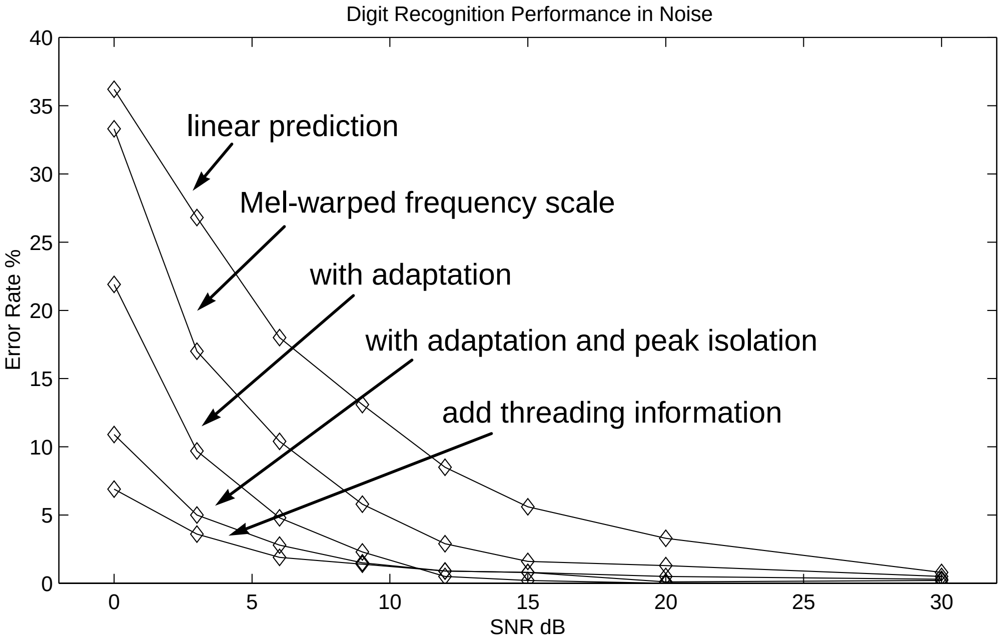


Figure 5.4 Baseline recognition performance.

Figure 5.4 also includes the performance with adaptation, adaptation and peak isolation, and then with adaptation, peak isolation and the threading processing. Each of these provides additional improvements in recognition performance in noise.

Figure 5.5 compares the baseline MFCC representation and the processing proposed here with other common techniques aimed at improving recognition robustness. Of the other techniques considered, only those which alter the dynamic range of the spectral representation (cepstral normalization and non-linear spectral scaling) showed considerable improvements in recognition robustness.

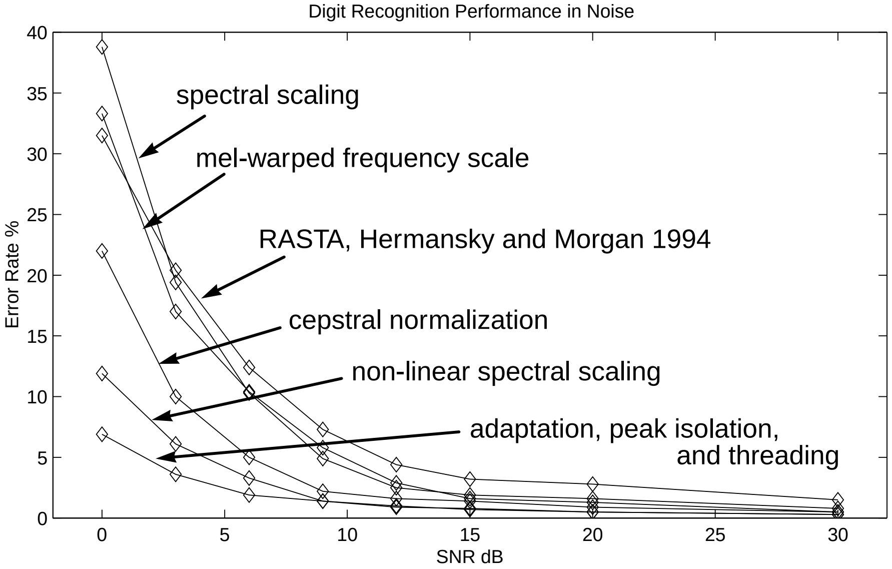


Figure 5.5 Recognition comparisons with other signal processing techniques.

Finally, Figure 5.6 shows the recognition performance when only the 5 threading features are used, and when voicing features described above are used with different pieces of the previous representations. When used alone, the 3 frequency positions and 2 frequency derivatives are insufficient to discriminate the current data. However, adding 6 voicing features to the 5 threading features reduces the error rate considerably. The error rate also decreases when the 6 voicing features are added to the 30-element feature vector that includes the cepstral and delta-cepstral representations after adaptation and peak isolation and the 5 threading features. This final system, incorporating the four processing mechanisms of this dissertation, provides a 1.5% error rate at 3 dB SNR, or more than an order of magnitude fewer errors than the typical representation used in ASR systems.

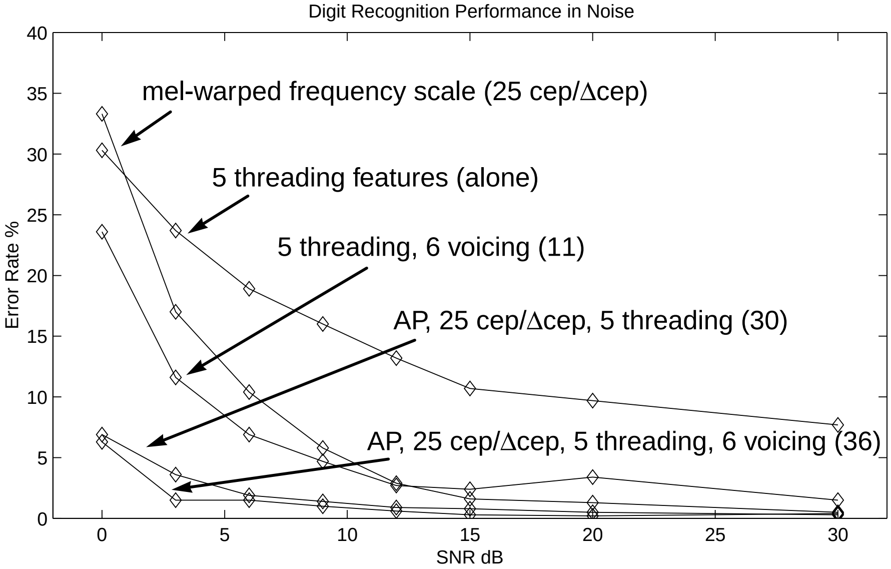


Figure 5.6 Evaluations with threading and voicing information. AP is adaptation and peak isolation, and Δcep refers to cepstral derivatives.

5.7 Interpreting the Results

The recognition task used here attempts to assess the potential noise-robustness improvements of the algorithms of the previous chapters in the current recognition paradigm. The task requires the recognition system to find and identify a word in background noise. This may be a reasonable approximation for many current limited-domain voice-control applications. Solutions for this task in clean environments have been available for years [Rabiner and Juang 1993], and as expected, most speech representations evaluated here lead to very little error at high SNRs. However, the results above show that mechanisms which incorporate aspects of auditory perception can dramatically reduce the performance degradation in noise.

Speech information is encoded in highly-redundant, multi-dimensional representations which range across many time scales. In addition to identifying aspects of auditory perception which are typically ignored in the representations of speech used for ASR, there are perhaps two consistent motivating ideas which helped ensure that the mechanisms described here were successful and complimentary. In general, each processing mechanism addressed a different dimension or time-scale (see Figure 1.8), and the processing for most mechanisms was de-coupled across frequency.

Consider the relevant dimensions for each mechanism. The adaptation mechanism emphasizes onsets and transitions in frequency which occur in the syllabic range, while the peak isolation mechanism enhances changes across the spectral range. The threading parameters in turn characterize the motion of the isolated spectral peaks in the syllabic range, while the voicing features characterize the voicing range. More specific to the task used above, the voicing features help distinguish speech from noise while the threading information helps discriminate words. Obviously, if these four mechanisms had addressed the same dimension, it is less likely that their combination would improve recognition results.

Motivations for de-coupling the representations of speech across frequency are described in [Allen 1994]. Because the feature vectors for ASR systems are almost always a function of the entire spectral range (recall that cepstral coefficients are the DCT of the entire log-magnitude spectral estimate), distortions in one spectral region influence the entire feature vector, reducing performance. Humans, on the other hand are much more immune to static disturbances in a particular spectral region. For an ASR system to use this approach, at least some feature vectors must be a function of specific spectral ranges. In the processing described in previous chapters, clearly the voicing features, the threading features, and the processing for the temporal adaptation are largely de-coupled across frequency.

However, the peak isolation mechanism is not similarly de-coupled. The spectral estimate after peak isolation starts with the spectral estimate after the cepstral vector has been weighted by a raised-sin function. Near zero, the raised sin function increases almost linearly, approaching the frequency response of a (spectral) derivative. As shown in Figure 2.10, the response after liftering is therefore more dominated by spectral changes. While this processing provides some de-coupling across frequency (spectral slope for instance is de-emphasized), the processing is still a function of the entire spectral estimate. Regional spectral estimates, regional cepstral representations, and regional cepstral liftering might lead to additional improvements. This will be one area of future work.

While this task may be a reasonable approximation for many current limited-domain voice-control applications, considerable work today addresses the transcription of large-vocabulary continuous and even spontaneous speech. To limit the “local domain” for these tasks, a hierarchy of HMMs are used. Phrases are modeled as groups of words which are modeled as sequences of phonemes. Alignment requires identifying the most probable sequence of phonemes, constrained by the probabilities of the word pronunciations, which are in turn constrained by the probabilities of the word sequences in the expected phrases. Final recognition performance for these systems becomes extremely dependent on the reliability of the statistical estimates for the higher-level sequences. Pronunciation and word-sequence (or grammar) models often limit performance. One significant question that remains is: Do the current processing improvements generalize to these more complicated tasks?

Recall that the recognition task used here included word-level models, and that two of the processing stages (temporal adaptation, and the frequency derivatives in the threading processing) provided a context-dependent response which can last for several frames. If word models were instead built from phoneme models, the processing here would therefore most likely provide different responses depending on the context of the phoneme. Tri-phone (and even quint-phone) models, or different models for each phoneme in every possible preceding and following context, are already commonly used [Woodland et al. 1998], and would appear to be necessary with the current processing. This will be a second area of future work.

In conclusion, the mechanisms described in the previous chapters each address a somewhat complimentary aspect of the speech signal, and together significantly decrease the error rate of a word recognition system in noise.

Chapter 6

Summary and Extensions

This dissertation provides evidence that advances in robust speech recognition can be made by incorporating mechanisms which approximate aspects of human auditory signal processing. Mechanisms including adaptation, peak isolation, an explicit parameterization of the position and motion of local spectral peaks, and a correlation-based analysis of perceptual voicing information are shown to improve recognition performance in noise.

This work suggests areas of future research in psychoacoustics, auditory physiology, and speech recognition.

6.1 Threads

As described in Chapter 3, the threading operation described here can be viewed as an early step in auditory scene analysis.

Traditional psychoacoustic modeling efforts usually focus on the perception of static sounds. Or sightly more generally, either variations across frequency (e.g. spectral masking), or variations across time (e.g. temporal masking) are considered. With spectral measurements, changes in the logarithmic output available across assumed auditory filters, or the spectral excitation pattern [e.g. Zwicker 1970], are found to correlate with perceptual performance. For temporal measurements, changes in the output of auditory filters with time, or temporal excitation patterns, are considered [e.g. Oxenham and Moore 1994]. For each of these, the subject's responses are assumed to be made as an ideal observer. That is, using the excitation pattern (usually corrupted by internal noise) as input, the subject chooses the option that has maximum likelihood.

Speech is non-stationary. Speech recognition systems therefore consider variations in excitation patterns in both time and frequency. But the concept is similar. The recognition process makes the maximum likelihood choice, now given 2-dimensional excitation patterns. The adaptation and peak isolation mechanisms described here change the characteristics of the excitation patterns, and the voicing detection adds (yet) another dimension.

But threading can be viewed differently. In addition to monitoring excitation patterns, threading assumes the subject is also *actively* piecing together higher-level structure that is also *observed*. Obviously, information is not manufactured by a receiver. That is, at first glance it might appear that an ideal observer would not do any better observing redundant representations derived from earlier representations. But if we assume the excitation patterns are analyzed for low-level structure (e.g. threaded) before some of the internal noise corrupts the observation of excitation patterns, and further that (independent) uncorrelated internal noise also corrupts the observation of the intermediate structure, then the ideal observer would use information from both stages to improve performance. Figure 6.1 shows an overview for this arrangement.

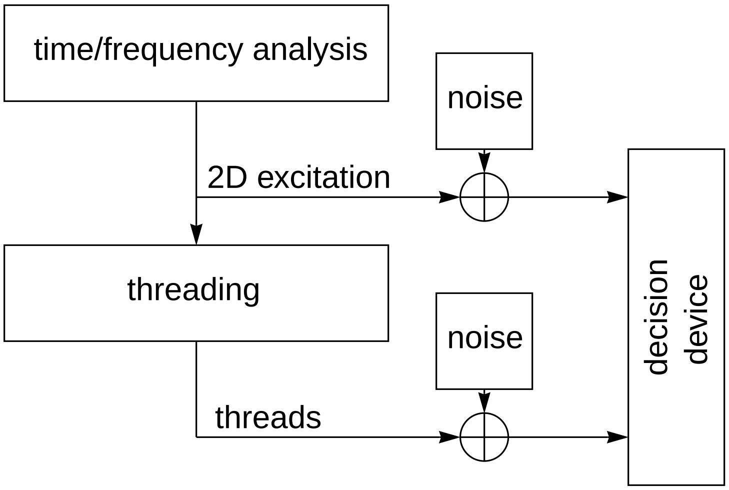


Figure 6.1 Threading in a two-stage model of auditory perception.

Unfortunately, other than the recognition improvements shown in this dissertation, only qualitative points can be made in support of this structure. However the evidence is considerable and growing. Many auditory scene illusions have similar visual analogies: when an edge is partially obscured, the observer fills in the missing piece; or when flickering lights are correlated in either time or space they can be grouped accordingly [Bregman 1990]. In fact, hearing (and seeing) might be best understood as piecing together partially obscured measurements to obtain information about objects in the environment. With this requirement, a full multi-dimensional excitation pattern representation of sound would be inefficient, and the neural representation is more likely to focus on threads, or aspects of the sound corresponding to the auditory analogy for visual edges. To understand the inefficiency of a 2-D excitation pattern, consider the size of the space required for 0.300 seconds of sound. Assuming 30 auditory channels, and temporal samples every 10 ms, this implies 900 nearly independent dimensions. Clearly we do not have the cognitive ability to maintain and recall arbitrary patterns within this space.

More directly, recent attempts to predict the perception of what appear to be very basic non-stationary experiments are not predicted assuming standard excitation patterns, and may instead be consistent with threading. When describing their ‘multi-look’ hypothesis [Viemeister and Wakefield 1991] (an ideal observer using multiple independent observations of a temporal excitation pattern), Viemeister and Wakefield also showed that when the temporal excitation pattern stayed the same, subjects were able to do better than an ideal observer with independent observations. When listening for a tone in noise, as the duration of the tone increases, the observer has more observations. If the observations are independent, thresholds should drop with the square root of the number of observations [Green and Swets 1966]. Instead, for tones in noise, thresholds drop with the total energy of the tone, or about twice as fast. If the ideal observer is using the excitation patterns together with some type of threading, then we would expect increased performance from observations that form strong threads (i.e. static tones).

When detecting short duration bursts of noise in a noise background, as the bandwidth of the noise increases, we might also expect thresholds to drop with the square root of the number of observations across frequency. Again, the auditory system does better, behaving more as if total intensity were integrated across frequency [Hant et al. 1997]. We should also consider the possibility that perceptual grouping of the discontinuity across frequency (a vertical thread) is increasing performance.

Finally other experiments show that intensity discrimination [Zeng 1994, Zeng 1998] and frequency selectivity [Hant et al. in press] are considerably degraded in the context of forward and/or backward temporal maskers. The temporal maskers would interfere with a threading mechanism, which could be causing the measured performance degradation.

6.2 Correlation Inconsistencies

In this dissertation we extended the application of Licklider's duplex theory to show that, in addition to predicting many aspects of pitch perception, this approach is also sufficient to predict TMTF data as well as the detection of voicing for high-pass filtered strident fricatives in noise. Such processing was further shown to be of some use for robust speech recognition. But there are at least two significant challenges which must be considered. First, the physiological mechanisms for implementing these measurements have not been identified. Second, measurements with electrical hearing have not supported the use of mid-frequency (300-1500 Hz) temporal information.

The first issue may still be a technology issue. The building blocks for autocorrelation type measurement: neural delays, multiplication (or coincidence detection) and low pass filtering, are widely available in the neural substrate. But finding evidence for the entire structure would require extensive (simultaneous, or at least well-synchronized) neural population measurements. Unfortunately, temporal measurements are often made from a single cell, and extensive population measurements are still not practical. As a starting point, population measurements have shown that the temporal information for this type of processing is available in the auditory nerve [Cariani and Delgutte 1996a-b]. As shown in Chapter 4, after low-pass filtering, the neural representation of the running autocorrelation could easily be down-sampled (to 25 Hz in the modeling in Chapter 4). It is also likely that neural processing would reduce redundancies in the representation across the time-delay variable τ, perhaps using an approximate cosine transform. These two stages would lead to single units which respond selectively to different modulation rates and have relatively low average firing rates. Such responses are measured in many areas of the auditory system [e.g. Langner 1992], but this provides only minimal support for the current theory.

Current non-evasive neural population measurement techniques are limited to monitoring average blood-flow rates and other very slow responses. While these measurements are helpful for mapping response areas, they show little promise for the types of measurements necessary to understand low-level temporal auditory processing. Unfortunately, these discoveries may have to wait until 3D electrical field measurements are available with cellular resolution in space and μsec resolution in time. Or at least, then we'd know for sure.

Perhaps the greater immediate challenge to Licklider's theory are pitch jnd measurements with cochlear implants. Briefly, cochlear implants use a series of electrodes near different places on the basilar membrane to provide (nearly) direct electrical stimulation of the auditory nerve. Subjects with cochlear implants are not able to use temporal information to discriminate fine pitch distinctions at 'normal' pitch ranges (80-500 Hz) [e.g. Townsend et al. 1987, Shannon 1992]. Maintaining a version of the duplex theory therefore requires that some of the differences in electrical hearing must confound the available temporal processing. The first and most likely culprit is that the neural fine structure associated with electrical stimulation is profoundly different from that available from acoustic hearing [Wilson et al. 1994]. (Perhaps Cariani and Delgutte's measurements should be reproduced using animals with cochlear implants.) A second possibility is that the neural processing which analyzes *population* interval information may rely on the precise phase relationships made available by the mechanical wave-guide in the cochlea, but absent in electrical hearing. In any case, the ability of some subjects with cochlear implants to use pitch rate temporal information might be best approximated by the ability of subjects with normal hearing to use pitch-rate temporal information with stochastic carriers (i.e. amplitude modulated wide-band noise). That is, if we assume electrical stimulation is at best providing no information at the carrier rate, then only envelope cues with stochastic (or at least useless) fine structure are available for neural processing. This is identical to the assumption used in recent models of the perception of low modulation rate (2-20 Hz) cues in electrical hearing [Shannon et al. 1995]. Therefore, without the reliable fine structure from deterministic carriers, we should expect the fact that pitch-related performance with electrical hearing only approaches that of normal hearing with stochastic carriers [Burns and Viemeister 1981].

6.3 Looking Forward

The extreme auditory periphery is well understood. Current and future work will focus on characterizing the functional significance of increasingly more central neural centers. Figure 6.2 shows an overview of the anatomy of the auditory neural pathway.

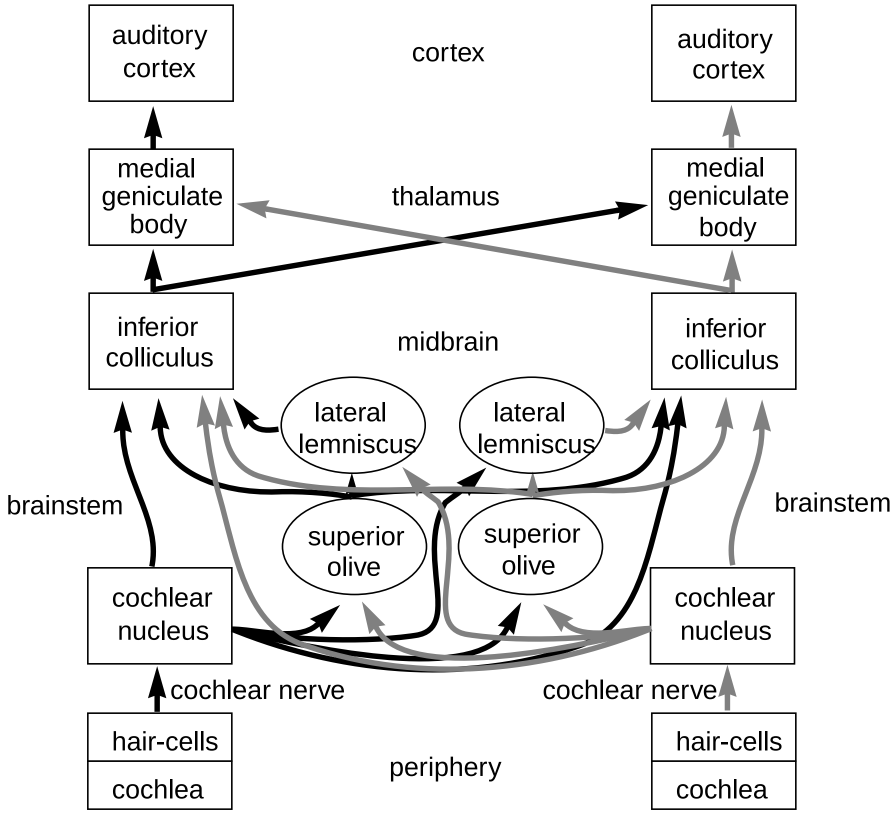


Figure 6.2 Overview of the auditory neural pathway.

After the initial anatomy is mapped, the challenges will be understanding the functionality and interactions of the various processing stages. At least four trends are already apparent: 1) Some form of the tonotopic map, consistent with the initial cochlear filtering, is maintained throughout most of the auditory system; 2) Average firing rates decrease considerably in increasingly higher levels (more central neural regions); 3) Many stages are connected for binaural comparisons; and 4) Feedback from higher levels to the periphery (efferents, not shown in Figure 6.2) may be as pervasive as signaling from the periphery to the higher levels (afferents).

Unfortunately, the challenge for the neurophysiologist is very analogous to asking a freshman engineer to figure out how a computer works using an oscilloscope, an Ohm meter, his best guesses for a block diagram, but with no schematics, no monitor, and no direct contact with the designer. To improve the analogy, the performance of the computer would change as the measurements were made, the computer would typically stop working after a few hours of measurements, and each new replacement computer would be slightly different. Both the engineer and the neurophysiologist are left to measure countless signals in hopes of finding the elusive insights to understanding. Progress will be slow.

From a psychoacoustic perspective, modeling efforts will address the perception of non-stationary sounds. Early work in this direction reveals that the effects of perceptual grouping and auditory scene analysis may have key roles with even the most basic non-stationary stimuli. As consistencies emerge between the measured physiology and the predictions of psychoacoustic models, engineers will translate these systems into increasingly robust speech processing applications.

Bibliography

Aikaiwa, K., and Saito, T. (1994). "Noise robust speech recognition using a dynamic-cepstrum," Proceedings ICSLP, Yokohama, 1579-1582.

Allen, J.B. (1994). "How do humans process and recognize speech?" IEEE Trans. Speech and Audio Proc., **2**, 567-577.

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.

Bacon, S.P., and Grantham, D.W. (1989). "Moulation masking: Effects of modulation frequency, depth, and phase," J. Acoust. Soc. Am., **85**.6, 2575-2580.

Bregman, A.S. (1990). *Auditory Scene Analysis*, MIT Press, Cambridge.

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.

Burns, E.M. and Viemeister, N.F. (1981). "Played-again SAM: observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am., **70**.6, 1655-1660.

Byrne, D., and 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.

Cariani, P.A. and Delgutte, B. (1996a). “Neural correlates of the pitch of complex tones: I. Pitch and pitch salience,” J. Neurophys. (Bethesda), **76**.3, 1698-1716.

Cariani, P.A. and Delgutte, B. (1996b). “Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch,” J. Neurophys. (Bethesda), **76**.3, 1717-1734.

de Cheveigne, A. (1998), “Cancellation model of pitch perception,” J. Acoust. Soc. Am., **103**.3, 1261-1271.

Cohen, J.R. (1989). “Application of an auditory model to speech recognition,” J. Acoust. Soc. Am., **85**, 2623-2629.

Dau, T., Kollmeier, B. and Kohlrausch, A. (1997a). “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**.5, 2892-2905.

Dau, T., Kollmeier, B. and Kohlrausch, A. (1997b). “Modeling auditory processing of amplitude modulation. II.Spectral and temporal integration," J. Acoust. Soc. Am. **102**.5, 2906-2919.

Dau, T. and Pueschel, D. (1996a). “A quantitative model of the ‘effective’ signal processing in the auditory system. I. Model structure,” J. Acoust. Soc. Am. **99**, 3615-3622.

Dau, T. and Pueschel, D. (1996b). “A quantitative model of the ‘effective’ signal processing in the auditory system. II. Simulations and measurements," J. Acoust. Soc. Am. **99**, 3623-3631.

Davis, S.B. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, and Sig. Proc., **28**, 357-366.

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.

Delgutte, B. (1996). "Physiological Models for Basic Auditory Percepts," in *Auditory Computation* (eds. H.L. Hawkins, T.A. McMullen, A.N. Popper, and R.R. Fay), Springer, New York, pp. 157-220.

Deng, L., and Geisler, C.D. (1987). "A composite model for processing speech sounds," J. Acoust. Soc. Am. **82**.6, 2001-2012.

Deng, L., and Kheirallah, I. (1993), "Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model," IEEE Trans. on Biomed. Engin., **40**.5, 456-467.

Deng, L. (1994). "Integrated optimization of dynamic feature parameters for hidden Markov modeling of speech." IEEE Signal Processing Letters, **1**.4, 66-69.

Deng, L, Aksmanovic, M., Xiaodong Sun, and Wu, C.F.J. (1994), "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," IEEE Transactions on Speech and Audio Proc., **2**.4, 507-527.

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**, 43P-44P.

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

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

Foote, J.T., Mashao, D.J., and Silverman, H.F. (1993), “Stop classification using DESA-1 high resolution formant tracking,” Proceedings of IEEE ICASSP, **2**.5, 720-723.

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

Gales, M.J.F. and Young, S.J. (1996). “Robust continuous speech recognition using parallel model combination,” IEEE Trans. on Speech and Audio Proc., **4**.5, 352-359.

Geisler, C.D. (1998). *From sound to synapse: Physiology of the mammalian ear*, Oxford University Press, New York.

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.

Glasberg, B.R. and Moore, B.C.J.(1990). "Derivation of auditory filter shapes from notched-noise data," Hearing Research, **47**, 103-138.

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.

Goldstein, J.L. and Srulovicz, P. (1977), "Auditory-Nerve spike intervals as an adequate basis for aural frequency measurement,” in *Psychophysics and Physiology of Hearing* (Eds. E. F. Evans and J. P. Wilson), Academic Press, London, 337-347.

Green, D.M. and Swets, J.A. (1966). Signal Detection Theory and Psychophysics. John Wiley and Sons, Inc. New York.

Hanson, H.M., Maragos, P., and Potamianos, A. (1994). “A system for finding speech formants and modulations via energy separation,” IEEE Trans. on Speech and Audio Proc., **2**.3, 436-443.

Hant, J.J., Strope, B., and Alwan, A. (1997). “A psychoacoustic model for the noise masking of plosive bursts,” J. Acoust. Soc. Am., **101**.5, 2789-2802.

Hant, J.J., Strope, B., and Alwan, A. (in press). “Variable-duration notched-noise experiments in a broadband noise context,” J. Acoust. Soc. Am.

Harris, D.M., and Dallos, P. (1979). “Forward Masking of Auditory Nerve Fiber Responses,” J. of Neural Phys., **42**.4, 1083-1107.

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

Hermansky, H., and Morgan, N. (1994). “RASTA Processing of Speech," IEEE Trans. on Speech and Audio Proc., **2**, pp. 578-589.

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.

Houtgast, T. (1989). “Frequency selectivity in amplitude-modulation detection,” J. Acoust. Soc. Am., **85**.4, 1676-1680.

Jankowski, C. R. Jr., Vo, Hoang-Doan H., and Lippman, R. P. (1995). “A Comparison of Signal Processing Front Ends for Automatic Word Recognition,” IEEE Trans. Speech and Audio Proc., **3**.4, 286-293.

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.

Joris P.X., Carney, L.H., Smith, P.H., and Yin, T.C.T. (1994). "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Response to tones at the characteristic frequency," J. Neurophys. (Bethesda), **71**, 1022-1051.

Juang, B.H. Rabiner, L.R. and Wilpon, J.G. (1987). “On the use of bandpass liftering in speech recognition,” IEEE Trans. Acoust., Speech, Signal Processing, **35**.7, 947-954.

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.

Kewley-Port, D. and Watson, C.S. (1994), "Formant-frequency discrimination for isolated English vowels," J. Acoust. Soc. Am. **95**.1, 485-496.

Kidd, G. Jr., and Feth, L.L. (1982). “Effects of masker duration in pure-tone forward masking,” J. Acoust. Soc. Am. **72**, 1364-1386.

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.

Kopec, G. (1986). “Formant tracking using hidden Markov models and vector quantization, IEEE Trans. Acoust., Speech, Sig. Proc., **34**.4, 709-729.

Langner, G. (1992). “Periodicity coding in the auditory system,” Hearing Res., **60**, 115-142.

Laprie, Y., and Berger, M.-O. (1994), “A new paradigm for reliable automatic formant tracking,” Proceedings of IEEE ICASSP, **2**, 201-204.

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.

Licklider, J.C.R. (1951). “A Duplex Theory of Pitch Perception,” Experientia **7**, 128-134.

Lippmann, R. P. (1997). “Speech recognition by machines and humans,” Speech Communication, **22**.1, pp. 1-15.

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. (1984), “Computational models of neural auditory processing,” in Proc. IEEE ICASSP, San Diego, 36.1.1-4.

Lyon, R. F., and Mead, C. (1988). “An Analog Electronic Cochlea,” IEEE Trans. on

Acoust., Speech, and Sig. Proc. **36**, 1119-1133.

Maragos, P., Kaiser, J.F., and Quatieri, T.F (1993). “On amplitude and frequency demodulation using energy operators,” IEEE Trans. on Signal Proc., **41**.4, 1532-1550.

McCandless, S.S. (1974). “An algorithm for automatic formant extraction using linear prediction spectra,” IEEE Trans. on Acoustics, Speech and Signal Proc., **22**.2, 135-141.

Meddis, R. and Hewitt, M.J. (1991a), “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,” J. Acoust .Soc. Am. **89**.6, 2866-2882.

Meddis, R. and Hewitt, M.J. (1991b), “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity,” J. Acoust. Soc. Am. **89**.6, 2883-2894.

Meddis, R. and O’Mard, L. (1997), “A unitary model of pitch perception,” J. Acoust. Soc. Am. **102**.3, 1811-1820.

Moore, B.C.J. (1973). “Frequency difference limens for short duration tones,” J. Acoust. Soc. Am., **54**, 610-619.

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.

Niranjan, M., Cox, I.J., and Hingorani, S. (1994). “Recursive tracking of formants in speech signals,” Proceedings of IEEE ICASSP, **2**, 205-208.

Oxenham, A.J. and Moore, B.C.J. (1994), “Modeling the additivity of nonsimultaneous masking,” Hearing Res., **80**, 105-118.

Oxenham, A.J. and Plack, C.J. (1996) “Peripheral origins of the upward spread of masking,” J. Acoust. Soc. Am., **99**, 2542.

Palmer, A.R. and Russel, I.J. (1986). “Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor ptoential of inner hair cells,” Hearing Res., **24**, 1-15.

Patterson, R.D. (1976). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. **50**, 1123-1125.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992), “Complex Sounds and Auditory Images,” in Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing, June 1991, (Eds. Y. Cazals and K. Horner), Pergamon Press, Oxford, pp. 429-446.

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.

Pirello, K. Blumstein, S.E., and Kurowski, K. (1997). “The characteristics of voicing in syllable-initial fricatives in American English,” J. Acoust. Soc. Am. **101**.6, 3754-3765.

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

Potamianos, A. and Maragos, P. (1996). “Speech formant frequency and bandwidth tracking using multiband energy demodulation,” J. of Acoust. Soc. Am., **99**.6, 3795-3806.

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

Rao, P. (1996). “A robust method for the estimation of formant frequency modulation in speech signals,” Proceedings IEEE ICASSP, **2**, 813-816.

Rigoll, G. (1986). “A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter,” Proceedings of IEEE-ICASSP, **2**, 1229-1232.

Schafer, R.W., and Rabiner, L.R. (1970). “System for automatic formant analysis of voiced speech," J. Acoust. Soc. of Am., **47**.2, 634-648.

Sellick, P., Patuzzi, R., and 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. (1988). "A joint synchrony/mean-rate model of auditory speech processing," J. Phonetics, **85**, 55-76.

Shannon, R.V. (1990). "A Model of Temporal Integration and Forward Masking for Electrical Stimulation of the Auditory Nerve," in *Cochlear Implants: Models of the Electrically Stimulated Ear* (Eds. J. M. Miller F. A. Spelman), 187-205.

Shannon, R.V. (1992). "Temporal modulation transfer functions in patients with cochlear implants," J. Acoust. Soc. Am., **91**, 2156-2164.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygnoski, J., and Ekelid, M. (1995). "Speech Recognition with Primarily Temporal Cues," Science, **270**, 303-304.

Siebert, W.M. (1968). "Stimulus transformation in the peripheral auditory system," in *Recognizing Patterns,* (eds. P.A. Kollers and M. Eden), MIT Press, Cambridge, pp. 104-133.

Smith, R.L., and Zwislocki, J.J. (1975). "Short-Term Adaptation and Incremental Responses of Single Auditory-Nerve Fibers," Biol. Cybernetics **17**, 169-182.

Stevens, S.S. (1956). "The direct estimation of sensory magnitudes--loudness," Am. J. Psychol. **69**, 1-25.

Stevens, K.N., Blumstein, S.E., Glicksman, L., Burton, M., and Kurowski, K. (1992). “Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters,” J. Acoust. Soc. Am., **91**.5, 2979-3000.

Strickland, E.A., and Viemeister, N.F. (1996). “Cues for discrimination of envelopes,” J. Acoust. Soc. Am., **99**.6, 3638-3646.

Strickland, E.A., and Viemeister, N.F. (1997). “The effects of frequency region and bandwidth on the temporal modulation transfer function,” J. Acoust. Soc. Am., **102**.3, 1799-1810.

Townsend, B., Cotter, N., Van Compernolle, D., and White, R.L. (1987). “Pitch perception by cochlear implant subjects,” J. Acoust. Soc. Am. **82**.1, 106-115.

Viemeister, N.F. (1979). “Temporal modulation transfer function based on modulation thresholds,” J. Acoust. Soc. Am., **66**, 1364-1380.

Viemeister, N.F. and Wakefield, G.H. (1991). “Temporal integration and multiple looks," J. Acoust. Soc. Am. **90**, 858-865.

Wang, K. and Shamma, S. (1994). “Self-Normalization and Noise-Robustness in Early Auditory Representations,” IEEE Trans. on Speech and Audio Processing, **2**.3, 412-435.

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

Wilson, B.S., Finley, C.C., Zerbi, M. and Lawson, D.T. (1994). *Speech processors for Auditory Prostheses*. Quarterly Progress Report No. **7**. Center for Auditory Prosthesis Research, Research Triangle Institute.

Woodland, P.C., Hain, T., Johnson, S.E., Nielser, T.R., Tuerk, A., and Young, S. (1998). "Experiments in broadcast news transcription," Proceedings of the IEEE ICASSP, **2**, pp. 909-912.

Zeng, Fan-Gang, (1994). “Loudness growth in forward masking: relation to intensity discrimination,” J. Acoust. Soc. of Am., **96**.4, 2127-2132.

Zeng, Fan-Gang, (1998). “Interactions of forward and simultaneous masking in intensity discrimination,” J. Acoust. Soc. of Am., **103**.4, 2021-2030.

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

Zwicker, E. (1970), “Masking and psychological excitation as consequences of the ear’s frequency analysis,” in *Frequency Analysis and Periodicity Detection in Hearing* (Eds. Plomp and Smoorenburg), Leiden, Sijthoh, 376-396.

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

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

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

Zwislocki, J.J. (1969). “Temporal summation of loudness: an analysis,” J. Acoust. Soc. Am., **46**, 431-441.