Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments

by

Brian E. D. Kingsbury

B.S. (Michigan State University) 1989

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA, BERKELEY

Committee in charge:

Professor Nelson Morgan, Chair
Dr. Steven Greenberg
Professor John Wawrzynek
Professor David Wessel

Fall 1998

The dissertation of Brian E. D. Kingsbury is approved:

Chair Date

Date

Date

Date

University of California, Berkeley

Fall 1998

Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments

Abstract

Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments

by

Brian E. D. Kingsbury
Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Nelson Morgan, Chair

Natural, hands-free interaction with computers is currently one of the great unfulfilled promises of automatic speech recognition (ASR), in part because ASR systems cannot reliably recognize speech under everyday, reverberant conditions that pose no problems for most human listeners. The specific properties of the auditory representation of speech likely contribute to reliable human speech recognition under such conditions. This dissertation explores the use of perceptually inspired signal-processing strategies—critical-band-like frequency analysis, an emphasis of slow changes in the spectral structure of the speech signal, adaptation, integration of phonetic information over syllabic durations, and use of multiple signal representations for recognition—in an ASR system to improve robustness to reverberation. The implementation of these strategies was optimized in a series of experiments on a small-vocabulary, continuous speech recognition task. The resulting speech representation, called the modulation-filtered spectrogram (MSG), provided relative improvements of 15–30% over a baseline recognizer in reverberant conditions, and also outperformed the baseline in other acoustically challenging conditions. The MSG and baseline recognizers may be combined to obtain more accurate recognition than is possible with either recognizer alone. Preliminary tests with the Broadcast News corpus indicate that the MSG representation is useful for large-vocabulary tasks as well.

Professor Nelson Morgan
Dissertation Committee Chair

For Linda and for my parents.

Contents

List of Figures

List of Tables

Acknowledgements

Although my name stands alone on the title page of this dissertation, the work it describes was possible only within a larger community of researchers and friends to whom I am deeply indebted.

I would first like to thank my advisor, Nelson Morgan, for his guidance and support over the course of my research. Through much effort (and consumption of large numbers of antacid tablets) he has assembled a top-notch group of speech researchers at ICSI and kept it going for over ten years now. His sense of humor and relaxed management style have made him a joy to work with. It's not going to be easy to leave!

Steven Greenberg has played a crucial role in my development as a speech scientist. Many of the ideas explored in this dissertation originated with him. He has instilled in me an appreciation of the extensive literature on psychoacoustics and auditory neurophysiology, and his careful reading of this dissertation as a member of my committee has greatly enhanced its clarity and flow.

I thank John Wawrzynek for serving on my thesis committee and for advising me in my first few years at Berkeley. His VLSI design class, which he talked me into taking in my first semester as a CS graduate student, was a particularly valuable experience. It reminded me that large-scale engineering projects can be great fun—an important lesson after a few too many dull, undergraduate-level classes.

Thanks to David Wessel for pointing me at Neil Todd's work on the modeling of rhythm perception and for serving on my thesis committee.

I have benefited greatly from my collaboration with Hynek Hermansky and Carlos Avendaño and their willingness to share wisdom, experimental results, and room impulse responses.

The development, care and feeding of a state-of-the art ASR system is too large a task for any one individual. My research would not have been possible without the efforts of current and former members of the ICSI Realization Group, a company of researchers as talented and supportive as I ever could have hoped for. Thanks to Eric Fosler-Lussier, my friend and workout partner, who provided valuable feedback on this dissertation and who has convinced me that there really are interesting things going on beyond the acoustic model; to Su-Lin Wu, whose work on syllable-based speech recognition provided an important testbed for my speech representations; to Nikki Mirghafori, organizer of the "lunch bunch" (a.k.a.

"the dissertation support group"); to Dan Ellis and Adam Janin for their extensive work on Broadcast News and their willingness to answer all of my questions about it; to Dan Gildea for his work on the Numbers 95 lexicon; to Jeff Bilmes and Mike Shire for their comments on my dissertation; and to Toshihiko Abe, Takayuki Arai, Joy Hollenback, Dan Jurafsky, Katrin Kirchhoff, Yochai Koenig, Warner Warren, Chuck Wooters and Geoff Zweig.

At last count, the experiments described in this dissertation required over 800 neural-network trainings, all of which were done on the Spert-II system using the T0 processor. The extensive exploration of the front-end design space that I performed would have been impossible without this hardware accelerator. On a more personal note, my own participation in the T0 project was, quite possibly, my single best experience at Berkeley. Thanks to Krste Asanović, the principal architect of the T0 microprocessor and my office-mate for nearly eight years; to Bertrand Irissou, for his work on the schematic design and layout of the chip, and for his introduction of rubber dart guns into the group; to Jim Beck, ICSI staff engineer *extraordinaire*, whose good humor and ability to build just about anything (from Spert boards to test rigs to electric barriers against slugs and snails) were vital to the success of the project; and to David Johnson, who wrote the neural-network training software used in all my experiments, and who taught me that "temporary" software has a way of becoming permanent, so it's best to do it right the first time.

Thanks to Kathryn Crabtree for shepherding me through the bureaucratic maze at Berkeley, and to Renee Reynolds, Nancy Shaw, Devra Pollack, and Elizabeth Weinstein, whose hard work as ICSI support staff help to keep the whole enterprise running.

Finally, thanks to my wife, Linda, who proofread the entire first draft of this thesis and who took over all of the household chores in the last few months of my thesis-writing, and thanks to my parents, who nurtured my sense of curiosity as I was growing up and who never asked me how much longer I planned to stay in school.

Funding for my work has come from a number of sources, including an NSF Graduate Fellowship, NSF grant MIP-8922354, NSF grant MIP-9311980, NSF grant MIP-9612778, a European Community Basic Research grant (Project SPRACH), and the International Computer Science Institute.

Chapter 1

Introduction

Automatic speech recognition has only recently emerged from the research laboratory as a viable technology. Currently, several companies are marketing document dictation software for desktop computers that can recognize tens of thousands of different words, and it is becoming more common to interact with automated systems over the telephone using speech-based interfaces with limited vocabularies. These two classes of recognizer mark opposite poles of the continuum of realizable automatic speech recognition (ASR) systems today. In order to reliably recognize speech, the large-vocabulary desktop systems rely on head-mounted close-talking microphones, a relatively quiet operating environment and considerable speaker adaptation. In contrast, the telephone-based systems can work reliably over a wide range of telephone channel conditions (although cellular telephones and speaker phones are still problematic) with minimal, if any, speaker adaptation; however, they must restrict the speech input by recognizing only isolated words instead of continuous speech, by limiting the vocabulary to a few thousand words, or by employing a constrained grammar. Significant obstacles still must be overcome to reach the ultimate goal of ASR, which is machine recognition of speech at levels comparable to human performance across the full range of possible speakers, vocabularies, and acoustic environments.

One of the key challenges in ASR research is the sensitivity of ASR systems to real-world levels of acoustic interference in the speech input. Ideally, a machine recognition system's accuracy should degrade in the presence of acoustic interference in the same way a human listener's would: gradually, gracefully and predictably. This is not true in practice. Tests on different state-of-the-art ASR systems carried out over a broad range of different vocabularies and acoustic conditions show that automatic recognizers typically commit at least ten times more errors than human listeners [Lip97]. ASR systems deployed in real-world applications must often be retrained on field data following their development in order to achieve intended levels of accuracy, even when their original training data were thought to adequately reflect field conditions [Tho97].

Acoustic interference can take many forms. The speech signal may contain extraneous sounds (additive noise) from the speaker's environment or the communication channel that transmits the speech to the recognizer, the signal may have some unknown spectral shaping or nonlinear distortion imposed on it by the microphone or communication channel, or the signal may include reverberation from the room in which the speaker is talking. Nor are these distortions mutually exclusive: the signal may be affected by all of them. The focus of this work is on the improvement of ASR accuracy in the presence of one specific form of acoustic interference — reverberation.

1.1 Reverberation

Reverberation is the name commonly given to the effect a room has on an acoustic signal produced within it. When speech or any other acoustic signal is produced in a room, it follows multiple paths from source to receiver. Some portion of the signal energy that reaches the receiver is transmitted directly through the air, while the remainder is reflected off of one or more surfaces in the room prior to reception. Usually the earliest reflections arrive discretely, while later reflections arrive in rapid succession or concurrently as the number of paths the sound may take increases. The reverberation process can be modeled as a convolution of the speech signal with a room impulse response. This model does ignore many effects, though. It ignores the fact that the characteristics of the transmission of sound from source to receiver can change significantly as the positions and orientations of the source and receiver vary [Mou85], as air currents in the room shift, and as objects change their positions (e.g., as doors open and close and as people move about). It also ignores the nonlinear properties of sound propagation within enclosures. Despite these shortcomings, the convolutional model is accurate enough to be useful for simulating many of the effects room reverberation and will be used in the present study.

Figure 1.1 illustrates the structure of a typical room impulse response. The important features of the impulse response are the initial direct response, the discrete early echoes and the reverberant tail, which is similar to exponentially decaying noise. The noise-like character of the tail is a consequence of the summation of a large number of transmission paths having different magnitudes and phases. The tail decays in an exponential manner because, with each reflection, some of the acoustic energy is absorbed by the reflecting surface. In a time-frequency representation the effect of reverberation is akin to a smearing of the signal along the time dimension, as illustrated in Figure 1.2. Reverberation can also alter the spectrum of the signal (even for a steady-state signal), as illustrated in Figure 1.3.

Spectral shaping (also known as spectral coloration) of sound is a linear, convolutional form of distortion. The characteristics of the receiver determine whether a convolutional distortion is better described as spectral shaping or reverberation. If the distorting impulse response distributes energy over a significantly longer duration than the temporal window of the receiver's spectral analyzer, the distortion is reverberant. If most of the distorting impulse response's energy falls within the temporal window of the receiver, the distortion is spectral shaping.

1.1.1 Characterizing Reverberation

The transmission of sound from a source to a receiver at fixed positions and orientations within a given room may be described by two parameters that are correlated with speech intelligibility, the reverberation time and the direct-to-reverberant energy ratio. The reverberation time, T_{60}, is the interval required for sound energy to decay by 60 dB after the sound source is turned off.[1] It may be based on a broadband measurement or it may be measured in restricted frequency bands, typically one octave in bandwidth. Because most surfaces reflect low-frequency acoustic energy more efficiently than high-frequency energy, and because the absorptive properties of air increase with frequency, the reverberation time is typically shorter at high frequencies than at low frequencies. Broadband reverberation time measurements are usually dominated by the low-frequency room response. Reverberation time is dependent upon the size of a room (smaller rooms typically have shorter reverberation times than larger rooms) and on the absorptive properties of the room surfaces. This can be seen by considering Sabine's approximation for computing reverberation

[1] In the room acoustics literature, the abbreviations $RT60$ and RT_{60} are also used.

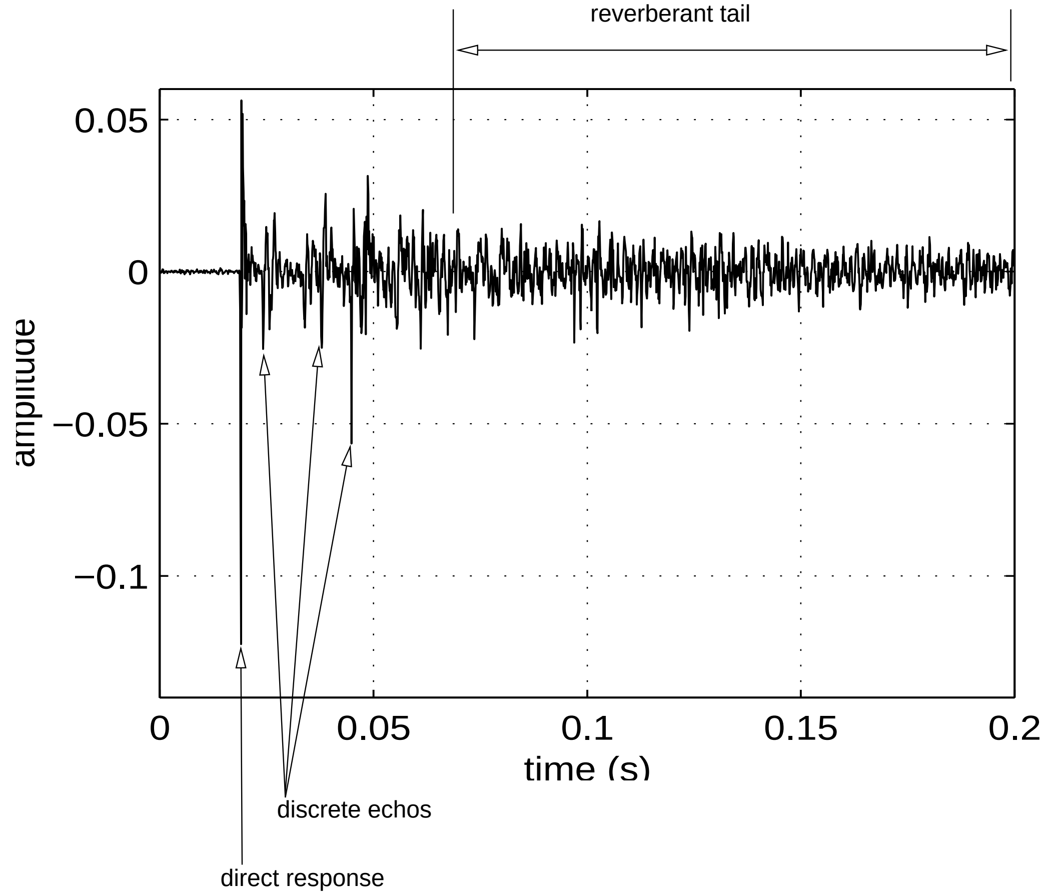


Figure 1.1: A typical room impulse response. The important features are the strong, initial response from the direct transmission path, a number of strong echoes in the first 100 ms of the impulse response (the strongest of which comes just before the 50 ms mark in this example), and the exponentially decaying reverberant tail of the response. It should be noted that the tail of the response has been truncated for clarity. The impulse response contains significant energy up to 0.9 s after the direct response.

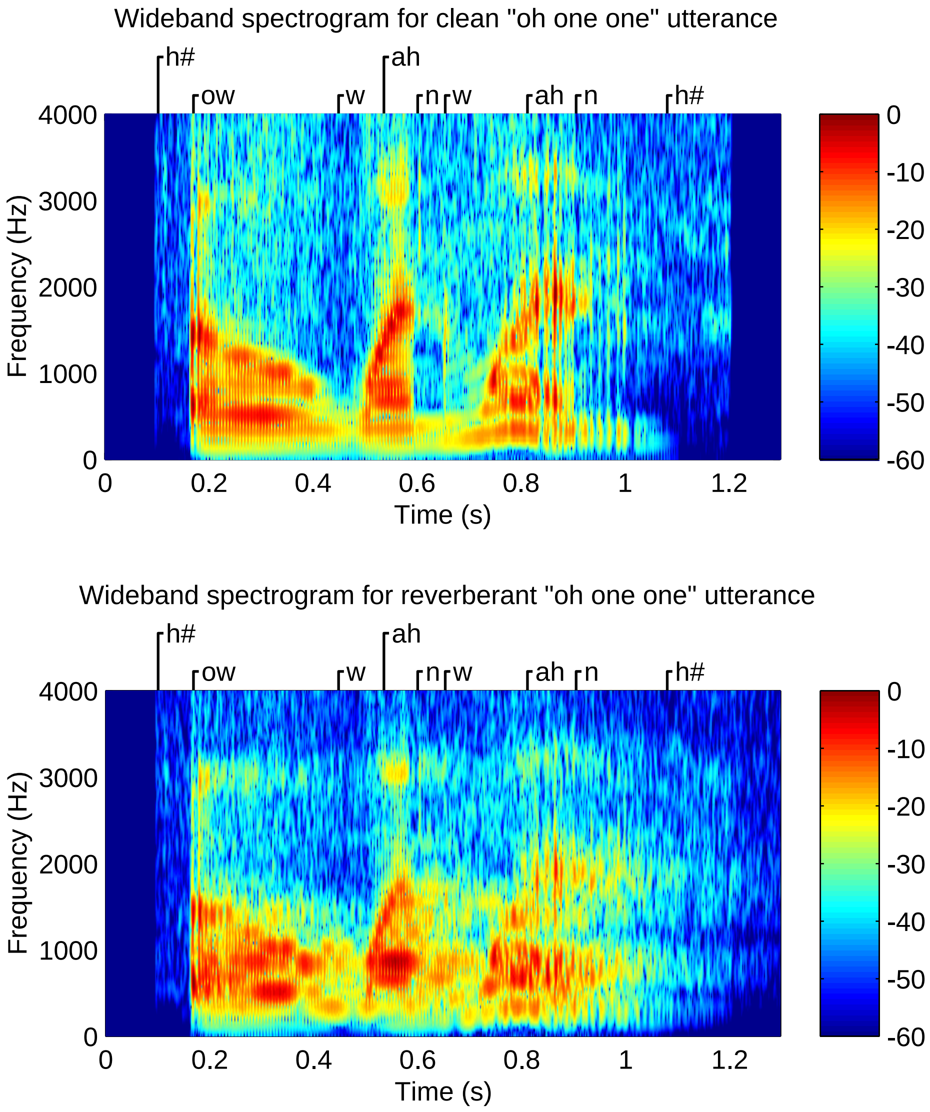


Figure 1.2: Wideband spectrograms for an adult female saying “oh one one” in clean and moderately reverberant conditions. The reverberant version of the utterance was generated by convolution with an impulse response characterized by a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 1 dB. The dominant effect of the reverberation is a temporal smearing of the signal, which is most evident in low-energy segments of the signal following high-energy segments (for example, the part between 0.6 and 0.7 s above). The signals are pre-emphasized with a filter, $H(z) = 1 - 0.94z^{-1}$, prior to the computation of the spectrograms. The spectrograms are based on 256-point FFTs computed from 8-ms segments of the signal weighted by a Hamming window function, using a window step of 2 ms. The energy scale is in dB relative to the peak level of the signal and has a lower bound of -60 dB.

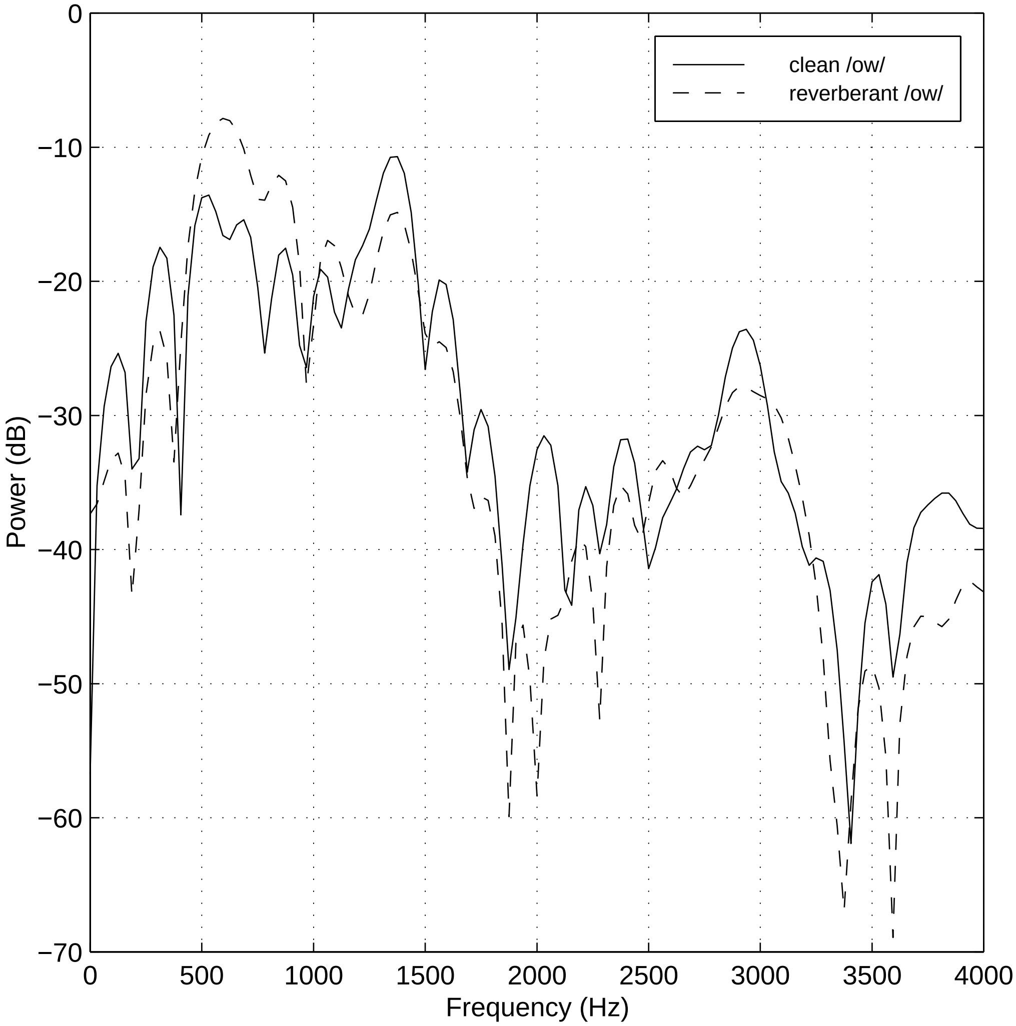


Figure 1.3: A comparison of short-time power spectra of the clean and reverberant signals portrayed in Figure 1.2. The plotted spectra are computed from 8-ms windows centered at 0.2 s. This time point is sufficiently early in the utterance that the major effect of the reverberation is in the form of spectral shaping rather than temporal smearing.

time [Sab22],

$$T_{60} \approx 0.163 \frac{V}{S\overline{\alpha}}$$

where V is the room volume in m^3, S is surface area of the room's walls, in m^2, and $\overline{\alpha}$ is the mean acoustic absorption coefficient of the room's walls. Typical reverberation times for average-sized offices are 0.4–0.6 s, while conference rooms usually have reverberation times of 0.8–1.2 s, and large auditoria may have reverberation times of 2 s or longer.

The direct-to-reverberant-energy ratio, which is usually expressed in decibels, is computed as

$$\frac{E_d}{E_r} = h(k_d)^2 / \sum_{i=k_d+1}^{k_{max}} h(i)^2$$

where $h(k)$ is the discrete-time room impulse response, k_d is the time of arrival for the direct sound, and k_{max} is the effective duration of the room impulse response, which is determined by the recording conditions and the noise floor of the measurement system. This ratio drops as the distance between speaker and receiver increases. For a given room and speaker position, the distance from the speaker at which the direct-to-reverberant energy ratio drops to 0 dB is called the critical distance. Critical distances of 0.5–1 m are typical of offices and conference rooms.

1.1.2 Performance of Human Listeners in Reverberation

Reverberation degrades speech recognition accuracy for human listeners. The degree of degradation increases with increasing reverberation time and decreasing direct-to-reverberant energy ratios. Monaural listening tests on young adults with normal hearing using relatively low-predictability speech material (words from the Modified Rhyme Test [HWHK65] embedded in a constant carrier phrase) show that recognition accuracy degrades from 99.7% correct for a reverberation time of 0.0 s (anechoic conditions) to 97.0%, 92.5%, and 88.7% correct for reverberation times of 0.4 s, 0.8 s and 1.2 s, respectively [NR82]. The ratio of direct to reverberant energy was not specified in these experiments. These levels of accuracy are sufficiently high to ensure that more natural, redundant speech material will be recognized reliably in everyday conditions. Binaural listening improves recognition accuracy in the presence of reverberation somewhat [MD67, NR82], while recognition accuracy decreases for children and elderly listeners [NR82], for fluent non-native listeners [ND84] and for hearing-impaired listeners [NP74].

Thus, reverberation reduces the intelligibility of speech for human listeners, but the impact of reverberation is usually not severe for natural speech materials presented to unimpaired listeners in typical environments. As will be shown in the next section, reverberation presents a much greater challenge for reliable automatic speech recognition.

1.1.3 Performance of ASR Systems in Reverberation

There are relatively few published data on the performance of ASR systems in the presence of reverberation, but the available data show that reverberation significantly reduces the accuracy of automatic recognizers. One recent study [GOS96] reports recognition results for simulated room reverberation with T_{60} ranging from 0.1–1 s using either a single omnidirectional microphone or an array of four omnidirectional microphones located 1.5 m from the speaker. The recognizer used state-of-the-art techniques: continuous-density hidden Markov models (HMMs) and a front end that produced eight mel-cepstral coefficients normalized via cepstral mean normalization, a normalized log-energy measure as well as first- and second-order temporal derivatives of all features. The system was trained on a clean set of phonetically diverse, Italian utterances collected with a close-talking microphone from both male and female speakers. The single-microphone results under simulated reverberant conditions show that recognition accuracy degrades from around 80% of words correct for $T_{60} = 0.1$ s to around 50% correct for $T_{60} = 0.3$ s, and to around 10% correct for $T_{60} = 0.5$ s. When an MAP re-estimation procedure [GL94] was used to perform HMM adaptation by adjusting the means of the Gaussian mixture components, system performance in reverberation improved to about 40% correct for $T_{60} = 0.5$ s. However, human listeners maintain an accuracy of better than 90% correct for reverberation times up to 0.8 s on more difficult test material.

Similar results were obtained in another study [San94, SG95] that compared the performance of recognizers using either a mel-cepstral front end or an auditory front end (the ensemble interval histogram (EIH) [Ghi86]) for the classification of a reduced set of phones in TIMIT utterances that had been downsampled to 8 kHz. The performance of the recognizers, which had been trained only on unreverberated utterances, was measured under both clean conditions and simulated room reverberation with a $T_{60} < 0.35$ s. A simplified classification task, in which the recognizer was provided with the locations of phone boundaries taken from hand transcriptions of the utterances, was used in order to simplify the analysis of the results by eliminating phone insertions and deletions. Although recognition performance using features from either the mel-cepstral or EIH front end (supplemented with first- and second-order differential features) was reasonably good for clean test data (phone classification accuracies of 66.2% and 57.6% for the mel-cepstral and EIH front ends respectively), performance under reverberant conditions was severely degraded (phone classification accuracies of 18.7% for the mel-cepstral front end and 17.3% for the EIH front end).

1.2 Scope of This Thesis

The goal of this thesis is to demonstrate that the performance of ASR systems in the presence of reverberation may be improved by making them more robust under reverberant conditions. An ASR system is robust if it can perform well in the presence of acoustic interference not represented in its training data. The approach explored in this work is the development of new signal-processing algorithms for the recognizer front end that are based on properties of human speech perception, are applicable to single-channel speech data and do not attempt to explicitly learn and invert the room impulse response.

The perceptual approach employed in the current work assumes that the reliability of human speech recognition is attributable, at least in part, to the characteristics of the auditory representation of speech, and that the use of similar representations in ASR systems may improve their reliability as well. Thus, the signal-processing strategies examined here were chosen because they are similar to those employed by the human auditory system for speech perception or because they are similar to those employed in the auditory systems of other organisms whose auditory processing is presumed to be similar to that of humans.

Although examination of human speech perception can suggest signal-processing strategies worth exploring, the details of the signal processing cannot be based solely on perceptual knowledge. Often, the available perceptual data are not sufficiently complete to provide all the necessary details. Also, the front-end signal processing must be compatible with the algorithms used by the ASR system. Therefore, the detailed implementation of the signal processing was guided by the results of automatic speech recognition experiments. The resulting algorithms are not intended to serve as detailed models of auditory processing. Instead, they follow only the general strategies employed by the auditory system for the robust representation of speech.

Only single-channel speech is considered in this study. While human listeners do recognize reverberant speech more accurately with binaural presentation than with monaural presentation, the "binaural advantage" in reverberation is not very large under most circumstances [NP74], and human monaural performance in reverberation is much better than ASR system performance under similar conditions. Moreover, a signal-processing method that improves the robustness of ASR systems for single-channel input could potentially be combined with multiple-microphone, beam-forming algorithms to achieve even greater robustness to reverberation.

The algorithms explored in the current work are "blind" in the sense that they do not attempt to determine or model the room impulse response and suppress the effect of reverberation via inverse filtering. Dereverberation of signals via inverse filtering is an extremely difficult task because of the non-stationary nature of the room response, because of the large number of parameters that must be estimated to properly characterize the room response and because room impulse responses are typically not minimum-phase and thus are not invertible using causal processing methods [NA79]. In addition, the extraction of reverberation-robust features for ASR that primarily describe the linguistic content of the speech signal may be a simpler task than the complete dereverberation of the speech signal because the extraction of the ASR features discards much of the spectro-temporal detail.

1.3 Overview

The rest of this thesis proceeds as follows. Chapter 2 provides an overview of the signal-processing strategies employed by the human auditory system that may contribute to the reliability of human speech perception in acoustically challenging conditions. Particular emphasis is given to temporal factors in the auditory processing of speech, such as adaptation and sensitivity to slow changes in the speech spectrum over time, because they may play a significant role in the robust auditory representation of speech. An overview of automatic speech recognition technology is provided in Chapter 3, with a focus on the speech-recognition system used in this work. A review of some current temporal processing approaches to robust speech recognition is also presented. Chapter 4 describes experiments on the visual display of speech and the automatic recognition of speech with a simple signal-processing system that uses some of the strategies described in the overview of human auditory processing of speech. Chapter 5 presents a series of experiments that led to improvements in the perceptually inspired signal processing, yielding both better recognizer performance under clean and reverberant conditions and a signal-processing system that operates on-line. The applicability of the signal-processing system developed in Chapter 5 to different acoustic conditions and to a different recognition task is tested in Chapter 6. Finally, Chapter 7 summarizes the work done in this thesis and discusses the broader lessons that may be learned from it.

Chapter 2

Speech Recognition by Humans

Signal processing in the human auditory system is complex, adaptive, and highly nonlinear, and it supports a wide variety of functions in addition to speech recognition (e.g., localization of sound sources and characterization of sources in terms of pitch and other parameters). Thus, results from perceptual and physiological studies of the auditory system must be interpreted with caution if they are to be applied to automatic speech recognition. The most applicable results are those obtained using speech stimuli or stimuli with speech-like spectral and temporal characteristics as test signals. Tests using simple pure tone or noise stimuli may not accurately predict auditory processing of speech signals. Likewise, results using experimental tasks that require signal identification are more likely to be relevant than results obtained with signal detection or discrimination tasks.[1]

From an engineering standpoint, the most immediately useful perceptual results may be those that indicate which aspects of the speech signal carry relatively little linguistic information and may therefore be excluded from the signal representation provided to an ASR system [Her97]. Such results may enable the development of representations with lower dimensionality and lower levels of variability attributable to nonlinguistic factors, leading to more reliable and robust ASR systems.

The potential benefits of perceptual processing in ASR are illustrated by considering the use of front-end speech representations having human-like frequency resolution.

[1] Detection, discrimination, and identification are all standard psychophysical tasks. In a detection task, subjects are asked to determine whether or not a test stimulus is present. In a discrimination task, subjects are asked to determine whether or not two test stimuli are different. In an identification task, subjects are asked to state the identity of test stimuli.

Many early ASR systems employed representations of the speech signal having a constant frequency resolution (a linear frequency scale). In contrast, the frequency resolution of the human auditory system decreases with increasing frequency. When speech representations with frequency resolution similar to that of humans were used, recognizer performance improved [DM80]. The representations with human-like frequency resolution give relatively precise locations for the first, second, and sometimes third formants (spectral prominences corresponding to resonances of the vocal tract) in the signal, while the locations of the higher formants are estimated with less precision. For many, but not all, voiced speech sounds, the location of the first two (or three, in some cases) formants depends strongly on the speech sound being produced, while the locations of higher formants are more strongly influenced by the geometry of the vocal tract of the individual speaker. The use of human-like auditory frequency resolution improved speaker-independent recognizer performance by enhancing the representation of a more speech-dependent aspect of the signal (the locations of the first and second formants) and blurring the representation of more speaker-dependent properties of the signal (such as the higher formants).

The linguistic content of the speech signal is encoded in both its spectral and temporal structure. Thus, the processing of speech in both frequency and time by the human auditory system may suggest signal-processing strategies useful for ASR. This chapter reviews the perceptual evidence for the various signal-processing strategies explored in this study. In the spectral domain, only the frequency resolution of the human auditory system will be considered, while three temporal processing schemes will be discussed: selectivity for changes in the spectral structure of incoming signals that occur at rates characteristic of speech, automatic gain control (adaptation) and integration of phonetic information over segments of the signal approximately 200–250 ms in duration. The use of multiple signal representations in perceptual processing will also be considered.

2.1 Frequency Resolution in Human Speech Perception

The processing of speech and other signals in the auditory system begins with a frequency analysis performed in the cochlea. This frequency analysis is preserved and enhanced in both the peripheral and central auditory systems, and may readily be measured via behavioral tests. The standard metric of auditory frequency resolution is the critical bandwidth: the minimum frequency separation for which components of a signal are processed in a relatively independent manner as derived from psychoacoustic measurements.

The term "critical bandwidth" was coined by Fletcher [Fle40] to describe the results of an experiment that measured the detection threshold for a sinusoidal probe tone in the presence of a variable-bandwidth, constant-power, bandlimited masking noise centered on the probe frequency. For any given masker power, the tone-detection threshold was constant for maskers narrower than the critical bandwidth, but decreased for masker bandwidths above this limit. The critical bandwidth may be interpreted as an estimate of the bandwidth of the highest-SNR frequency channel. As the bandwidth of the masker is increased, the SNR in the highest-SNR frequency channel remains constant and thus subjects' detection thresholds remain unchanged as long as the noise bandwidth is less than the channel bandwidth. Once the noise bandwidth exceeds the bandwidth of the highest-SNR channel, the SNR in that channel will increase as some of the noise power falls outside the channel passband, and subjects' performance on the task will improve.

Fletcher also noticed that the critical bandwidth was frequency dependent, with narrower bandwidths at low frequencies and broader bandwidths at high frequencies. Other researchers, using different experimental methods such as tone detection thresholds in notched (bandstop) noise [Pat76] or loudness summation [ZFS57], have obtained similar estimates of auditory frequency selectivity. Although the signals used in the measurement of critical bandwidths are simple, the measured critical bandwidths are applicable to speech. The masked intelligibility of speech whose spectral structure has been smoothed is not affected unless the smearing window exceeds the critical bandwidth [tKFP92].

Although measured human auditory filters behave in a nonlinear manner, with the slope of the low-frequency cutoff decreasing as the filter output level increases [RB94], auditory frequency analysis is most frequently modeled by a bank of linear, bandpass filters whose bandwidths increase with increasing frequency because the linear approximation is much simpler to implement.

Because of its importance for the modeling of auditory processing, the relationship between critical bandwidth and frequency has received considerable attention. Of the many auditory frequency scales that have been proposed, three scales are used in this study:

1. The Bark scale defines the critical bandwidth as 1 Bark, where the relationship be-

1/3-octave	$0.231f$
1/4-octave	$0.173f$
Bark	$0.167\sqrt{f^2+360000}$
Greenwood	$0.124f+20.6$

Table 2.1: Expressions for estimated critical bandwidth as a function of center frequency, f (in Hz), for two constant-Q scales, the Bark scale and Greenwood's cochlear frequency-position function.

tween frequency in Hz and Barks is designed to match published measurements as of 1960 [Zwi61]. The present study uses a mapping proposed by Schroeder [FAF$^+$77],

$$b = 6\sinh^{-1}\left(\frac{f}{600}\right)$$

where f is frequency in Hz and b is frequency in Barks.

2. Greenwood's cochlear frequency-position function [Gre61, Gre90] relates frequency in Hz to position along the cochlear partition in mm as follows:

$$f = A\left(10^{ax} - k\right)$$

where f is frequency, x is position, and the species-dependent parameter values are $a = 0.06$, $k = 1.0$, and $A = 165.4$ for humans. It is assumed that the critical bandwidth corresponds to equal spatial intervals along the cochlea of approximately 0.9 mm.

3. The logarithmic frequency scale approximates critical bandwidths by constant intervals in logarithmic frequency (a "constant-Q" approximation). One-third of an octave is a commonly used interval, while the current work uses one-quarter of an octave.

Figure 2.1 shows the relationship between estimated critical bandwidth and center frequency for four different scales of auditory frequency resolution—the Bark scale, Greenwood's function, a one-third octave scale and a quarter-octave scale—while Table 2.1 gives the corresponding mathematical expressions. Note that for all the scales the analysis bandwidth approaches a constant fraction of the center frequency as the center frequency becomes large. That is, all four scales are approximately logarithmic for a sufficiently large f. At low frequencies the Bark and Greenwood scales have approximately constant frequency resolutions, while the constant-Q scales are logarithmic for all center frequencies.

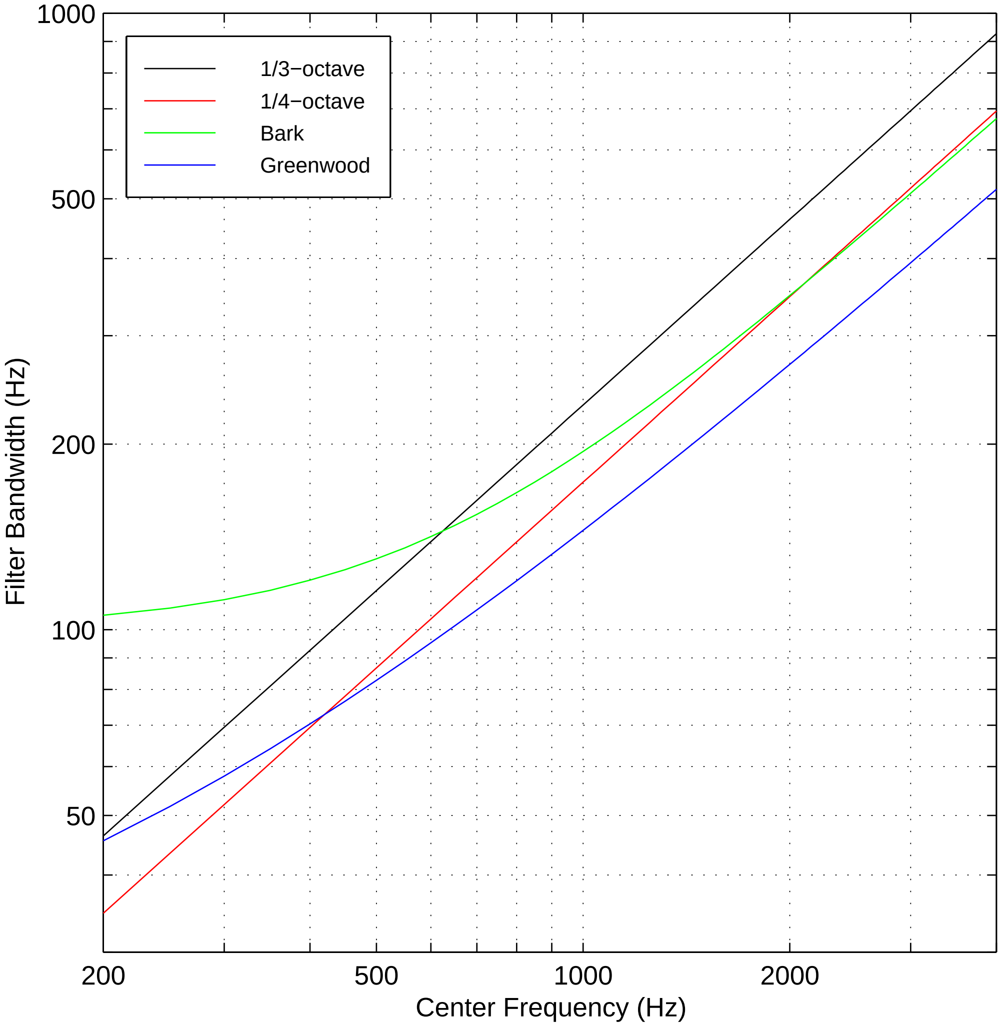


Figure 2.1: Plots of estimated critical bandwidth as a function of center frequency for two constant-Q scales, the Bark scale and Greenwood's cochlear frequency-position function.

Of the four scales reviewed, Greenwood's function is the best match to both physiological [Lib82] and recent psychoacoustic [Gre90, MG83] measurements of auditory frequency resolution.

2.2 Temporal Analysis in Human Speech Perception

2.2.1 The Importance of Slow Modulations for Speech Intelligibility

A key aspect of human speech perception that has important implications for speech processing, including the design of front-end feature extraction stages for ASR systems, is that the bulk of the linguistic information is encoded in relatively slow changes (below about 16 Hz) in the spectral structure of the speech signal.

Evidence for this perspective first emerged from the work of Homer Dudley and his colleagues at Bell Labs on the development of the channel vocoder. The channel vocoder models the production of the speech signal as the filtering of a source signal by a time-varying filter. The source may be either a periodic "buzz" signal that approximates the sound produced by the periodic opening and closing of the glottis during voiced speech segments, or it may be an aperiodic noise signal that approximates the hiss of air passing through a constriction in the vocal tract during production of unvoiced sounds. The time-varying filter models the influence of the vocal tract upon the source signal. The vocoder transmits speech by extracting parameters for the production model from a speech signal, transmitting those parameters to a receiver/synthesizer unit and then resynthesizing the speech signal from the parameters. Dudley and his colleagues found that they could obtain highly intelligible speech using filter control signals that had been lowpass filtered with a 25-Hz cutoff frequency and that the dominant frequencies in the filter control signals were 10 Hz and below [Dud39].

The Modulation Spectrum of Speech

The concept of spectral change over time has received a formal treatment in the study of room acoustics with the measurement of the modulation spectrum of speech and the characterization of sound propagation in rooms in terms of their modulation transfer functions [HS72, HS73]. The modulation spectrum, $|m(f)|$, is a characterization of the way

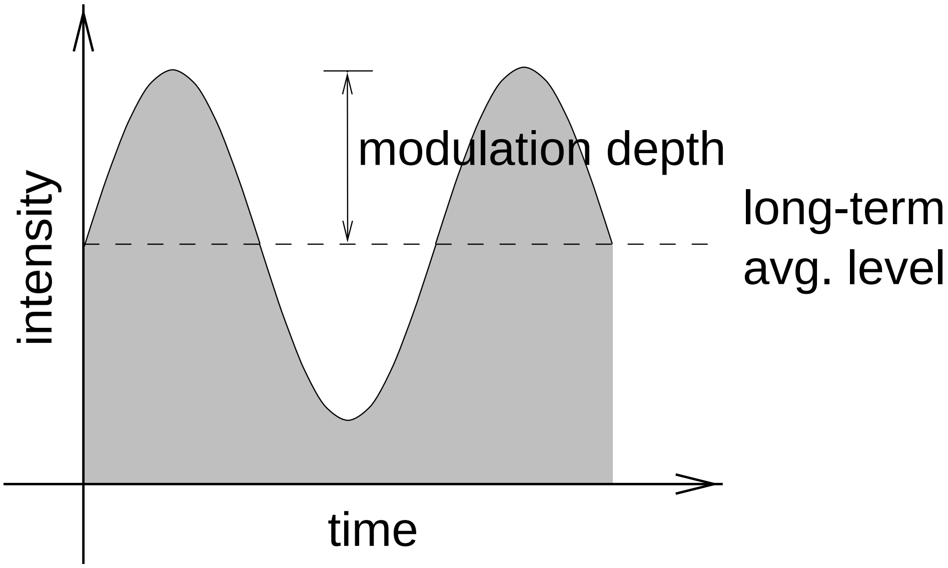


Figure 2.2: The modulation index is a measure of the change in a signal's energy over time that is computed by taking the ratio of modulation depth and the average level of a signal's energy envelope. A modulation index of 1 means that the dips in signal energy go all the way down to zero and the peaks go to twice the average energy level, while a modulation index of 0 means that the signal energy is constant.

a signal's energy changes over time. It is computed by performing a spectral analysis of the signal's energy envelope and normalizing by the average energy of the signal. That is, the modulation spectrum of a signal is computed as

$$|m(f)| = \frac{1}{\langle x(t) \rangle} \left| \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt \right|$$

where $x(t)$ is the energy envelope of the signal and $\langle x(t) \rangle$ is the average value of $x(t)$. The energy envelope of a signal may be computed by filtering the signal squared with a lowpass filter having a cutoff frequency just above the highest modulation frequency of interest. Note that the modulation frequencies measured must be well below the frequencies contained in the signal being analyzed. For all f, $0 \leq |m(f)| \leq 1$, so the modulation spectrum is an estimate of the modulation index (see Figure 2.2) of each sinusoidal component of a signal's energy envelope.

The modulation spectrum provides a statistical characterization of the temporal structure of a signal. For a very simple signal the modulation spectrum may be an exact characterization of the change in the signal over time; however, for more complex signals, such as speech, the modulation spectrum should be considered a description of *average* temporal structure. For example, the modulation spectrum of a sinusoid amplitude-modulated by the signal $x(t) = \sqrt{1 + m\cos(2\pi ft + \phi)}$, where f is the modulation frequency (in Hz), ϕ is the phase of the modulation, and m is the modulation index, is

$$|m(\omega)| = \delta(\omega) + \frac{m}{2}\delta(\omega - 2\pi f)$$

(if windowing effects and other details of the spectral analysis are ignored). A Gaussian noise signal with the same amplitude modulation has the same modulation spectrum, but the modulation spectrum only reaches this ideal in the limit, when it is averaged over an infinite number of sample points.

In a spectrally complex signal, such as speech, different frequency bands may behave in a relatively uncorrelated manner, such that the full-bandwidth energy envelope of the signal does not adequately reflect its temporal variation.[2] Thus, for spectrally complex signals, a modulation-spectral analysis is more appropriately applied to band-limited versions of the signal. One-octave bands are often used for speech. The modulations in these bands have two principal origins. The first is changes in the overall level of energy in the speech signal that correspond to the alternation between voiced segments, unvoiced segments and silence. The second is changes in the spectral distribution of speech energy, such as formant movements.

Figure 2.3 illustrates typical modulation spectra for speech. The modulation spectra are lowpass in form, with the roll-off beginning around 4 Hz.[3] The 4–8 kHz band is more strongly modulated over the range of modulation frequencies measured than are the other bands. The reason for this is that speech energy falls into the 4–8 kHz band more intermittently than it does into the other bands, with bursts and fricatives being the only speech sounds with significant energy in the 4–8 kHz range. The relatively high cutoff fre-

[2] For an extreme example, consider a signal synthesized by summing two noise bands which have the same level, do not overlap in frequency and which are each modulated by a square-wave signal. If the two modulating signals have the same frequency and opposite phases, then the energy envelope of the final signal will not fluctuate, but the signal itself will still have a distinct temporal structure.

[3] It should be noted that the modulation spectra in Figure 2.3 look quite different from those reported by Houtgast and Steeneken, who showed modulation spectra with a bandpass shape having a peak around 4 Hz. The reason for this difference is that Houtgast and Steeneken performed the spectral analysis of the energy envelope with a one-third-octave filterbank in which each filter had the same peak magnitude response. Such a filterbank does not produce a spectrum in the usual sense, that is, an estimate of energy density as a function of frequency, unless the output of each filter in the filterbank is weighted in inverse proportion to its bandwidth. This compensation was not performed, as is apparent in Figure 2 in [HS72], which shows the modulation spectrum for a one-octave band of Gaussian white noise as having a +3 dB/octave slope. In Figure 2.3, the modulation spectrum of a one-octave band of Gaussian white noise would be flat.

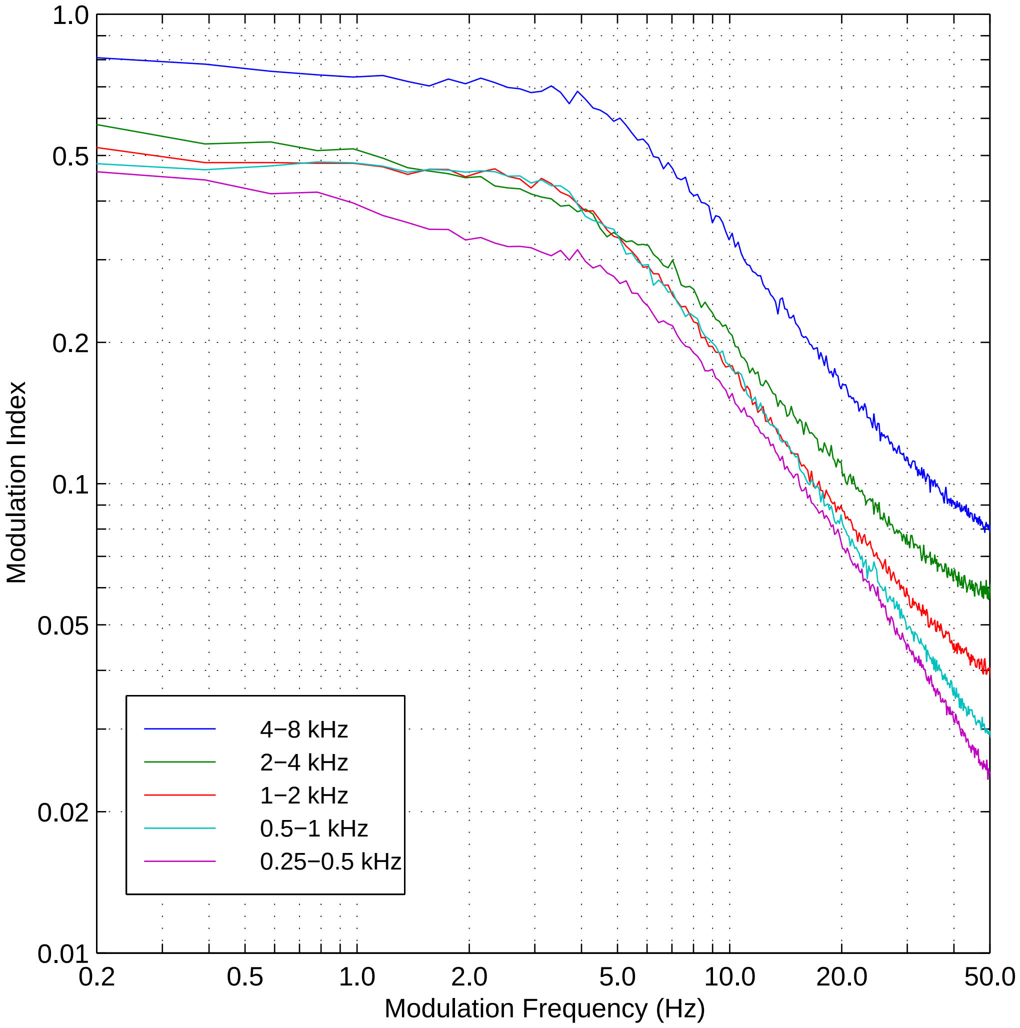


Figure 2.3: Modulation spectra for one-octave bands from 0.25–8 kHz, computed from a 206-s segment of speech taken from the Broadcast News corpus. The segment is from one female speaker giving a weather report.

quency for the modulation spectrum of speech distinguishes it from many common noise signals, which are not significantly modulated at rates above ca. 1 Hz.

The Modulation Transfer Function of a Channel

Because changes in the speech spectrum over time are responsible for conveying linguistic information to listeners, the fidelity with which a channel (such as a room) transmits these changes will determine the intelligibility of the transmitted speech. A channel may be characterized by its modulation transfer function, which is simply the ratio between the modulation spectrum of a transmitted signal and the modulation spectrum of the original input signal. The modulation transfer function (MTF) is a useful characterization of a channel because it concisely summarizes the effects of multiple forms of acoustic distortion that may be imposed on a signal by the channel.

For a noisy channel, the modulation transfer function, $|H_m(\omega)|$, is simply a constant attenuation factor independent of modulation frequency:

$$|H_m(\omega)| = \frac{i_s}{i_s + i_n}$$

where i_s is the average energy of the signal and i_n is the average energy of the noise. Nonlinear distortion may be modeled as signal-correlated noise with an expected level that is some proportion of the signal level that may differ across frequency bands [SH80].

A channel with an impulse response $h(t)$ has a modulation transfer function [Sch81]

$$|H_m(\omega)| = \left| \frac{\int_0^\infty h^2(t) e^{-j\omega t} dt}{\int_0^\infty h^2(t) dt} \right|$$

That is, the modulation transfer function is the magnitude of the Fourier transform of the squared impulse response divided by the total energy of the impulse response. Thus, for a room with an "idealized" impulse response of an exponentially decaying white noise with reverberation time, T_{60}, the modulation transfer function is lowpass:

$$|H_m(\omega)| = \left[1 + \left(\frac{\omega T_{60}}{13.8} \right)^2 \right]^{-\frac{1}{2}}$$

As illustrated in Figure 2.4, actual rooms have lowpass MTFs as well, but the modulation attenuation flattens for sufficiently high modulation rates. For combinations of distortions

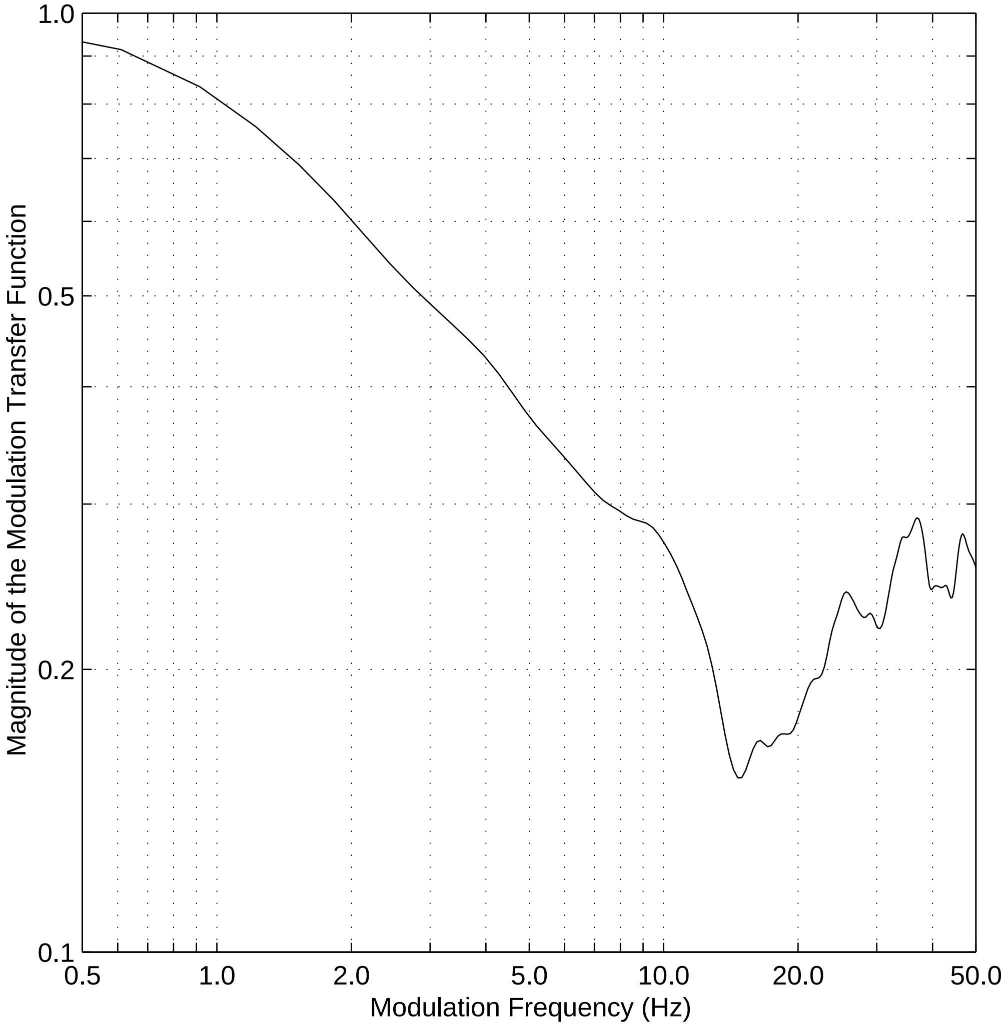


Figure 2.4: Modulation transfer function for the room impulse response illustrated in Figure 1.1. The impulse response is characterized by a T_{60} of 0.9 s and a direct-to-reverberant energy ratio of -9 dB.

(i.e., a noisy channel with impulse response, $h(t)$), the MTFs for the individual distortions combine multiplicatively.

The predictive power of the modulation transfer function for speech intelligibility has been demonstrated with the speech transmission index (STI) [SH80], a simple numerical score that summarizes the MTF of a channel for frequencies of 125–8000 Hz and modulation frequencies of 0.625–12.5 Hz. The STI has been shown to be highly predictive of speech intelligibility for reverberant and noisy auditoria tested with multiple languages [SH82] and for communication channels with noise, spectral shaping, and nonlinear distortion [SH80]. The broad applicability of the STI measure for predicting speech intelligibility is further evidence for the view that the slow changes in the speech spectrum carry the linguistic information.

Perceptual Studies on the Intelligibility of Temporally Smeared Speech

More direct evidence for the importance of slow changes in the speech signal for conveying linguistic information comes from recent perceptual experiments which measure the intelligibility of speech processed to suppress modulations above a given rate.

One such study [DFP94] used a vocoder-like analysis-synthesis system to study the intelligibility of temporally smeared Dutch sentences in noise and temporally smeared Dutch CVC and VCV syllables[4] in quiet. The temporal modification of the speech material was performed using the signal-processing system illustrated in Figure 2.5. The input speech signal was analyzed into separate frequency channels using a constant-Q FIR (finite impulse response) filterbank. In each channel an amplitude envelope was determined by computing the magnitude of the analytic signal associated with the filterbank output.[5] Each amplitude envelope signal is smoothed by processing it non-causally with a lowpass FIR filter. The filters have odd-length, symmetric impulse responses which are centered on the current sample such that, if the filter length is $2m + 1$, there is an m-point lookahead into the future. Next, a temporally smoothed version of the filterbank output is synthesized via pointwise multiplication of the original filterbank output and the ratio of the filtered and original amplitude envelopes. In effect, the division of the original filterbank output by the

[4] That is, consonant-vowel-consonant and vowel-consonant-vowel syllables.

[5] Given a real-valued signal $s(t)$, the analytic signal associated with $s(t)$ is $\tilde{s}(t) = s(t) + j\hat{s}(t)$ where $\hat{s}(t)$ is the Hilbert transform of $s(t)$.

original amplitude envelope flattens the filterbank output so that it may be given a smoothed amplitude envelope via pointwise multiplication by the filtered envelope signal. There was sufficient noise in the input speech (due to tape hiss) that the original amplitude envelope signal was always greater than zero. Finally, the speech is resynthesized by summing the smoothed subband signals and normalizing the sum to have the same level as the original input signal.

For tests of speech intelligibility in noise, the speech reception threshold (SRT) was measured for Dutch sentences processed with filterbanks having one-octave, half-octave, or quarter-octave filter bandwidths covering the frequency range 125–4000 Hz and with lowpass envelope filters having cutoff frequencies of 64, 32, 16, 8, 4, 2, 1, 0.5, or 0 Hz. For the 0-Hz case, each envelope signal was replaced by its average value for the utterance, which is equivalent to infinite peak-clipping of the signal with preservation of its average energy. The SRT is a standard measure of speech intelligibility and is defined as the signal-to-noise ratio for which 50% of the speech material is correctly recognized. In these tests, the masking noise has the same spectrum as the average spectrum of the test sentences.

For envelope filters having cutoff frequencies of 2 Hz or below, the processed sentences were not completely intelligible in quiet, precluding the use of the SRT measure. For these sentences, intelligibility in quiet was measured instead. Intelligibility was highest for the 2-Hz envelope filters and the one-octave filterbank, and decreased with decreasing cutoff frequency and filterbank bandwidth. For reasons explained below, the intelligibility of the sentences processed with the 0-Hz filters was quite high for some test conditions. The intelligibility of sentences processed with the one-octave filterbank was 90% with the 2-Hz envelope filter and dropped to 80% for the 0-Hz envelope filter, while the intelligibility of sentences processed with the quarter-octave filterbank was 20% with the 2-Hz envelope filter and dropped to under 5% for the 0-Hz envelope filter.

The comparatively high intelligibility of the sentences processed with the one-octave filterbank may be attributed to the relatively wide bandwidth of the spectral analysis compared to critical bandwidths, which are narrower than one-third of an octave over much of the audio frequency range (see Figure 2.1). Even if all amplitude fluctuations are eliminated for a one-octave bandwidth (as in the 0-Hz case above), changes in the amplitude of one-third octave bands within the one-octave bandwidth may remain.[6] As the analysis

[6] This may been seen by considering an octave-wide noise band synthesized by summing three non-

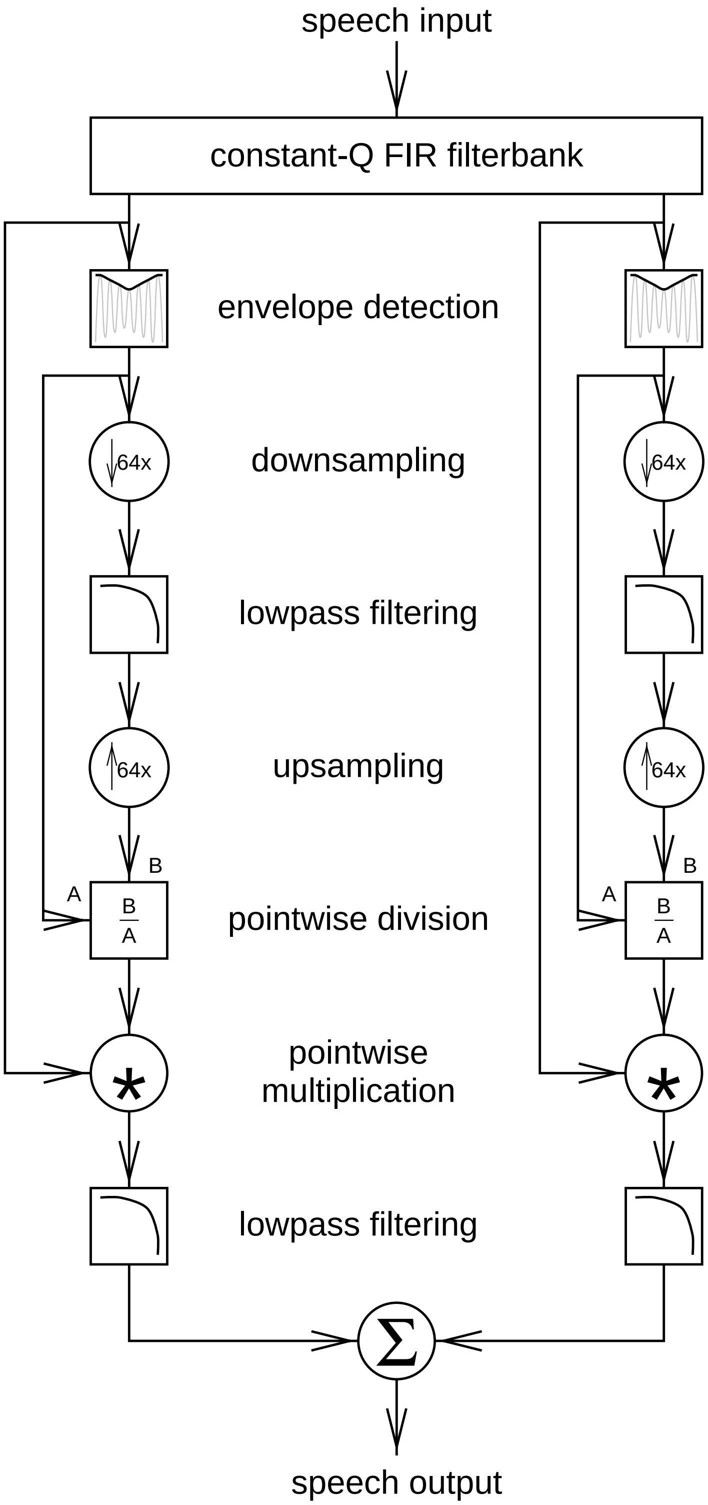


Figure 2.5: The vocoder-like signal processing system used by Drullman and colleagues to synthesize speech without high-frequency modulations.

bandwidth used for the temporal smearing approaches that of the human auditory system, however, the temporal fluctuations of the speech signal that carry linguistic information are obliterated by the smearing process.

For envelope filter cutoff frequencies of 4 Hz and above there was no significant effect of analysis bandwidth, although there was a significant effect of envelope filter cutoff frequency. For envelope filter cutoffs of 16 Hz and above, a performance ceiling was reached. Thus, the envelope fluctuations of 16 Hz and above did not contribute to speech intelligibility in noise. The SRT for the 8-Hz envelope filter was significantly higher than that for the 16-Hz filter, and the SRT of the 8-Hz envelope filter was significantly higher than that for the 4-Hz filter. Nevertheless, even with the 4-Hz envelope filter, speech intelligibility was quite good, with an SRT of 0 dB. This is only 5.6 dB higher than the SRT of unprocessed sentences.

Analysis of phone recognition scores for the processed CVC and VCV syllables demonstrated that the effects of the temporal smearing are akin to those of room reverberation. Vowels are recognized more reliably than consonants, with stops being more vulnerable than other consonants. Vowel confusions are predominantly diphthong-to-monophthong, long-to-short, and short-to-long confusions. Unlike reverberation, however, the recognition of medial and final consonants is as accurate as the recognition of initial consonants in the temporally smeared, isolated syllables. This difference is due to the non-causal filtering of the amplitude envelopes in the temporal-smearing process.

Similar results have also been obtained for Japanese V and CV syllables processed with an analysis-synthesis system that temporally smoothes the LPC cepstral coefficients of the input signal [APHA96]. In this system, the zero-order cepstral coefficient is not filtered, so in each analysis frame the input and output signals will have exactly the same total energy. Thus, the system will not suppress modulations caused by changes in the overall level of the speech signal, but will suppress modulations caused by changes in the spectral distribution of speech energy. Filtering cepstral coefficients instead of amplitude envelopes simplifies the study of highpass and bandpass filtering, because negative values for cepstral coefficients are valid.[7] As in the earlier study, it was found that changes in

overlapping, equal-energy noise bands that are one-third octave in width, where each band is modulated by a pulse train with a 33% duty cycle and the pulses in the three modulating pulse trains do not overlap in time. The energy of the one-octave-wide noise band will not fluctuate, but a finer frequency analysis will reveal a high degree of temporal structure in the signal.

[7] The application of highpass or bandpass filtering in the vocoder-like scheme employed in [DFP94] is

the spectral shape of the speech signal occurring at rates above 16 Hz are not required for speech intelligibility. It was also found that changes at rates below 1 Hz were not required.

More recent studies have highlighted the importance of the relative timing of these changes across frequency bands in conveying linguistic information. One set of experiments measured the intelligibility of utterances from the TIMIT corpus that were processed so only a few spectral "slits" remained [GAS98]. The speech material was filtered through an FIR filterbank into fourteen 1/3-octave-wide channels, and only the four channels covering 298–375 Hz, 750–945 Hz, 1890–2381 Hz, and 4762–6000 Hz were retained.

Test utterances were then synthesized by summing together the four channels with groups of one to three channels delayed with respect to the other channels by 25, 50, or 75 ms. If no delay was introduced, the intelligibility of the test utterances was 88.3%. As the delay increased, intelligibility decreased. The intelligibility of the utterances was between 70.7% and 80.4% for delays of 25 ms, between 53.6% and 65.0% for delays of 50 ms, and between 40.7% and 58.9% for delays of 75 ms. The degradation of speech intelligibility caused by the introduction of the delays indicates that the temporal relationship between the different channels carries crucial information.

There is an apparent contradiction between these results and the results of an earlier study [AG98] that measured the intelligibility of TIMIT utterances processed to desynchronize them across frequency. The utterances were filtered into nineteen frequency bands in which the lowest channel covered the 0–265 Hz range and the others were 1/4-octave in width. Test stimuli were synthesized by delaying the channels by randomly selected amounts between 0 ms and a variable maximum delay that ranged from 60 ms to 240 ms in 20-ms steps and then adding the delayed channels together. The delays were chosen so that any pair of adjacent frequency channels were shifted with respect to one another by more than one quarter of the maximum delay.

In this study it was found that the intelligibility of the stimuli was 80% or more for maximum delays of up to 140 ms. This result would seem to suggest that the relative timing of different frequency channels is not important for speech intelligibility, contradicting the more recent study. There may be no contradiction, however, because a re-analysis of the

problematic because negative amplitude envelope values, which correspond to negative values for signal energy, are meaningless and it is unclear how they should be handled. This difficulty complicates the interpretation of experiments with highpass and bandpass filtering using the vocoder-like processing.

stimuli used in [AG98] showed that they could contain sets of channels distributed across frequency for which there were relatively low degrees of desynchronization. If the auditory system is able to identify and use these synchronous sets, then the apparent contradiction is likely to be resolved.

Summary

The design parameters for the channel vocoder, the successful use of the temporal modulation transfer function and STI for predicting speech intelligibility across a wide range of rooms and communication channels, and the perceptual results for the intelligibility of temporally smeared speech all support the view that the primary carrier of linguistic information in the speech signal is changes in the spectral structure of speech occurring at rates between 1 and 16 Hz and that an adequate spectral resolution for characterizing these changes is the critical bandwidth.[8] Additional support for this account may be found in the study of speech production, where it has been shown that the motions of the lips and jaw during articulation may be described by a linear second-order system with a damped frequency between 2 and 12 Hz [SBMK93], and in neurophysiological measurements that have shown the existence of large populations of primary auditory cortical neurons selective for amplitude modulations at relatively slow rates: 4–15 Hz in laboratory rats [GO95] and under 20 Hz in cats [SU88].

2.2.2 Adaptation

Auditory adaptation also contributes to the reliable recognition of speech by human listeners in different acoustic environments. Fibers of the auditory nerve respond most strongly to signal onsets. Following onset, their firing rate gradually drops to a lower, steady-state level. Similar adaptive behavior is observed in more central auditory nuclei, such that at the level of the inferior colliculus (of anesthetized cats) strong responses to a speech signal are found only for syllable onsets and stop-consonant bursts [DHC99].

This adaptive processing is functionally similar to automatic gain control applied in a frequency-local manner, and it has important consequences for human speech perception.

[8] A very similar signal-processing architecture has been proposed as a model for rhythm and prosody perception [Tod94].

Most notably, speech intelligibility for human listeners is quite insensitive to the frequency response of the channel transmitting the speech. The speech reception threshold for Dutch sentences is constant for channels with spectrally sloped frequency responses over a range of slopes from -6 to +9 dB/octave. Moreover, listeners perform well even when the slope is time-varying, for rates of variation up to 1 Hz [vDAP87]. Both peripheral adaptation and more specific, central compensation mechanisms appear to contribute to the insensitivity of human listeners to static or slowly changing channel frequency responses [Wat91].

2.2.3 Perceptual Processing Time and Units of Recognition

The integration of phonetic information over time by the auditory system may also contribute to reliable human speech recognition. Most accounts of human auditory processing hypothesize a pre-perceptual auditory memory (sometimes called the "echoic memory") that stores a relatively detailed and unprocessed form of the recent auditory input. Such a store is believed to be necessary because of the transient, dynamic nature of auditory input. Many auditory judgments require integration of information over the recent past, and thus need some sort of buffering of the auditory input in order to function. The storage capacity of the pre-perceptual store is of interest because it may constrain the duration of the speech signal which may contribute directly to a phonetic judgment at a given time. A number of studies using different experimental protocols estimate the pre-perceptual auditory capacity at ca. 200–250 ms. These studies, which are reviewed below, include investigations of backwards recognition masking, measurements of the intelligibility of interrupted or temporally segmented speech, and studies on the vowel sequence illusion.

Backwards Recognition Masking

In a backwards recognition masking task, subjects are presented with a target stimulus, followed by a variable-duration silent interval and a long-duration masker stimulus. The subjects' task is to identify the target. Targets and maskers used in such tests include high and low pure tones with an intermediate-frequency pure tone masker, brief vowels with vowel or vowel-like maskers and synthesized English CV syllables with synthesized CV syllable maskers [Mas72, Mas74]. These studies have three results that are relevant to the question at hand:

1. Identification accuracy increases as the silent interval between the target and masker increases, up to a silent interval of roughly 150–250 ms. Beyond this duration, recognition accuracy reaches an asymptote.

2. The identity of the masker has only a small effect on recognition accuracy. Thus, for example, an [i] target may be masked as effectively by an [i] as by an [a].

3. If the silent interval is set to zero and the duration of the target is varied, recognition accuracy increases as the target duration increases, up to about 200 ms. Increments in the target duration beyond 200 ms provide no improvement in accuracy.

These results suggest that auditory processing of the target may require up to ca. 200 ms (based, perhaps, on a pre-perceptual auditory store with a 200 ms capacity) and that the processing of the target stimulus may somehow be disrupted by the presentation of a second stimulus within the 200-ms time window. The precise nature of this disruption cannot be deduced from these results, however. The presentation of the masker stimulus may halt processing of the target stimulus, or it may alter or overwrite the representation of the target in early auditory memory.

The Intelligibility of Interrupted or Temporally Segmented Speech

Studies on the intelligibility of interrupted speech [ML50], and temporally segmented speech [Hug75] shed light on the nature of the disruption found in the backwards recognition masking experiments. An interrupted speech signal is one in which portions of the signal are replaced by silence, so not all of the speech is heard. A temporally segmented signal is one into which silent gaps have been inserted. In a temporally segmented speech signal, all of the speech is heard. Studies of the intelligibility of these signals demonstrate a deleterious effect on the intelligibility of speech only when the duration of the speech fragments falls below 180 ms and the spacing of successive fragment onsets exceeds 180 ms. If the speech fragments are relatively short, but longer than about 33 ms, and if two or more fragments are heard every 180 ms, or if the speech fragments are at least 180 ms in duration, the speech is quite intelligible.

With these interrupted or segmented stimuli, listeners are able to integrate information from speech fragments separated by silent intervals, provided that the interval is short enough. This result makes it unlikely that backwards recognition masking occurs because processing of the target is terminated by presentation of the masker. If the presentation of the masker did indeed halt processing of the target, then with the interrupted or segmented speech stimuli one would expect an improvement in intelligibility for speech fragments separated by more than ca. 200 ms. Instead the converse is true: speech intelligibility improves when the fragments are separated by silent intervals less than 180 ms. It is therefore more likely that the auditory system integrates sounds over a window of ca. 200 ms, and that this obligatory integration interferes with target identification in the backwards recognition masking experiments, although it improves recognition accuracy in the interrupted and temporally segmented speech studies.

The Vowel Sequence Illusion

Given the backwards recognition masking results, it is (perhaps) not surprising that listeners presented with repeated sequences of six steady-state vowels are unable to report the identity and order of the vowels if their durations are under 100 ms. If the vowel durations are over 200 ms, however, the task is simple [THCG70, CEST77]. When listeners are presented with sequences in which the vowel durations are under 100 ms, they typically report hearing two simultaneous voices with different timbres each repeating a different series of two or three syllables [WHC96]. This effect is called the "vowel sequence illusion." The syllable sequences reported by a given listener for a given vowel sequence are stable for tests performed one week apart. Moreover, different subjects report hearing the same syllable sequences for the same vowel sequences if a 250-ms silence is inserted between repetitions of the vowel sequence, and even if no silence is inserted between sequence repetitions, subjects are able to match the illusory syllable sequences reported by other subjects to their corresponding vowel sequences.

Given that the mean duration of phones in conversational speech is 72 ms [GHE96], these results call into question the popular model of human speech recognition as a process that begins with the identification of phones in the speech signal. When presented with stimuli that have the spectro-temporal characteristics of speech but that are not interpretable as words, listeners hear syllables, not phones.

2.3 The Role of Multiple Representations in Perceptual Systems

A recurring motif in sensory systems is the use of multiple representations of the input as the basis for perceptual processing. This strategy is employed in a diverse array of organisms and sensory systems. Examples include

- the electrosensory system of gymnotiform fish, where the amplitude and phase of electric organ discharges are processed in separate loci prior to central integration [Hei89],
- the auditory system of the barn owl, which processes interaural intensity differences and interaural time differences in separate paths prior to integration in the inferior colliculus [TMK84], and
- the visual system of the macaque, in which form, motion, color, and stereoptic information are processed in a somewhat independent, parallel streams within a complex, hierarchical network of cortical centers [DE88, FE91].

Human speech perception is also, most likely, based on multiple representations of the speech signal. One way in which a multiplicity of representations may be generated is by processing frequency-limited bands of the speech signal separately. Such a model for human speech recognition was originally proposed by Fletcher to explain human recognition of nonsense syllables transmitted via a wide variety of channel characteristics, and was more recently restated by Allen [All94]. Concrete evidence for this hypothesis may have been found in the experiments on the vowel sequence illusion [WHC96] described in Section 2.2.3. In these experiments, listeners frequently hear *two* voices, each with a different timbre, repeating different words or syllables. The two voices correspond to separate spectral regions, one is ca. 300–1200 Hz and the other is ca. 1500–3500 Hz [CW94]. This spectral fissioning may reflect an additional processing step taken by the auditory system in the face of an ambiguous input, as proposed by Chalikia and Warren, or it may reflect the premature termination of the speech recognition process, based on the integration of partial recognition results derived from independent processing of different frequency channels.

The reasons for the use of multiple representations in sensory systems are manifold. In some sensory systems, averaging of many estimates of a perceptual variable may be required to overcome the relatively low precision afforded by individual neurons [Hei89, Cal83]. The important features of the input may be present at many scales, such that processing at a single scale is not efficient, or feature detection may rely on agreement from processing performed at more than a single scale [Mar76, MH80]. The robust representation of one form of information about an input may require the suppression of other, relevant information, necessitating multiple representations. Finally, different representations of the input may be insensitive to different forms of distortion, and thus basing perception on combinations of multiple representations may lead to more reliable sensory processing.

2.4 Summary

The reliability of human speech recognition in the face of a wide range of speaker characteristics and acoustic environments is attributable, at least in part, to the auditory representation of the speech signal. There are a number of strategies used by the human auditory system that appear to contribute to the relatively invariant representation of features of the speech signal that convey linguistic information and that may also be used to improve the reliability of ASR systems. These strategies include the following:

Critical-band frequency resolution. The use of critical-band-like frequency resolution in automatic speech recognition systems is already widespread because it reduces the recognizer's sensitivity to speaker-dependent signal characteristics and enhances its sensitivity to speech-dependent signal characteristics. An ASR system's performance in the presence of noise may also be improved because a critical-band-like frequency analysis enhances the representation of the lower frequency portions of the signal which contain the bulk of the signal energy.

A focus on slow changes in the spectral structure of the speech signal. Change in the spectral structure of the speech signal occurring at rates between 1 and 16 Hz appears to be the primary carrier of linguistic information in speech. Processing that suppresses changes outside of this important band may act as a sort of matched filtering operation for speech and should improve the reliability of ASR systems.

Adaptation. The inclusion of frequency-local adaptation in an ASR system should reduce the system's sensitivity to unknown spectral shaping of its input. As shown in Chapter 3, there are already front ends for ASR systems which do this (e.g., cepstral mean normalization and RASTA-PLP).

Integration of phonetic information over syllabic durations. The integration of phonetic information over syllabic durations in the speech signal may improve ASR reliability in the presence of intermittent interference by averaging judgments over a relatively long period of time. Integration of information over syllabic time scales is also likely to be an effective strategy because coarticulation appears to distribute information about phonetic identity over entire syllables [MSE93].

The use of multiple signal representations in the recognition process. The use of multiple signal representations in the speech recognition process may provide another means for improving the reliability of ASR systems. An improvement in reliability is possible if the different representations have markedly different properties such that phonetic decisions based on individual representations have different error patterns, and if a means can be found to combine the decisions such that correct decisions tend to overrule incorrect decisions.

The application of perceptual signal-processing strategies to automatic speech recognition may improve the reliability of ASR systems in the presence of acoustic interference, such as reverberation, if they are implemented in a manner that is compatible with the limitations of the recognition algorithms currently in use. Experiments in automatic speech recognition testing different implementations of these perceptual strategies are required to ensure compatibility.

Chapter 3

Speech Recognition by Machines

Recognizing speech is essentially a process of finding the sequence of words that best corresponds to a given acoustic sequence, based on some similarity metric which should account not only for acoustic similarity but also for higher-level constraints such as phonotactics, syntax, and pragmatics. Like the vast majority of modern automatic speech recognizers, the ASR system described in this thesis is based on statistical pattern recognition techniques that use probability as a similarity metric. Basing recognition algorithms on probability theory is useful because it provides a well-defined framework for making decisions in the face of uncertainty.

Ideally, an ASR system based on statistical pattern recognition techniques will recognize a sequence of acoustic vectors $X = (x_1, x_2, \ldots, x_t)$ by finding the most probable sequence of models $\tilde{M} = (m_1, m_2, \ldots, m_n)$ given the acoustics, X, and a set of parameters, $\tilde{\Theta}$, for the set of models from which elements of $\tilde{M}$ are drawn. That is, to recognize an acoustic sequence, X, an ASR system should compute

$$\tilde{M} = \operatorname*{argmax}_{M \in \mathcal{L}} P(M|X, \tilde{\Theta})$$

where $\mathcal{L}$ is the set of all possible model sequences. The models that are matched to the acoustics by the recognition process may model sequences of words, individual words, or basic sound elements such as syllables or phones. Prior to recognition, the model parameters $\tilde{\Theta}$ should be learned by computing

$$\tilde{\Theta} = \operatorname*{argmax}_{\Theta} \prod_{k=1}^{K} P(M_k|X_k, \Theta)$$

where each (M_k, X_k) pair is drawn from a training data set in which all acoustic sequences are labeled with their corresponding model sequence. These recognition and training methods are called MAP (maximum *a posteriori*) methods because they maximize model probabilities given observed data and model parameters — *a posteriori* probabilities. Using the MAP criterion minimizes the probability of recognition errors.

The training of ASR systems is usually based on the maximum likelihood (ML) criterion instead of the MAP criterion. While there exist training algorithms based on the MAP criterion for ASR systems [BBdSM86, KBM96], the ML-based training algorithms require significantly less computation. The ML criterion is derived by applying Bayes' rule and making a set of simplifying assumptions. Thus,

$$P(M|X) = \frac{P(X|M)P(M)}{P(X)}$$

The simplifying assumption made in ML training is that the prior probability of the acoustic sequence, $P(X)$, is constant. While this assumption is true during recognition, it is not during training. The prior probability of a model sequence, $P(M)$, is estimated by a statistical model, called the language model, which is usually trained independently of the other parts of the system. The computation of $P(X|M)$, which is a computation of probabilities of sequences, is performed using hidden Markov models (HMMs).

HMMs model the speech signal (the sequence of acoustic vectors) as the output of a stochastic system that can be described by a set of states, $\{q^1, q^2, \ldots, q^p\}$, a set of time-independent transition probabilities between the states, $p(q_{t+1} = q^j | q_t = q^i)$, and a probability distribution for each state describing the distribution of acoustic vectors, $p(x|q)$. Figure 3.1 illustrates a simple two-state HMM. An HMM is described as "hidden" because the process that produces the sequence of acoustic vectors (the sequence of HMM states) is not directly observable and must be inferred from the acoustics.

As mentioned above, the HMMs used in an ASR system may correspond to words, sequences of words, or parts of words. The specific identity of the HMMs and HMM states is a design decision that varies from system to system. In many systems, the HMMs model basic sound elements in speech such as phones, diphones,[1] or syllables and the HMM states

[1] Diphones represent transitions between successive phones, and are usually defined as going from the midpoint of the steady-state part of one phone to the midpoint of the steady-state part of the following phone. Thus, if a system defines N different phones, there are N^2 possible diphones, although not all diphones will occur in practice due to phonotactic constraints.

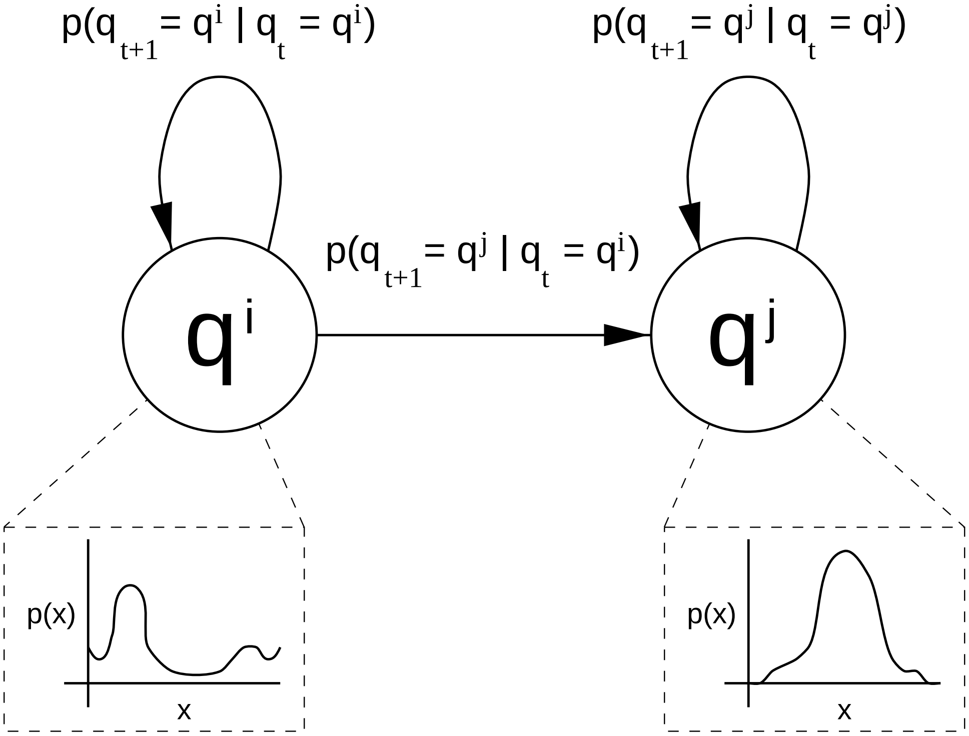

p(q_{t+1} = q^i | q_t = q^i)
p(q_{t+1} = q^j | q_t = q^j)
p(q_{t+1} = q^j | q_t = q^i)
q^i
q^j
p(x)
x
p(x)
x

Figure 3.1: A hidden Markov model is a stochastic, finite-state automaton that is typically defined by a set of states, $\{q^1, q^2, \ldots q^p\}$, transition probabilities between the states, $p(q_{t+1} = q^j | q_t = q^i)$ and a probability distribution, $p(x|q_i)$, that describes the acoustic vectors associated with that state.

correspond to short segments of these sound elements. In such systems, words are modeled by concatenating the HMMs for the more basic speech sounds. In systems where the HMMs represent specific sound units, the models may be context-independent or context-dependent. In a context-independent system, there is a single model for each sound unit, while in a context-dependent system there are multiple models for each sound unit, with each one modeling the sound given the identity of one or more of its neighboring units. For example, a context-independent, phone-based system would use the same model for the "a" sound in the words "cat," "rack," and "gab," while a context-dependent system might use three different "a" models. Other systems use HMMs to model whole words. In these systems the HMM states correspond to segments of the words that do not have any specific linguistic identity. Instead, the correspondence between states and acoustics is determined automatically during training.

Choosing the identity of the HMMs and HMM states involves trade-offs between model complexity, model accuracy, and training data requirements. If the HMMs model whole words or context-dependent sound units, then the complexity of the model for each state may be lower because it models a specific sound in a specific context. This potential for model simplicity in the states comes at the price of more states in the system, increasing the number of parameters in the system and, therefore, increasing the need for training data. In contrast, a context-independent system may require more complex models at the level of individual states to account for the variability introduced by different acoustic contexts, but will require fewer states overall, and may therefore require less training data.

3.1 Implementation of ASR Systems

Most automatic speech recognition systems break the recognition process down into a series of steps:

feature extraction in which the speech signal is processed to produce a set of features that describe the signal, usually in terms of spectral shape,

acoustic modeling in which the likelihoods that the acoustic features were produced by the different HMM states are computed,

language modeling in which the prior probability of sequences of words is estimated, and

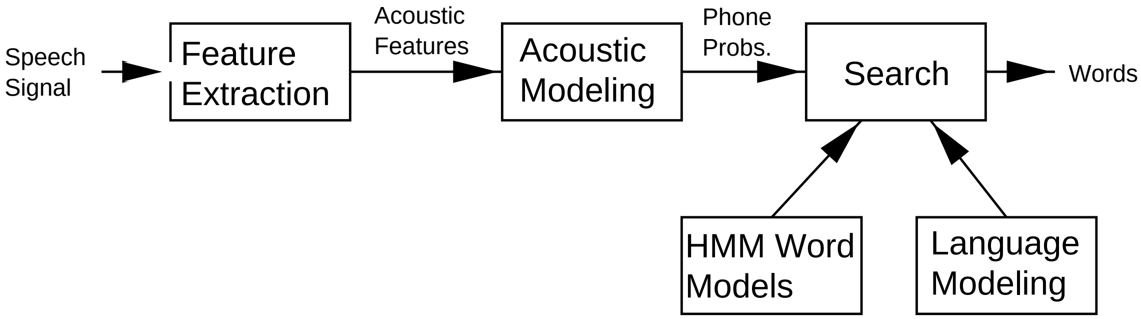


Figure 3.2: Structure of a typical HMM-based automatic speech recognition system. The processing is broken down into a series of stages: feature extraction, acoustic modeling, word modeling, language modeling, and search.

search in which the evidence from the acoustic model, language model, and lexicon of HMM word models is combined to determine the most probable utterance.

This structure is illustrated in Figure 3.2. The implementation of these different stages is discussed below, with a particular focus on the ASR system used in this thesis.

The primary recognition task used in the current work is a small-vocabulary task, namely recognition of continuous, spontaneously spoken numbers over the telephone. More details about this task are presented in Chapter 4. In Chapter 6 some preliminary results are described for a large-vocabulary recognition task (Broadcast News).

3.1.1 Feature Extraction

The purpose of the front-end feature extraction stage of an ASR system is to produce a description of the incoming signal that carries as much information as possible about the linguistic content of the signal (the phonetic identity of the incoming speech) and suppresses as much of the non-linguistic content of the signal as possible. This non-linguistic content includes information about speaker identity, such as gender, vocal tract length, accent and age, as well as information about the acoustic environment and the channel carrying the speech to the ASR system, such as background noise, filtering and reverberation.

Most ASR systems use feature extraction algorithms that produce a description of the spectral shape of the incoming speech, measured over segments (called “frames”) of around 16–32 ms and updated every 8–16 ms. These spectral shape features are often supplemented with differential features (also called “delta features”) that describe the change in spectral shape over time. Currently, the most popular features for ASR systems are mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) [Her90] coefficients. Both representations generate a description of the short-time spectral shape of the input signal that is based, in part, on auditory-like signal processing. While their details differ somewhat, the MFCC and PLP algorithms are sufficiently similar that they may be described together:

1. Both algorithms begin with the computation of the short-time power spectrum of the input signal. This is done by applying the FFT to windowed frames of the input. In the computation of MFCCs, the spectrum of the input signal is usually flattened prior to the computation of the power spectrum by filtering the signal with a pre-emphasis filter.

2. Next, a critical-band-like spectrum is derived by convolving the power spectrum with a bank of filters. For the computation of MFCCs, a typical design may use around twenty overlapping filters (for telephone-bandwidth speech) having triangular magnitude responses, constant bandwidths for frequencies below 1 kHz and bandwidths proportional to the filter center frequencies for frequencies above 1 kHz. For the computation of PLP coefficients, the filterbank contains a set of overlapping, filters that are equally spaced on the Bark frequency scale with bandwidths and center frequency spacings of about 1 Bark, high-frequency rolloffs of -25 dB/Bark and low-frequency rolloffs of -10 dB/Bark. The implementation of PLP used in this thesis used filters having a trapezoidal magnitude response, while other implementations use triangular filters.

3. The dynamic range of the critical-band-like spectrum is compressed. For the computation of MFCCs, the logarithm of the spectrum is computed (this is also crucial if homomorphic processing such as cepstral mean normalization is to be applied to the cepstral coefficients). For the computation of PLP features, the critical-band-like power spectra are warped to a scale that is similar to perceptual loudness by equalizing them according to a static equal-loudness weighting and taking the cube root.

4. Cepstral coefficients are then computed. For the computation of MFCCs, this is done directly, by computing the discrete cosine transform (DCT) of each log-compressed,

critical-band-like power spectrum. For the computation of PLP features, the perceptually warped spectrum is approximated by an autoregressive all-pole model, and the resulting LPC coefficients are transformed into cepstral coefficients. The autoregressive modeling serves to emphasize peaks in the spectrum of the input.

5. Typically, for both MFCCs and PLP features, only the eight to twelve lowest-order coefficients are used (for telephone-bandwidth speech). The zero-order coefficient, which is simply a measure of the total frame energy, is often discarded because, if it is not normalized in some way, it primarily conveys information about the overall energy level of the utterance.

In the experiments described in this work, the focus will be on different perceptually inspired feature extraction algorithms, with comparisons against PLP features and various forms of RASTA-PLP features (described in Section 3.2.5), an extension to PLP that improves robustness to unknown spectral shaping or joint spectral shaping and additive noise.

3.1.2 Acoustic Modeling

The purpose of the acoustic modeling stage in an ASR system is to estimate local acoustic likelihoods (HMM emission probabilities) $p(x|q_t = q^i)$ — the probability of acoustic features given an HMM state. This estimation is most often performed using a Gaussian mixture model trained using the maximum-likelihood criterion. The recognizer used in this work, however, is a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) system [BM94] that estimates the acoustic likelihoods using a multilayer perceptron. The MLP used for acoustic modeling in these experiments is a 2-layer, feedforward MLP, as illustrated in Figure 3.3. The input layer presents a context window of $2c + 1$ consecutive frames, centered on the current frame. In other words, the MLP input is the sequence of acoustic vectors $x_{t-c}, x_{t-c+1}, \ldots, x_t, x_{t+1}, \ldots, x_{t+c}$, where x_t is the vector of acoustic features for the current frame. A context window of nine frames is used in most of the experiments in this thesis. The MLP contains a variable number of hidden units (usually in the range of 300–600 in this thesis), and as many output units are needed to represent the context-independent phones that compose the recognizer vocabulary. The vocabularies used in the study required between thirty-two and fifty-four phone units. The hidden units

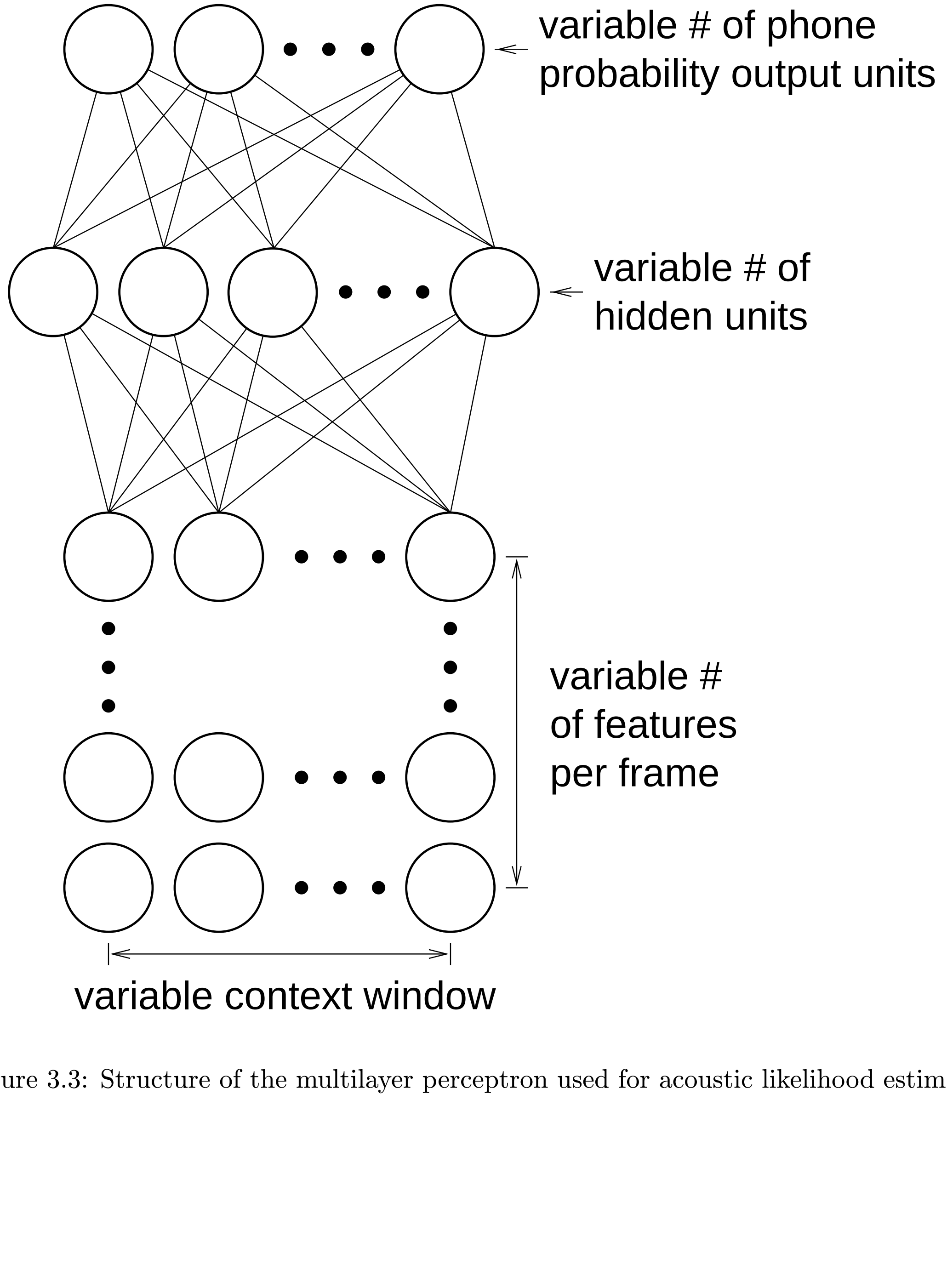


Figure 3.3: Structure of the multilayer perceptron used for acoustic likelihood estimation.

are standard sigmoid units which compute the output activation y as

$$y(x) = \frac{1}{1 + e^{-w^T x}}$$

where w is the unit's weight vector and the input, x, is augmented with a constant element having a value of 1 so that $w^T x$ is an affine transform of x. The output layer of the MLP is a softmax layer [Bri90], in which the k-th output unit computes its activation, g_k, as

$$g_k = \frac{e^{w_k^T x}}{\sum_{i=1}^{K} e^{w_i^T x}}$$

where w_k is the weight vector for the k-th output unit and the input vector, x, is augmented with a constant element having a value of 1.

The features are normalized to have zero mean and unit variance before they are input to the MLP. The mean and standard deviation for each feature is estimated from the MLP training set, and the features are normalized by subtracting the means and dividing by the standard deviations. This step speeds up the MLP training because it ensures that the majority of the inputs to the MLP units (the $w^T x$ values) will fall into the high-gain region of the units' nonlinearities.

The MLP is trained using on-line error backpropagation with a cross-entropy error criterion. The training targets are hard targets, so that in a given frame the target output for the unit corresponding to the frame label is 1, while the target outputs for all other units are 0. The use of these targets along with error-backpropagation training and a cross-entropy error criterion ensures that the MLP estimates posterior probabilities, $p(q_t = q^i | x_{t-c}, x_{t-c+1}, \ldots x_{t+c})$ [BW89, RL91]. The estimates of probabilities from the MLP will not be completely accurate because

1. the training procedure will not necessarily reach the global error minimum,

2. the training data set may not be completely representative of the potential range of inputs, and

3. the accuracy of the estimates will be limited by the size of the MLP.

Given these limitations, the softmax normalization is useful because it ensures that the MLP outputs will sum to one. Also the softmax function is the correct form for the *a posteriori* probability density for a wide range of class-conditional probability density functions [Jor95].

The posterior probabilities estimated by the MLP are converted to scaled likelihoods by dividing them by class prior probabilities, $p(q_t = q^i)$. The priors are estimated by counting the labels in the training data.

To prevent overfitting on the training data, roughly ten percent of the training utterances are set aside as a cross-validation set and classification performance on this set is used to control the training process. The learning rate for the error-backpropagation training is initially set at 0.008. Once the frame classification accuracy on the cross-validation set does not improve by at least 0.5% (absolute) in an epoch of training, the learning rate is halved for each subsequent training epoch. When the classification accuracy on the cross-validation set again does not improve by at least 0.5% the training is halted. This particular MLP training schedule was developed in earlier work at ICSI [BM94] and has proven to be a reasonable one for the tasks described in this thesis.

The training targets for the MLP may be based on hand transcriptions of the training data or on a labeling of the training data generated by another ASR system using a forced-alignment procedure. Forced alignment is described in more detail in Section 3.1.6.

MLP training is a computationally demanding task. The MLPs used in this study typically have on the order of 100,000 weights, and they are trained in seven to ten epochs over about two hours' worth of training data (675,000 training patterns). A special-purpose hardware accelerator, the SPERT-II system [WAK$^+$96, WAK$^+$95] was used to speed up the training (by a factor of 4–10 over available workstations). The SPERT-II system integrates a full-custom, fixed-point vector microprocessor (called "T0") [AKB$^+$96, ABI$^+$95], 8 MB of static RAM for the T0 processor, and a Xilinx FPGA on a double-wide Sbus card. SPERT-II functions as an attached coprocessor in a host Sun-compatible workstation. The MLP training and forward pass operations were performed using the QNTRAIN and QNFWD programs, written by David Johnson of ICSI.

3.1.3 The Lexicon

The lexicon of HMM word models used in this study contains multiple pronunciations for each word in the recognizer's vocabulary. Minimum phone duration constraints are enforced by repeating states in the HMMs. The HMM transition probabilities in the lexicon are not trained. Instead, they are fixed to $1/T$, where T is the number of transitions leaving a given state (including the self-loop). This is done because the scaled acoustic likelihoods from the MLP have a much greater dynamic range than do the transition probabilities, and they therefore dominate the calculation of the acoustic sequence likelihoods. The lexicon was generated as follows:

1. A set of pronunciations that covered about 90% of the pronunciations in the training set was derived from the phonetic transcriptions of the training utterances.[2]

2. Average context-dependent phone durations were calculated from the hand transcriptions of the training data.

3. An initial HMM lexicon was generated that incorporated all of the pronunciations determined in step 1 and the context-dependent phone durations found in step 2. The durations are included in the model by repeating states such that each phone is modeled by a sequence of n states, where $n = d/(2s)$, d is the average duration of the phone, and s is the frame step time. Thus, an [a] sound having an average duration of 80 ms would be modeled by a sequence of four [a] states in a system with a 10-ms frame step. This repetition of states matches the expected duration of the modeled phone to the average phone duration measured from the training set.

4. An ASR system was then trained using the hand transcriptions of the training set as targets and the lexicon generated in step 3. The resulting recognizer was then used to relabel the training data via forced alignment.

5. The final lexicon was generated by eliminating any of the pronunciations found in step 1 that were not used in the relabeling of the training set, computing new phone durations from the relabeling and compiling a new lexicon using the pruned set of pronunciations and the newly derived durations.

The iterative procedure described above may be repeated multiple times; however, a single iteration was sufficient (in most cases) to give good performance for the current study.

[2] Thanks to Dan Gildea of ICSI for deriving the pronunciations.

3.1.4 Language Modeling

The language model estimates the prior probability of a sequence of words, $M = m_1, m_2, \ldots, m_n$. The recognizer used in this work uses a bigram grammar which approximates $P(M)$ as

$$P(M) \approx P(m_1|s)P(e|m_n)\prod_{i=2}^{n} P(m_i|m_{i-1})$$

where s is the start symbol, $P(m_1|s)$ is the probability that m_1 is the first word of an utterance, e is the end symbol, and $P(e|m_n)$ is the probability that m_n is the last word of an utterance. The bigram probabilities, $P(m_i|m_{i-1})$, are estimated by counting word pairs in the training data. The probabilities of word pairs that never occur in the training set or do not occur frequently enough to permit reliable estimation of the conditional probability are approximated using a simple backoff method:

$$P(y|x) \approx b_x P(y)$$

where $P(y)$ is the prior probability of word y and b_x is the "backoff weight" for word x. The prior probabilities for the individual words are calculated by counting words in the training set. Methods for calculating backoff weights are described in [CG91].

3.1.5 Search

Recall that speech recognition in a statistical framework is accomplished by finding the model sequence, $\tilde{M}$, such that

$$\tilde{M} = \operatorname*{argmax}_{M \in \mathcal{L}} P(X|M)P(M)$$

where $\mathcal{L}$ is the set of all possible model sequences, X is the sequence of acoustic feature vectors to be recognized, $P(M)$ is the prior probability of a model sequence calculated by the language model, and $P(X|M)$ is the probability that model sequence, M, produced the acoustic sequence, X. The search stage of an ASR system is responsible for computing $\tilde{M}$, given the stream of acoustic likelihoods from the acoustic model, the lexicon of HMM word models and the language model. This search process is also referred to as "decoding."

The computation of $P(X|M)$ requires a summation over all possible state sequences corresponding to model sequence M. While it is possible to perform this summation in a computationally efficient manner, $P(X|M)$ is frequently approximated by the probability of the most likely state sequence through M — the summation in the computation of $P(X|M)$ is replaced by a maximization. In other words, the best state sequence is found instead of the best model sequence. This approximation is known as the Viterbi approximation.

The use of various scaling factors and penalties in the decoding process is a second common approximation in speech recognition systems. According to theory, the final score for a hypothesized word string should be the sum of the log likelihoods from the acoustic model and the language model.[3] In practice, most ASR systems compute a score for a hypothesis that is not an actual likelihood, but rather is a sum of the acoustic model's log likelihood, the language model's log likelihood multiplied by a language-model scaling factor, and additional penalty terms. A common penalty term is the word-transition penalty, which is a fixed value that is added to the score for a hypothesis once for each word-to-word transition in the hypothesized word string. These scaling factors and penalties are usually set empirically to minimize the word error rate on a given collection of development test utterances.

The recognizer used in this thesis is based on Y0, a decoder that uses a dynamic-programming search to compute the best state sequence (the Viterbi approximation) given the scaled acoustic likelihoods from the MLP, the lexicon of HMM models and the bigram probabilities from the language model. While Y0 is capable of speeding up the search by only considering a limited set of high-scoring hypotheses at each point in the search (a technique known as Viterbi beam search), most of the recognition tasks used in the current study were small enough that it was practical to employ a full Viterbi search.

3.1.6 Forced Alignment

The Y0 decoder, like many other speech decoding programs, may also be used for forced alignment. In the forced alignment procedure the decoder produces the most likely sequence of states for an utterance, given the stream of scaled acoustic likelihoods from the MLP, the lexicon of HMM models and the actual word sequence in the utterance. The correct word sequence functions as a trivial form of language model in forced alignment,

[3] ASR systems typically do not represent probabilities directly, combining them by multiplication, but instead work with log probabilities and combine them by addition. Use of log probabilities can speed up the search process (if additions require less time than multiplications) and can reduce (but not eliminate) underflow problems caused by the multiplication of many small numbers.

with the correct model sequence having a probability of 1 and all other sequences having probabilities of 0. Labels from forced alignment may be used as training targets for a new MLP and for generation of new lexicons (as described above). The process of iteratively training a recognizer, using it in forced-alignment mode to relabel the training data and then training a new recognizer based on the relabeling of the training data is often called embedded training. The embedded training process is useful for training recognizers on data for which no hand transcription of phonetic labels is available and for obtaining the best possible recognition performance by ensuring consistency between the acoustic models (the MLP, in this instance) and the HMM lexicon.

3.1.7 Combining Recognizers

As described in Chapter 2, many perceptual systems use multiple representations of their input as the basis of processing. A similar strategy is also useful for automatic pattern classification and recognition systems, including ASR systems. The reasons for using multiple representations in automatic systems are essentially the same as those proposed for perceptual systems. Multiple representations of the input may be needed because important features exist at different scales in the input or because processing to reliably represent one important feature obscures the representation of another. Better overall system performance may be obtained by combining decisions from sub-recognizers that use different input representations if the sub-recognizers tend to make different errors and the combination method allows correct decisions to override incorrect ones.

There are a number of levels at which different input representations or decisions based on those representations may be combined within an ASR system. In this thesis, only the two simplest approaches are considered—more than one speech representation may be supplied as input to the MLP, or scaled log acoustic likelihoods from more than one MLP may be combined by averaging them. The presentation of multiple representations to the same MLP may improve system performance if all the representations provide useful information about the linguistic content of the signal, if the information provided by each representation is somehow different from that provided by the others and if the MLP is able to learn to integrate the information from the representations. The combination of MLPs by averaging their estimates of acoustic log likelihoods may improve system performance if the distribution of likelihoods from each MLP tends to be relatively uniform (that is, if the entropy of the distribution is relatively high) when its likelihood estimates are unreliable and if the MLPs have different error patterns. A common form of this combination strategy is the use of both static and time-differentiated features (described in more detail in Section 3.2.3) in recognition systems. In hybrid HMM/MLP systems and in continuous-density HMM systems, the feature and differential feature vectors are frequently concatenated and presented as a single input to the acoustic model. In many discrete HMM systems, vector quantization is applied to the feature vector and differential feature vector separately and the resulting acoustic log likelihoods are combined by summation. It is also possible to combine recognizers at higher levels in the recognition process. For example, acoustic likelihoods from different recognizers may be averaged at syllable boundaries in the decoder or at the ends of entire utterances [Wu98, WKMG98a, WKMG98b].

3.1.8 Evaluating Recognizer Performance

The standard measure of an ASR system's performance is its word error rate measured on a specified test set — some collection of utterances on which it was not trained. Word error rate is computed on a test set by finding the best alignment between the word string hypothesized by the ASR system and the actual spoken word string for each utterance using a dynamic-programming search, then counting the number of word substitutions, deletions, and insertions between all the aligned word strings and normalizing by the total number of words in the test set. A substitution is counted when one word appears in the correct word string and a different word appears in the hypothesized word string. A deletion is counted when a word appears in the correct word string but has no corresponding word in the hypothesized word string. An insertion is counted when a word appears in the hypothesized word string but has no corresponding word in the correct word string. Note that it is possible to have a word error rate of more than 100% because insertions are included in the measure.

When comparing the word error rates for two different recognizers, measured on the same test set, it is necessary to account for some randomness in the performance of the two systems. In hybrid HMM/MLP systems this randomness arises from the training of the MLPs. The MLP weights are initialized to small, random values prior to backpropagation training, and the on-line MLP training algorithm selects examples in a random order from the training set. The word error rate for a recognizer should therefore be considered as an individual sample from a distribution of possible word error rates, and not as a fixed value. To account for the stochastic nature of the word error rate, recognizer performance measurements are compared using a one-tailed statistical significance test with a significance level of $p < 0.05$.

3.2 Robustness

One of the primary advantages of the statistical approach to automatic speech recognition is that it permits the development of usable recognition systems without requiring complete, detailed understanding of how linguistic information is encoded in the speech signal. Instead, relatively general and powerful learning methods are applied to large corpora of training data to automatically find patterns that enable recognition of speech. This data-driven approach has a weakness, however, in that it is difficult to ensure that the patterns a trained system learns will generalize to input not seen during training. An ASR system will usually not work as reliably on input that is not well-represented by its training data as it does on input that is similar to the training data. The problem of minimizing this degradation in performance is the problem of robustness.

There are a number of approaches that may be used to enhance the robustness of an ASR system. One of the simplest and most effective is to increase the size and diversity of the training set. This is currently an indispensable step for deploying an ASR system in a real-world application [Tho97]. Unfortunately, gathering additional training data is often time-consuming and expensive, and it may be difficult to completely characterize potential input variability in the training set. A second approach is to modify the recognition system so that it can model input variability and, by modeling it, compensate for it. Parallel model combination [GY92] and vocal-tract-length normalization [CKA94, LC95] are examples of this approach to enhancing ASR system robustness. The limitation of this approach is that the variability being compensated for must be modeled by the recognizer. This can increase the amount of processing time required by the recognizer. Also, the model of undesired variability may require its own training set. The final approach, which is the one explored in this thesis, is to make the recognition system insensitive to undesired variability by focusing the recognizer on essential features of the input. This may be done at the level of the classifier by, for example, using missing-data techniques [CGC94, MCG98] or using a distance metric that is insensitive to interference, but most commonly changes are made to the front-end signal processing to try to produce a robust representation of the linguistic information in the speech input.

These three approaches to robustness—using larger and more diverse training data, modeling undesired variability and using robust features and classification metrics—are complementary. The most successful systems will often use all three methods together. For example, the development of potentially usable speaker-independent ASR systems for American English was driven by the collection of large, multi-speaker corpora for ASR training and testing, by the development of systems that perform speaker clustering or vocal tract normalization, and by the use of front-end processing that incorporates auditory-like spectral analysis to reduce system sensitivity to speaker characteristics.

3.2.1 Temporal-processing Approaches to Robust Feature Extraction

Because it is a fundamental, long-standing problem in ASR research, the literature on robustness is vast, and a full review is beyond the scope of this dissertation. For a comprehensive review of the literature on robustness to noise in ASR systems, the reader is referred to [Gon95]. The various temporal approaches to robust feature extraction that have preceded and inspired the approaches explored in the current work are reviewed here, however, to place it in perspective.

The key idea behind temporal approaches to ASR robustness is that the spectrum of the speech signal changes at rates that are distinctive from the rates at which potential forms of acoustic interference are likely to change. When this is true, filtering of the time sequences of spectral parameters (the spectral trajectories) of the input signal in a domain in which the speech and interference are (approximately) additive can suppress the effects of the interference. This approach is essentially an extension of homomorphic filtering [OSTGS68].

Thus, it is possible to suppress additive noise whose spectrum changes more slowly or more rapidly than do the portions of the speech signal that carry linguistic information by filtering power spectral trajectories [HMR91] because the speech and noise signals are additive in the power spectral domain. This idea is similar to that of spectral subtraction [Bol79], but does not require speech detection.

Similarly, unknown spectral shaping of the speech signal may be suppressed by filtering log power spectral trajectories [HMBK91, HM94]. If a speech signal, $s(t)$ with short-time Fourier transform $S(n, \omega)$ is filtered by an unknown filter with impulse response $h(t)$ and transfer function $H(\omega)$, then the resulting signal will be $x(t) = h(t) * s(t)$, the convolution of $h(t)$ and $s(t)$. The short-time Fourier transform of $x(t)$ will be $X(n, \omega) \approx H(w)S(n, \omega)$, provided that $h(t)$ is short compared to the length of the windowing function used in the short-time Fourier transform [Ave97b]. In the log power spectral domain $S(n, \omega)$ and $H(\omega)$ will be approximately additive, and filtering of the log spectral trajectories will be effective for suppressing the effects of the unknown filtering, provided that $H(\omega)$ is stationary or changes at rates outside the linguistically important 1–16 Hz range, that $h(t)$ is short compared to the window used in the short-time spectral analysis, and that $H(\omega)$ does not contain any spectral zeroes. If the filtering of the log spectral trajectories includes a highpass component that suppresses changes at rates below about 1 Hz, it will also suppress the average spectrum of the speech signal, which can improve the speaker independence of the ASR system [NPLJ97].

More generally, a highpass component in the filtering of spectral trajectories in any given domain will tend to equalize the modulation spectrum of the features presented to the ASR system [NJ94]. This equalization can produce features whose temporal statistics better match HMMs than do those of the unfiltered features [NPLJ97]. A lowpass component in the filtering of spectral trajectories that suppresses changes at rates above 16 Hz or so can also improve ASR system performance because these changes do not carry significant linguistic information and because changes at these rates may not be accurately characterized by the front-end signal processing [NJ94, NPLJ97].

3.2.2 Cepstral Mean Normalization

One of the oldest and most widespread temporal-processing methods for compensating for spectral shaping of speech by an unknown channel characteristic in ASR systems is cepstral mean normalization (CMN) [Ata74]. In CMN, the average cepstrum for an utterance, $\hat{c}$, is calculated as

$$\hat{c} = \frac{1}{T} \sum_{n=1}^{T} c_n$$

where c_n is the cepstrum of the n-th frame and T is the number of frames in the utterance. The average cepstral vector, $\hat{c}$ is then subtracted from each c_n, and recognition is performed on these normalized cepstra. Because the cepstrum is a linear transform of the log power spectrum of the input signal, this operation normalizes any spectral shaping imposed on the input speech, as well as the average spectrum of the speech. CMN is simple to implement, computationally inexpensive, and reasonably effective at compensating for unknown spectral shaping. Its primary disadvantages are that its use increases the response latency of an ASR system because recognition cannot begin until an entire utterance is received and that CMN may introduce some undesirable variability in the input because the modulation frequency response of the filtering performed by CMN is dependent on utterance length. Both of these problems may be alleviated by normalization with an average cepstrum computed over a fixed-duration, sliding window on the input signal.

3.2.3 Delta Features

Delta features [Fur81, Fur86b] are a second widespread temporal-processing method for improving the robustness of ASR systems to noise and spectral shaping. The first, and occasionally second and higher-order, time derivatives of the feature vectors may be calculated using either

1. a regression analysis, e.g.

$$\Delta c_n = \frac{\sum_{i=-k}^{k} c_{n+i} i}{\sum_{i=-k}^{k} i^2}$$

   where Δc_n is the first derivative of the n-th feature vector c_n and the regression is performed over a $2k+1$-frame window centered on c_n, or

2. using finite differences, e.g.

$$\Delta c_n = c_{n+d} - c_{n-e}$$

   where d and e specify the location of the points, relative to c_n, used to estimate the first derivative.

The time window used to estimate the differential features (using either the regression or finite difference methods) may range from 30 ms to 100 ms. Differential features are usually supplied to a recognizer as additional features, although other approaches in which the features and differential features are combined in a weighted sum or the differential features are used to weight distances in a dynamic time warping recognizer have also been explored [EB82].

When the feature vectors from which the deltas are computed are cepstral vectors, the use of delta features will tend to enhance the robustness of a recognizer to spectral shaping of the input, much in the way that CMN does. Delta features, however, appear to be more generally useful. They usually improve recognition accuracy on clean speech, most likely because they enhance the representation of changes in the speech signal that appear to be particularly important for carrying phonetic information [Fur86a]. Delta features have also proven to be useful for improving recognizer accuracy on noisy speech and speech influenced by the Lombard effect [HA90].[4]

3.2.4 Basis Functions for Spectral Trajectories

The representation of the temporal structure of the speech signal in an ASR system may also be enhanced by describing a sequence of feature vectors in terms of a set of orthogonal basis functions. The most common version of this approach is the use of two-dimensional cepstra [ASNS89], computed by performing a two-dimensional discrete cosine transform (DCT) of a matrix representing a sequence of spectra, usually with auditory-like frequency resolution. For the DCT, the time sequence of spectra is segmented into blocks 50–70 ms in duration, usually with an overlap of 50% between adjacent blocks. The transformed features are spectro-temporally smoothed by truncation; only the matrix elements with low-order transformed frequency (quefrency) and low-order transformed time indices are output as features. Other basis functions for the time dimension have also proven to be useful for improving ASR accuracy [Mil96].

Like delta features, two-dimensional cepstra and other representations that expand the time sequence of spectral vectors in terms of a set of basis functions generally improve the performance of ASR systems. By enhancing the representation of the dynamics of the

[4] When a speaker is talking in the presence of background noise, the level of vocal effort is generally higher than when speaking in a quiet environment. This increment in vocal effort in response to noise is known as the Lombard effect, and is named for the French otorhinolaryngologist Etienne Lombard who first noted the effect in his patients in 1909 and reported on it in 1911 [Lom11] (cited in Lane and Tranel's comprehensive review on the Lombard effect [LT71]). Changes in vocal effort can significantly alter the characteristics of the speech signal [Sch85, HHP88], causing problems for automatic speech recognition systems.

speech signal in a relatively compact manner, their use increases recognizer accuracy on clean speech. The robustness of a recognizer to unknown spectral shaping may be improved by omitting the matrix elements with transformed time indices of zero (which are most influenced by spectral shaping) from the feature set.

3.2.5 Modulation Filtering

All three of the previously described methods for enhancing robustness—cepstral mean normalization, delta features, and expansion of spectral trajectories in terms of basis functions—perform a linear filtering operation on features derived from the speech spectrum (usually cepstral features), with the form of the filter rather strictly defined by the processing method. It can be advantageous, however, to treat the design of the filters more generally, because the filter parameters and the domain in which the filtering is performed may be optimized for a particular task and class of acoustic distortions.

Speech Enhancement Via Modulation Filtering

Modulation filtering was originally tested as a speech enhancement method to improve the intelligibility of speech corrupted by additive noise or reverberation [LS82]. Filtering was applied to critical-band power spectral or log-power spectral trajectories. The filter was designed to be the inverse of the ideal, theoretically derived modulation transfer function of a specific noise or reverberant condition [HS73], but modified in order that the filtering did not enhance very rapid fluctuations. Resynthesis of the signal from the processed power spectra was accomplished via an overlap-and-add procedure. This method was tested both as a pre-processing step, applied before distortion was imposed, and as a post-processing step. An improvement in intelligibility was observed only for the case of filtering in the log-power spectral domain prior to the addition of white noise. Somewhat greater improvements in intelligibility were then achieved by performing the filtering in a nonlinear domain which was approximately logarithmic for low amplitudes and approximately linear for higher amplitudes. The use of this hybrid nonlinearity reduced the occurrence of strong, annoying peaks in the resynthesized output.

A more successful method for enhancing the intelligibility of reverberant speech was a post-processing method that performed filtering in the power spectral domain, based on an auditory-like, time-frequency representation and resynthesized the speech from the processed spectra with a set of frequency- and amplitude-modulated sinusoids [Sch89]. The filtering was performed using a filter designed to invert the smoothed, averaged modulation transfer function of a room, to try to achieve position-independent compensation for reverberation within a specific room.

Highpass Filtering of Spectral Trajectories for Robust ASR

Modulation filtering was first applied to ASR to improve the robustness of a recognizer to room reverberation [Hir88, Hir92]. The filtering was performed using a highpass FIR filter applied to power spectral trajectories. The processed power spectra were then converted to cepstra and used as features in a speaker-dependent, isolated-word recognizer with a vocabulary of forty-three short, monosyllabic German words selected for maximal variability in their phonetic sequences. For tests with artificial reverberation, the modulation filtering improved the recognizer error rate from about 50% to about 20% for a reverberation time of 2.5 s and from about 35% to about 2% for a reverberation time of 1.0 s, using a filter optimized for a reverberation time of 1.2 s. Later work [HMR91] demonstrated that high-pass filtering in the critical-band power spectral domain could improve the robustness of a speaker-independent, isolated-word recognizer to additive white noise and car noise, and that high-pass filtering in the log critical-band power spectral domain could improve the performance of a speaker-independent, continuous German digit recognizer on a corpus collected under a diverse range of acoustic environments (e.g., an anechoic chamber and several offices) with different levels of background noise, using different microphones. It was also observed that the high-pass filtering did not improve the performance of a speaker-dependent recognizer with matched acoustic conditions for training and test data.

RASTA-PLP

A set of robust front ends based on modulation filtering that are of particular relevance to this work are the RASTA-PLP (**r**el**a**tive **s**pec**t**r**a**l **p**erceptual **l**inear **p**rediction) front ends [HM94, MH92, HMBK91]. They are reviewed here in detail because they are in many ways the direct precursors of the front ends explored in this study, are reasonably

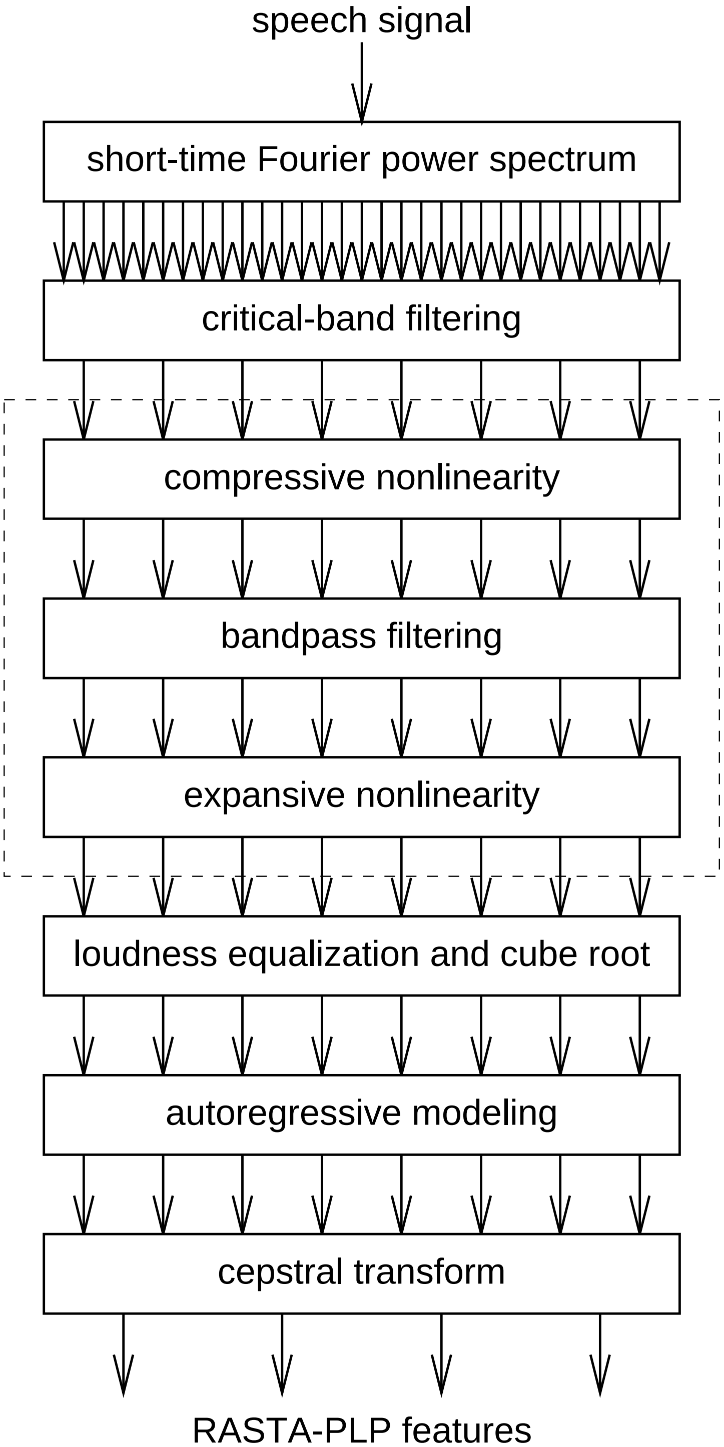


Figure 3.4: RASTA-PLP signal flow. The steps enclosed in the dashed box (compressive nonlinearity, bandpass filtering and expansive nonlinearity) are added to the PLP algorithm to compute RASTA-PLP features.

representative of most modulation filtering methods, and are the front ends against which the experimental front ends in this thesis are compared. RASTA-PLP is an extension to PLP that incorporates modulation filtering to compensate for unknown spectral shaping (log-RASTA-PLP) or to compensate jointly for unknown spectral shaping and additive noise (J-RASTA-PLP). The compensation is accomplished by filtering the critical-band power spectral trajectories in a domain appropriate for separating the speech and distortion. The general RASTA-PLP algorithm, which is summarized in Figure 3.4, proceeds as follows:

1. Critical-band-like power spectra are computed as for PLP, then the output of each critical-band filter is processed through a compressive, memoryless nonlinearity. In log-RASTA processing this nonlinearity is $y = \ln x$. In J-RASTA it is $y = \ln(1 + Jx)$, which is approximately linear for small Jx and approximately logarithmic for large Jx. During recognition, J is varied in inverse proportion to an estimate of the noise power in the incoming speech, so that channels with low power relative to the estimated noise are processed to suppress the noise, and channels with high power relative to the estimated noise are processed to suppress spectral coloration.

2. The nonlinearly transformed, critical-band power spectral coefficients are filtered through an IIR (infinite impulse response) bandpass filter with a passband between 1 and 12 Hz. This filtering emphasizes those parts of the signal that are changing at rates characteristic of speech, while suppressing elements changing at slower or faster rates. The RASTA filter, designed for a sampling rate of 100 Hz, is

$$H(z) = 0.1 * \frac{2z^2 + z - z - 2z^{-2}}{1 - 0.94z^{-1}}$$

3. The filtered power spectral coefficients are processed through an expansive, memoryless nonlinearity. In log-RASTA processing this nonlinearity is $x = e^y$, while in J-RASTA processing it is $x = e^y/J$. Finally, the PLP processing continues with the conversion of the power spectra to a loudness-like scale, autoregressive modeling and generation of cepstral coefficients.

Data-driven Filter Design

The general design of the filter used in the RASTA-PLP front end was chosen *a priori*, based on the idea that differentiation followed by leaky integration would be an effective strategy for reducing sensitivity to relatively stationary effects such as spectral shaping of the speech signal by an unknown channel. The specific parameters of the RASTA filter were set to optimize the performance of an ASR system that was trained on speech with one form of spectral shaping and tested on speech with a different spectral shaping.

It is also possible, however, to design the filters automatically from training data. Data-driven methods for designing filters have been successfully applied to both speech enhancement [HWA95, AH96, Ave97b] and to robust ASR [AvVH96, HAvVT97, Ave97b]. The signal-processing systems used for both the speech enhancement and speech recognition tasks are very similar:

1. A short-time Fourier transform is performed on the input signal and the resulting complex spectrum is split into magnitude and phase components. For the speech recognition task the phase is discarded. The magnitude coefficients may be processed individually or they may be integrated into a critical-band-like spectrum, as in the RASTA-PLP front end.

2. The magnitude coefficients are processed through a memoryless nonlinearity of the form $y = x^{\alpha}$, where α is a design parameter for the system.

3. The trajectories of the nonlinearly processed spectral magnitudes are filtered with an FIR filter that was designed automatically from training data. Depending on the application, there may be a different filter design for each spectral channel, or a single filter design may be used across the spectrum.

4. The filtered spectral trajectories are processed through the inverse nonlinearity $x = y^{1/\alpha}$.

5. In a speech enhancement application, the speech signal is resynthesized from the processed magnitude and unprocessed phase data using an overlap-and-add or filterbank summation technique. In a speech recognition application, the processed magnitudes may be used directly as speech features or they may be the basis of additional signal processing such as the autoregressive modeling in RASTA-PLP.

For both speech enhancement and speech recognition tasks the filters may be designed to minimize the differences between filtered, nonlinearly processed spectral magnitude trajectories computed from clean and distorted versions of the same utterances. If a least-squares criterion is used, the trajectory filters may be determined by solving the Wiener-Hopf equation for each spectral channel [AH96] or by solving a nonlinear, constrained optimization problem [AvVH96].

If the goal is to design a front end for robust ASR, the filters may instead be designed using linear discriminant analysis (LDA) [AvVH96]. In this case, the training data must be phonetically labeled. To derive filters for a spectral channel, fixed-length segments of nonlinearly processed spectral magnitude trajectories from that channel are assigned to different classes (based on the phonetic labeling of the data), and then an LDA procedure is run to find a set of filters that maximize the discriminability of the different segment classes. This procedure may be performed for training data for a single acoustic condition or the training data may include multiple conditions.

Filters designed using all three methods show strong similarities to the RASTA filter. They tend to be bandpass in form, with passbands covering the 1–16 Hz range. Unlike the RASTA filter, the automatically derived filters do not have zero response at 0 Hz. Instead, they usually have only 5–10 dB attenuation at 0 Hz.

A Long-time Technique for Reverberation-robust ASR

Although reverberation may be modeled as a form of convolutional distortion, temporal-processing methods, such as cepstral mean normalization and RASTA-PLP, generally have not been very effective for reducing the impact of reverberation on ASR performance (as demonstrated for RASTA-PLP in Chapter 4). One reason that CMN and RASTA-PLP are relatively ineffective for reverberation is that the impulse response associated with room reverberation is typically 0.5–2 s long, while the window used for spectral analysis in an ASR front end is much shorter, usually 16–32 ms in duration. Convolutional distortion is only (approximately) additive in the log spectral domain if the spectral analysis window is two to four times longer than the distorting impulse response.

A normalization technique that suppresses reverberation by operating on a long-time spectral representation of the speech signal has recently been proposed [Ave97b, HAvVT97]. This technique works as follows:

1. A long-time spectral representation of the speech signal, based on a 2-s analysis window, is generated using a DFT-based, critically sampled filterbank. The resulting

time-frequency representation is split into magnitude and phase components and the logarithm of the spectral magnitudes is computed.

2. Channel normalization is performed by computing the mean log magnitude in each channel over a 10-s, sliding window and subtracting this mean level.
3. The complex, long-time time-frequency representation is reconstituted from the processed log magnitude and unprocessed phase data. Then, a partial resynthesis of the original signal is performed, producing a short-time spectral representation of the input signal which is suitable for use as input features to an ASR system.

This dereverberation technique was tested on the same telephone-quality, continuous numbers recognition task used in this thesis. To test it, two hybrid HMM/MLP recognition systems were trained, one using RASTA-PLP features with eighth-order autoregressive analysis and first-order delta features and a second using features from the long-time dereverberation processing. The two systems were tested on a clean test set and on a reverberant test set that was generated by convolving the clean test set with a room impulse response having a 0.5 s reverberation time (T_{60}) and a 1-dB direct-to-reverberant energy ratio. The system using the long-time dereverberation processing had a word error rate of 22.8% on the reverberant test set, which was a significant improvement over the 34.8% word error rate obtained by the RASTA-PLP system on the reverberant test set. This performance improvement under reverberant conditions came at the cost of less accurate recognition for clean conditions, however. On the clean test the RASTA-PLP system had a word error rate of 8.6%, while the system using the long-time dereverberation processing had a word error rate of 13.5%.

3.3 Summary

Like most modern ASR systems, the recognition system used in this thesis is based on statistical pattern recognition techniques. The task of recognizing speech is broken down into a series of steps:

1. The front-end signal processing, which attempts to derive features from the speech signal which carry as much information as possible about the linguistic content of the signal and as little information as possible about the non-linguistic content.

2. The acoustic model, which estimates the probability of the acoustic features' belonging to different phonetic classes. The recognizer used in this work performs the classification using a multilayer perceptron, while most other state-of-the-art recognizers use Gaussian mixture models.

3. The HMM lexicon, which models the pronunciations and temporal characteristics of the words in the recognition system's vocabulary.

4. The language model, which estimates the prior probabilities of word sequences. The recognizer in the current work uses a backoff-bigram grammar for language modeling.

5. The decoder, which determines the most likely word sequence for a given acoustic input from the stream of acoustic likelihoods from the acoustic model, the constraints embodied in the HMM lexicon and the estimates of word sequence prior probabilities from the language model. The decoder used in this work performs a Viterbi search using dynamic programming.

The application of statistical methods to ASR has enabled the development of practical, usable recognition systems in the absence of a comprehensive model accounting for the encoding of linguistic information in the speech signal. Because ASR systems must be trained on some finite set of data, though, they tend to perform poorly on input which was not well-represented in the training set. The problem of reducing the degradation in performance caused by such input is the problem of robustness in ASR. An ASR system's robustness may be improved by making changes in any or all of the processing stages outlined above, but much of the work in robustness has focused on the front-end signal processing. Temporal-processing approaches to ASR robustness are of particular interest in this work because they have proven to be useful strategies for dealing with many forms of acoustic distortion and because there are interesting parallels between them and human auditory processing of speech.

Chapter 4

Initial Experiments with a Modulation-based Representation

As illustrated in Chapter 2, the robustness of human speech recognition to various forms of acoustic interference arises in part from specific properties of the auditory cortical representation of speech. Critical-band filtering, sensitivity to slow modulations, adaptation and the use of multiple representations of the input appear to be particularly important. Work on robust ASR has confirmed the utility of some of these strategies for improving the performance of automatic recognition systems in the presence of different forms of acoustic variability. By devoting more frequency resolution to the range of frequencies into which the first and second formants usually fall and having a coarser resolution in the higher frequencies, speech representations with auditory-like spectral resolution, such as mel-frequency cepstral coefficients and PLP, improve the performance of speaker-independent ASR systems. The robust temporal-processing methods that operate in the log spectral or cepstral domains perform an adaptive, automatic-gain-control function, and all of the temporal-processing strategies reviewed alter a recognizer's sensitivity to different modulation rates, although they may not exhibit a sensitivity similar to that observed in humans or other mammals. The combination of multiple representations of the input (and of multiple decisions based on different input representations) has been shown to improve ASR accuracy in a number of acoustic conditions [Wu98].

This chapter describes experiments with an initial, simple signal-processing system that incorporates some of the promising perceptually inspired strategies reviewed above. The representation was first developed to generate visual displays of speech that are stable across a range of acoustic distortions. This representation has been named the **m**odulation-filtered **s**pectro**g**ram (MSG).[1] The signal-processing system developed on the basis of the visual displays was tested as a front-end processor for an ASR system. It was found that the new MSG representation was significantly better than the PLP, log-RASTA-PLP, and J-RASTA-PLP front ends for a highly reverberant recognition test, but that it was significantly worse than any of the RASTA-PLP front ends for a clean recognition test. The intelligibility of the reverberant utterances for human listeners was then measured because relatively poor performance (word error rates on the order of 70%) was observed for all the ASR systems on the reverberant test. The reverberant test was found to be challenging for human listeners, but by no means impossible. The average word error rate for human listeners on the reverberant test was 6.1%. Next, variations on the original MSG processing were tested with the goal of improving ASR performance and determining how the different processing steps contribute to robustness to reverberation. Finally, an improved version of the MSG front end was tested on its own and in combination with the different RASTA-PLP front ends for a different, more moderate and more realistic reverberant test set and for a range of noisy test conditions.

4.1 Visualization Experiments

To try to gain insight into the effects of different perceptually inspired signal-processing strategies on the representation of speech, an initial speech visualization study was performed [GK97]. A simple and flexible signal-processing system that produces spectrographic-format displays of speech and includes auditory-like frequency resolution, adaptation, sensitivity to slow modulations, and emphasis of spectro-temporal peaks was developed, and the effects of different signal-processing parameters on the visual representation of clean and corrupted speech were examined. The generation of a modulation-filtered spectrogram, which is illustrated in Figure 4.1, proceeds as follows:

[1] This name represents a slight departure from previously published work [KMG98, Wu98, WKMG98a, WKMG98b, KMG97, KM97, GK97] that used the name "modulation spectrogram" instead of "modulation-filtered spectrogram." The name of the representation was changed because the name "modulation spectrogram" often caused listeners to mistakenly assume that the representation would explicitly portray the modulation spectrum of the signal at any given time.

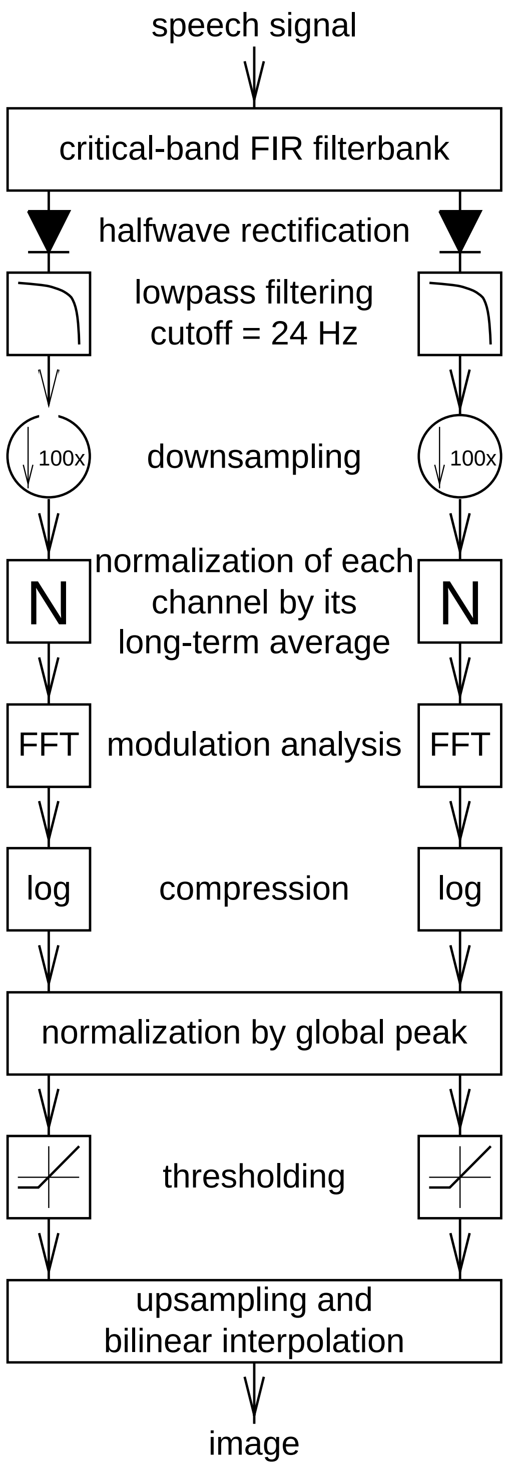


Figure 4.1: Signal-processing system used in initial visualization studies.

1. A speech signal (sampled at 8 kHz for all the experiments reported here) is analyzed into eighteen approximately critical-band-wide channels using an FIR filterbank. The filters were designed using a Kaiser window and were specified to have a trapezoidal magnitude response, 40 dB of stopband rejection, and minimal overlap between adjacent filters. The filterbank was designed to approximately cover the telephone bandwidth (300-3300 Hz) and, with the exception of the lowest two filters, which had bandwidths of 50 Hz, the filter passband edges were set to correspond to critical bands, defined as corresponding to ca. 0.9 mm segments of the basilar membrane based on the following modified form of Greenwood's function [Gre95]:

$$f = 160 \left(10^{0.06x} - 0.4\right)$$

2. Following the spectral analysis, the amplitude envelope is computed in each channel by half-wave rectifying and lowpass filtering the filterbank output. The lowpass filter is a linear-phase FIR filter with a cutoff frequency of 24 Hz and 40 dB of stopband rejection. To reduce the computational costs of the signal processing, and to match typical data rates into ASR systems, the envelope signals are downsampled by a factor of 100 to a sampling frequency of 80 Hz.

3. The amplitude envelopes are then normalized by computing the long-term average level (over an entire utterance) in each channel and then dividing each envelope signal by its average value. This normalization, which is analogous to cepstral mean normalization, is intended to capture some aspects of auditory adaptation, namely insensitivity to overall signal energy, and insensitivity to spectral shaping of the speech signal. It does not, however, provide the enhancement of signal onsets that is seen in the auditory system or in representations such as RASTA-PLP that use on-line automatic gain control.

4. The modulation frequency content of the normalized amplitude envelope signals is then analyzed by performing an FFT over a sliding Hamming window on each signal. The log magnitude of each result is computed, and one of the bins is selected for display. The duration of the modulation analysis window and the displayed FFT bin were experimental parameters. This processing step is intended to model the selectivity for slow modulation frequencies observed in the auditory system.

Frequency Range	T_{60}
0–250 Hz	3.1 s
250–500 Hz	2.6 s
500–1000 Hz	2.2 s
1000–2000 Hz	1.6 s
2000–4000 Hz	1.4 s

Table 4.1: Estimated reverberation times (T_{60}) in different frequency bands for a highly reverberant hallway.

5. The log magnitudes from the selected FFT bin are normalized by locating the maximum over all channels and all times in a given utterance, and this maximum level is subtracted from all of the log magnitudes. A thresholding operation is then applied in which all normalized log magnitudes equal to or below a given threshold are set equal to the threshold level, with the threshold level being an experimental parameter. The thresholding operation serves to emphasize spectro-temporal peaks in the signal by eliminating the representation of portions of the signal below the threshold.

6. The processed log magnitudes are then plotted in a spectrographic format, on the time-frequency plane, with bilinear smoothing used to produce the final image.

By plotting clean and corrupted samples of speech in this format and comparing their representations, the stability of the representation produced with different settings of the parameters was tested. The forms of acoustic interference that were examined were additive pink noise[2] at different signal-to-noise ratios, and severe reverberation. Using a 250-ms (20-sample) Hamming window for the modulation analysis, selecting the 4-Hz bin from the modulation analysis and thresholding all points 30 dB or more below the peak level in an utterance produced a relatively robust representation.

The severe reverberation was imposed by convolving clean speech samples with an impulse response designed to match the gross acoustic properties of a hallway approximately 6.1 m long, 2.4 m high, and 1.7 m wide with a floor, ceiling, and walls of concrete. The reverberation time of the hallway in different frequency bands was estimated from a recording of speech produced in the hallway and simultaneously recorded onto digital tape

[2] The power spectral density of pink noise is constant on a logarithmic frequency axis, above some specified cutoff frequency. The pink noise used in this study was taken from the NOISEX CD-ROM [VS93].

with a head-mounted, close-talking microphone and with an omnidirectional microphone located on the floor about 2.5 m from the talker. The estimated reverberation times are listed in Table 4.1. To synthesize the reverberant tail of the impulse response, a Gaussian white noise sample was filtered into subbands matching those used in the estimation of the hallway's reverberation times. Each subband was modulated with a decaying exponential envelope matched to the reverberation time for that subband. The modulated noise bands were then added together. The early reflections in the impulse response were estimated using a time-domain image expansion simulation [AB79], while the direct-to-reverberant energy ratio was manually adjusted to match the original recording. The final impulse response has an overall reverberation time of about 2.2 s and a direct-to-reverberant energy ratio of -16 dB. While a direct-to-reverberant energy ratio of -16 dB seems very severe (speech-shaped additive noise with an SNR of -16 dB would greatly reduce speech intelligibility), it is important to note that the direct-to-reverberant energy ratio is not equivalent to a signal-to-noise ratio because reflections arriving within about 80 ms of the direct sound contribute to speech intelligibility rather than detracting from it. The "early-to-late energy ratio" for this impulse response, which counts all energy arriving within 80 ms of the direct sound as contributing to intelligibility, is -2 dB.

To illustrate the properties of the modulation-filtered spectrogram representation, standard wideband spectrograms and modulation-filtered spectrograms are plotted for the utterance "two oh five," collected from a female speaker over the telephone, for the following five acoustic conditions:

clean conditions are shown in Figure 4.2.

moderately noisy conditions are shown in Figure 4.3. The moderately noisy utterance was generated by adding babble noise from the NOISEX CD-ROM to the clean utterance at an SNR of 20 dB, measured over the entire utterance.

very noisy conditions are shown in Figure 4.4. The very noisy utterance was generated by adding babble noise from the NOISEX CD-ROM to the clean utterance at an SNR of 0 dB, measured over the entire utterance.

moderately reverberant conditions are shown in Figure 4.5. The moderately reverberant utterance was generated by convolving the clean utterance with a room impulse response with T_{60}=0.5 s and a direct-to-reverberant energy ratio of 1 dB.

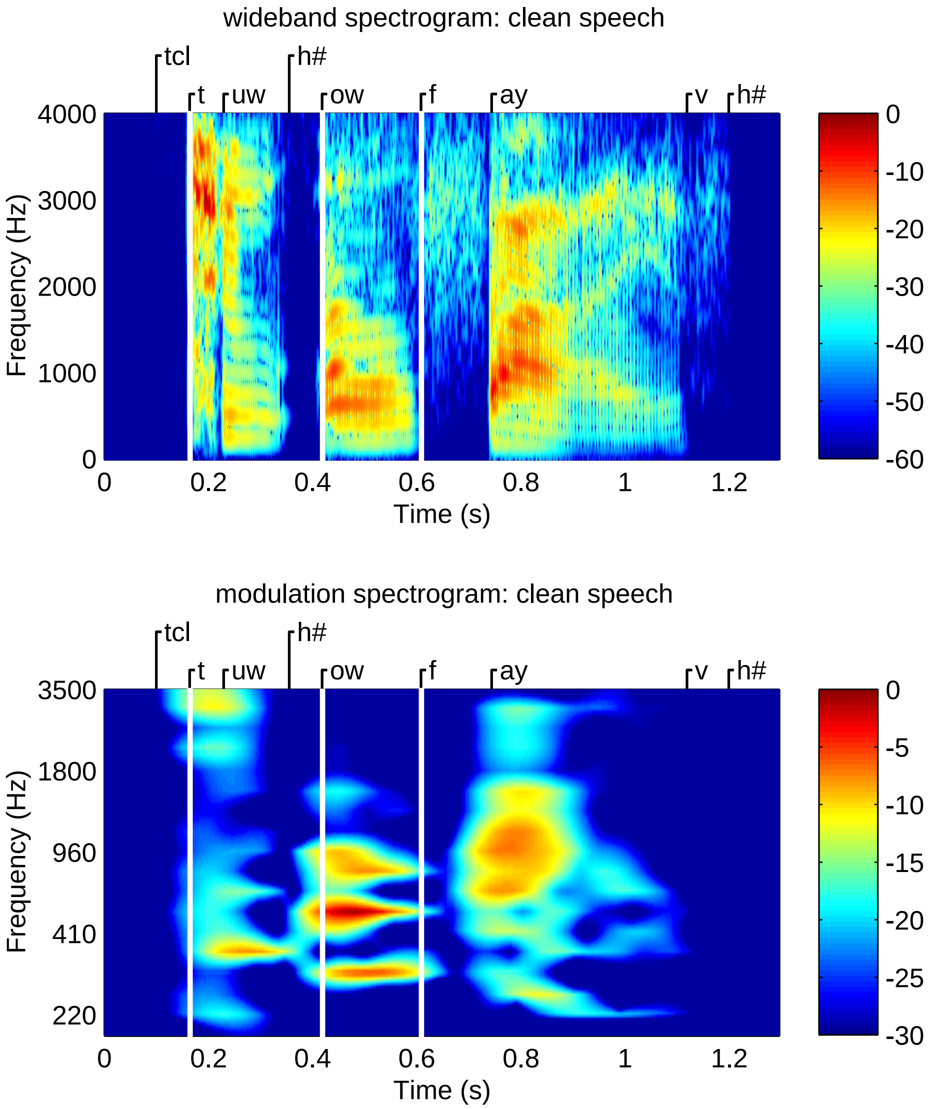


Figure 4.2: Wideband spectrogram and modulation-filtered spectrogram for the clean version of the utterance "two oh five," collected from a female speaker over the telephone.

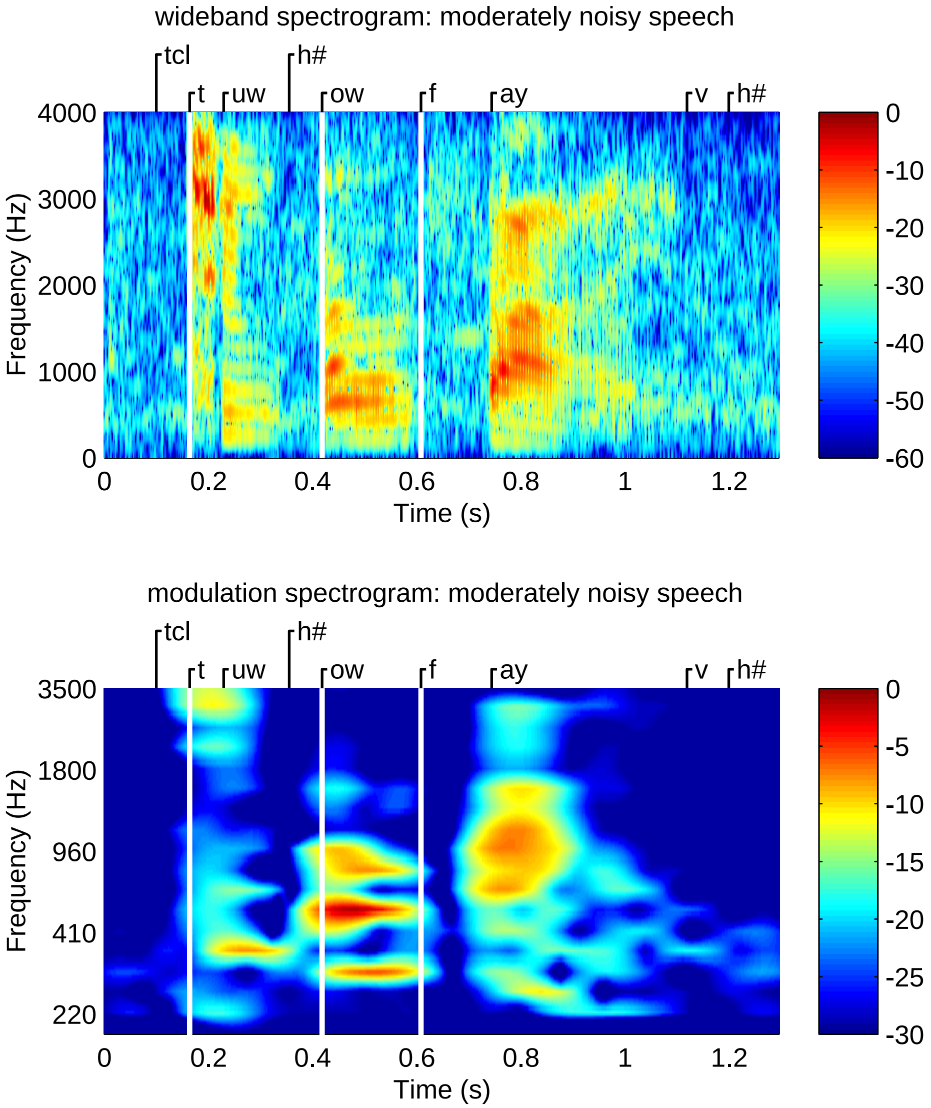


Figure 4.3: Wideband spectrogram and modulation-filtered spectrogram for the moderately noisy version of the utterance "two oh five," collected from a female speaker over the telephone. To generate this utterance, babble noise from the NOISEX CD-ROM was added to the clean utterance at an SNR of 20 dB, measured over the entire utterance.

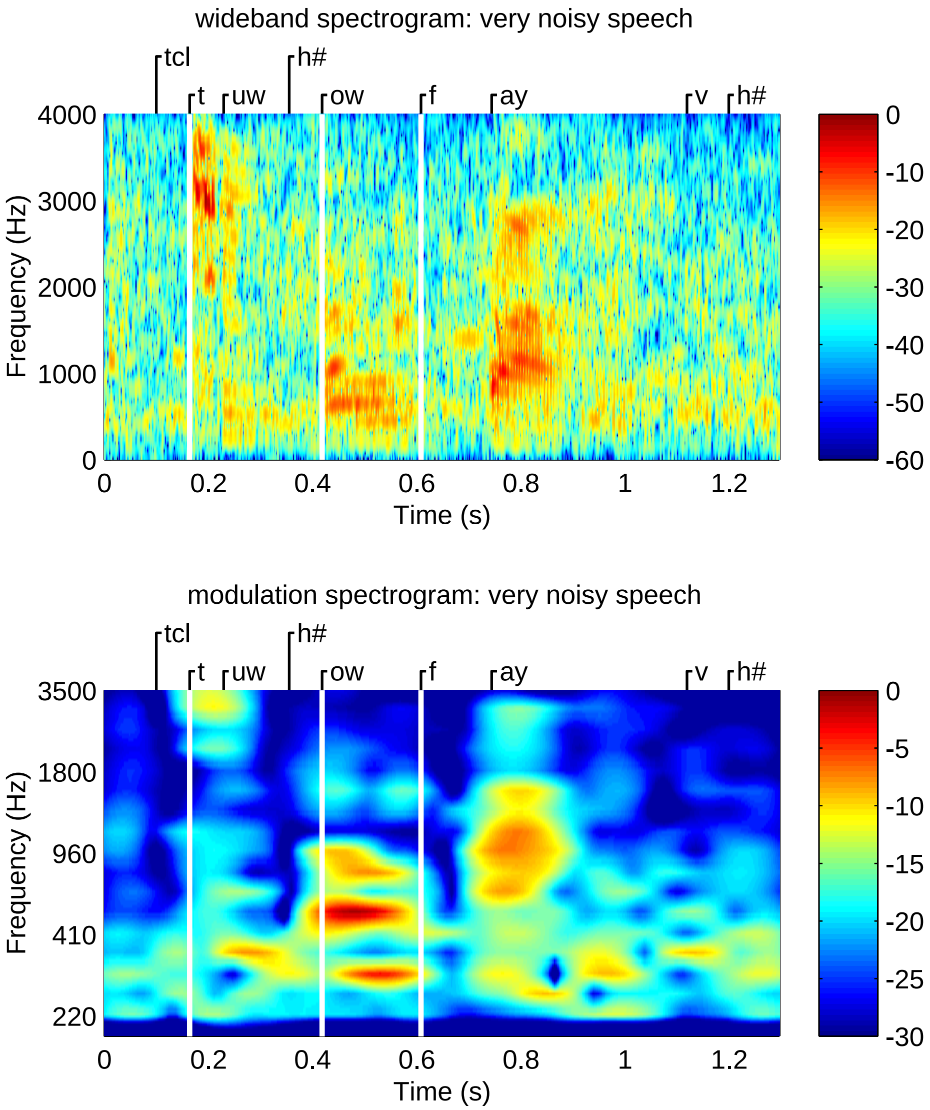


Figure 4.4: Wideband spectrogram and modulation-filtered spectrogram for the very noisy version of the utterance "two oh five," collected from a female speaker over the telephone. To generate this utterance, babble noise from the NOISEX CD-ROM was added to the clean utterance at an SNR of 0 dB, measured over the entire utterance.

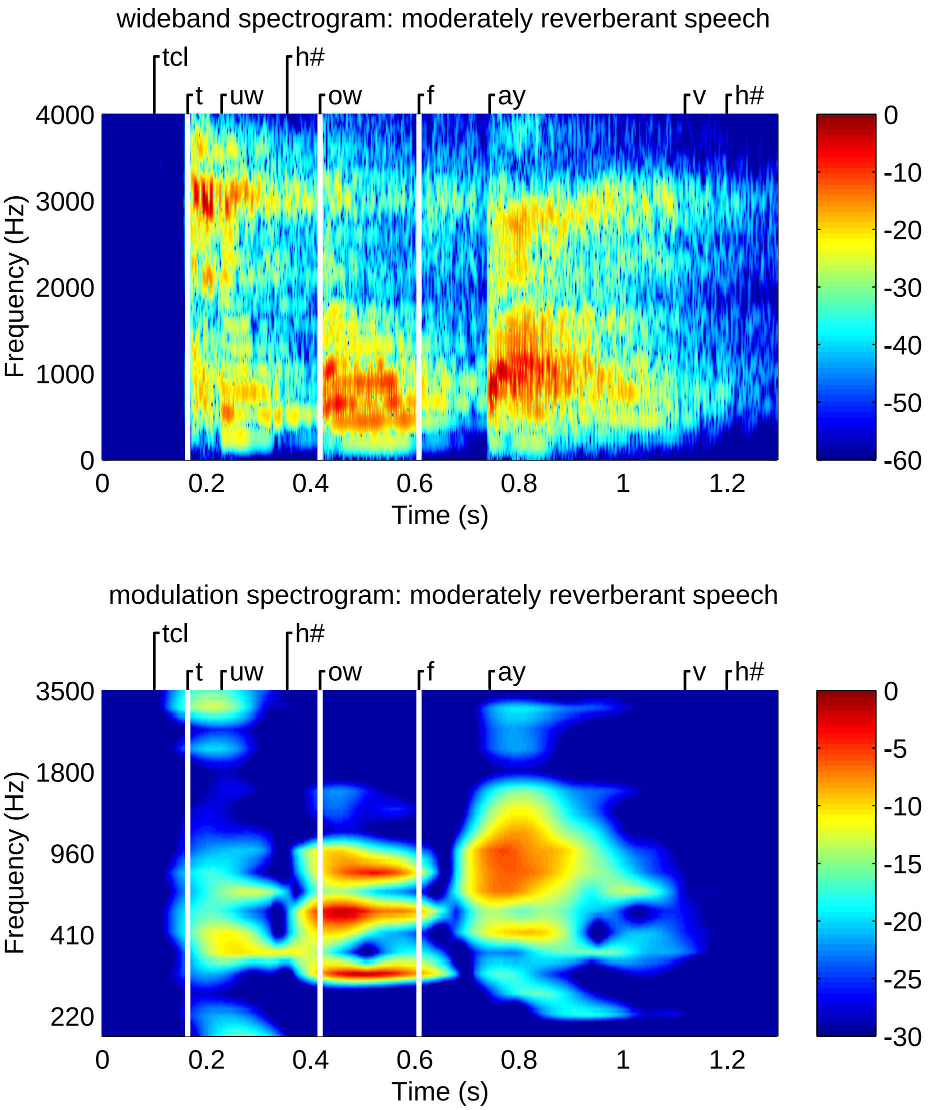


Figure 4.5: Wideband spectrogram and modulation-filtered spectrogram for the moderately reverberant version of the utterance “two oh five,” collected from a female speaker over the telephone. To generate this utterance, the clean utterance was convolved with a room impulse response with T_{60}=0.5 s and a direct-to-reverberant energy ratio of 1 dB.

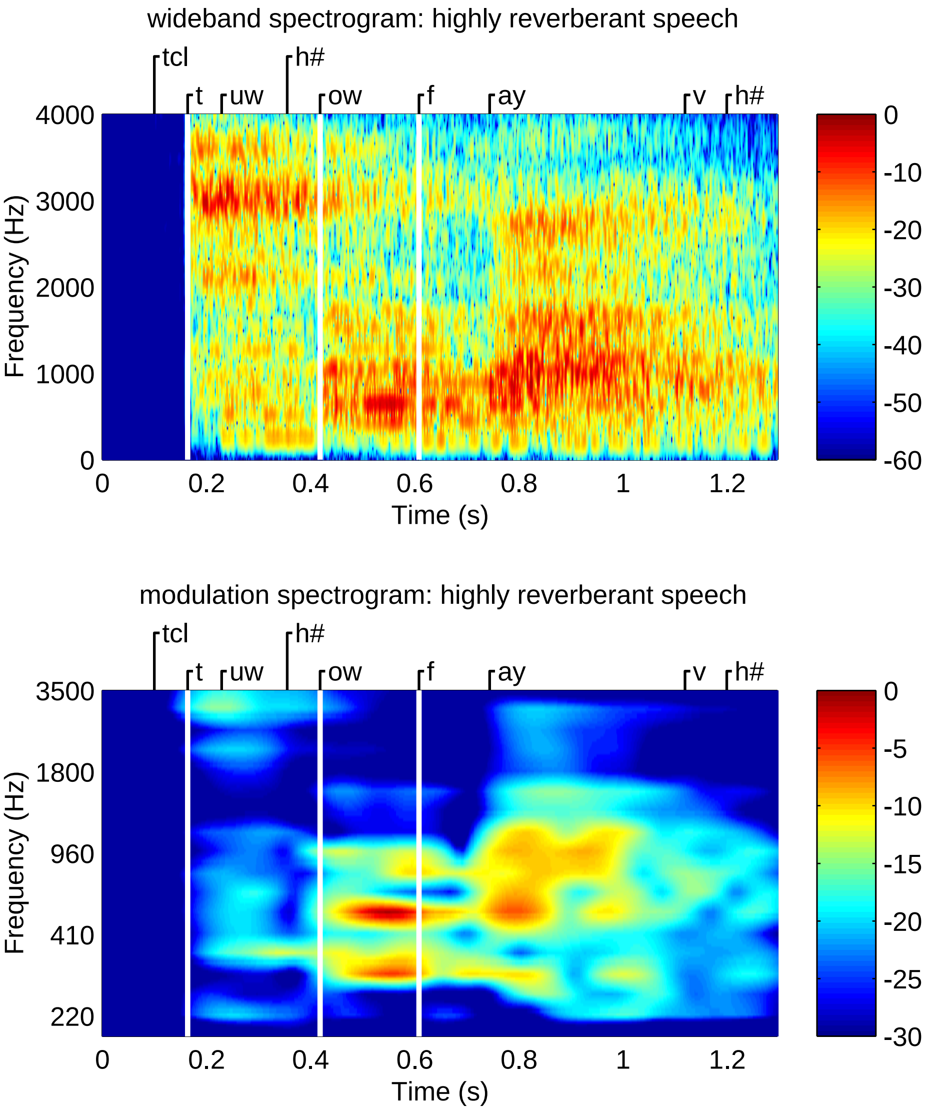


Figure 4.6: Wideband spectrogram and modulation-filtered spectrogram for the highly reverberant version of the utterance “two oh five,” collected from a female speaker over the telephone. To generate this utterance, the clean utterance was convolved with a room impulse response with T_{60}=2.2 s and a direct-to-reverberant energy ratio of -16 dB.

highly reverberant conditions are shown in Figure 4.6. The very reverberant utterance was generated by convolving the clean utterance with a room impulse response characterized by a T_{60}=2.2 s and a direct-to-reverberant energy ratio of -16 dB.

The displayed utterance was hand-labeled by an experienced phonetic transcriber. The phonetic labeling for the utterance is shown along the top of each display, and the syllable onsets are indicated by vertical bars on each display. The wideband spectrograms were calculated as for Figure 1.2. The speech signal was pre-emphasized with a filter, $H(z) = 1 - 0.94z^{-1}$, and then segmented into 8-ms windows with a 2-ms window step. Next, power spectra were calculated using 256-point FFTs. The power spectra were normalized with respect to the peak level in the signal and then plotted on a color scale with a lower threshold of -60 dB.

The wideband spectrogram and modulation-filtered spectrogram representations are very different from one another. The wideband spectrogram of the clean utterance shows a great deal of spectro-temporal detail in the signal, such as pitch pulses in voiced segments, sharp onsets and formant trajectories. In contrast, the modulation-filtered spectrogram of the clean utterance shows only the gross distribution of slowly modulated speech energy in time and frequency, with a warping of the frequency axis that expands the display of the lower frequencies and compresses the display of the higher frequencies. The fine spectro-temporal detail that appears in the wideband spectrogram for the clean utterance is gradually obscured as the level of noise or reverberation increases. Thus, the wideband spectrogram is not an especially stable speech representation.[3] On the other hand, the modulation-filtered spectrogram is quite stable, with the displays of the clean, moderately noisy and moderately reverberant utterances being nearly identical to one another and the displays of the very noisy and highly reverberant utterances showing strong similarities to the other modulation-filtered spectrograms.

4.2 ASR Experiments with the Visual Features

While it is possible to demonstrate representational stability in visual displays of speech, it is not so simple to show that the linguistically relevant information required

[3] It should be noted that the stability of the wideband spectrograms could be enhanced by using a lower threshold of -30 dB with respect to the global peak, as in the modulation-filtered spectrogram processing. A -60 dB threshold was used for the wideband spectrograms because it is a more standard value for speech displays.

zero
oh
one
two
three
four
five
six
seven
eight
nine
ten
eleven
twelve
thirteen
fourteen
fifteen
sixteen
seventeen
eighteen
nineteen
twenty
thirty
forty
fifty
sixty
seventy
eighty
ninety
hundred
thousand
a
and
double
dash
hyphen

Table 4.2: Vocabulary for the Numbers 93 subset used in initial ASR experiments.

for speech recognition is present using such displays.[4] To see if the processing preserves linguistic information, and to determine if the visual stability of the modulation-filtered spectrogram translates into improved ASR robustness, it was necessary to perform experiments in automatic speech recognition using the modulation-filtered spectrogram as an ASR front end. This section describes these ASR experiments.

4.2.1 Experimental Speech Material

The recognition experiments described in this section were all performed using material from the Numbers 93 subset of the Numbers corpus [CNLD95] from the Center for Spoken Language Understanding at the Oregon Graduate Institute. Numbers is a collection of continuous, naturally spoken numbers excised from spontaneous responses to various census-related queries (e.g., for street addresses, ZIP codes and telephone numbers). A sample utterance from the corpus is “nine double oh one eight.” The responses were recorded from a diverse population over local and long-distance telephone lines, and were digitally sampled at 8 kHz with 16-bit resolution. The Numbers 93 subset had a vocabulary of thirty-six different words. Listed in Table 4.2, they are primarily numbers, including confusable sets such as “six,” “sixteen,” and “sixty,” with a few additional words. The Numbers 93 subset was selected for these experiments because it is a relatively small data set and thus these preliminary experiments could proceed quickly. The utterances used for these experiments were partitioned into a set of 875 training utterances (containing a total of 3315 words) and 657 test utterances (containing a total of 2426 words). Two test sets were used, an unaltered, clean version and a reverberant version created by convolving all of the

[4] After all, a very stable display could be generated by multiplying the input signal by 0, but one would be hard-pressed to recognize speech using such a “representation.”

utterances in the clean version with the impulse response used in the visualization study described in Section 4.1.

4.2.2 Structure of the Experimental Recognizers

The automatic recognition system used in the initial experiments was a hybrid HMM/MLP recognizer, as described in Chapter 3. A multilayer perceptron with a single hidden layer was used to estimate phonetic probabilities from the acoustic input. Unless otherwise stated, the size of the hidden layer was set so that every MLP contained approximately 90,000 weights. The labeling of the training data was optimized using iterative embedded Viterbi training. Forty context-independent phone units were used in the word models. Language modeling was done with a class bigram grammar trained on the utterances used for recognizer training. The language model scaling factor and word transition penalty were optimized over the portion of the training data that was reserved for cross-validation during MLP training.

To speed up the recognition process in these initial experiments, the following simplifications were made to the recognizer:

- An input context window of fifteen frames centered on the current frame was used, and no delta features were used.
- The output layer was a set of sigmoidal units, not a softmax layer.
- The lexicon was simpler than the one described in Chapter 3, having only a single pronunciation for each word in the vocabulary and using fixed minimum phone durations of two states.

These changes reduced the time needed to compute features, to train the MLP, and to perform recognition; however, they also reduced the recognizer accuracy. Thus, only the relative performance of the different recognizers in this set of experiments should be considered. Later experiments, described in Section 4.3 and Chapters 5 and 6, used a more complex recognition system capable of matching the best performance reported for context-independent recognizers on the Numbers corpus.

The PLP, log-RASTA-PLP, and J-RASTA-PLP front ends were all tested in this series of experiments. For all three feature sets, the initial FFT power spectrum was com-

Features	Clean test error rate				Reverberant test error rate			
	total	sub.	del.	ins.	total	sub.	del.	ins.
PLP	15.8%	9.2%	3.2%	3.5%	70.1%	33.5%	33.8%	2.7%
log-RASTA-PLP	14.5%	8.9%	3.0%	2.5%	72.7%	39.3%	31.4%	2.0%
J-RASTA-PLP	15.1%	10.1%	3.2%	1.8%	77.3%	44.5%	30.0%	2.9%
MSG	30.1%	21.0%	6.6%	2.5%	65.2%	41.3%	20.5%	3.4%

Table 4.3: Word error rates on the clean and reverberant Numbers 93 test sets for PLP, log-RASTA-PLP, J-RASTA-PLP, and MSG features. The total error rates are presented and are also broken down in terms of substitutions (sub.), deletions (del.), and insertions (ins.).

puted over a 25-ms window with a 12.5-ms window step. The nine lowest-order cepstral coefficients, including the zero-order coefficient, were used as features for recognition. No delta features were used.

The generation of the MSG features differed slightly from the generation of the modulation-filtered spectrogram displays of speech. The filterbank used for spectral analysis was a constant-Q filterbank covering the frequency range 297–4000 Hz with fifteen quarter-octave bandwidth filters, and the modulation analysis was performed using a complex, lowpass FIR filter with a cutoff frequency of 8 Hz instead of using an FFT and retaining only one frequency bin. These changes do not greatly alter the representation of speech, but they did simplify the design of subsequent ASR experiments. The constant-Q filterbank was simpler to parameterize than a filterbank based on Greenwood's cochlear place-to-frequency map, although it produced a representation with finer resolution in the low frequencies than is necessary or psychoacoustically justifiable. The complex filter analysis was more efficient than the FFT-based analysis. Like the RASTA-PLP front ends, the MSG processing produced output vectors at a rate of one every 12.5 ms.

4.2.3 Baseline Recognition Results

Two sets of experiments were performed to establish initial baseline performance measurements. In the first set, a recognizer was trained on clean training data for each of the four front ends and then tested on clean and reverberant versions of the test set. The results of the clean test indicate how well a front end represents speech under matched acoustic

Features	Clean test error rate				Reverberant test error rate			
	total	sub.	del.	ins.	total	sub.	del.	ins.
PLP	72.5%	34.1%	4.2%	34.2%	48.5%	27.5%	13.4%	7.6%
MSG	45.4%	33.8%	6.8%	4.8%	43.5%	30.0%	10.4%	3.2%

Table 4.4: Word error rates on the clean and reverberant Numbers 93 test sets for PLP and MSG recognizers trained on reverberant data. In this case the reverberant test is the condition matched to the training data.

conditions, for a given MLP, lexicon and language model. The results of the reverberant test indicate how well a front end suppresses the reverberation in its representation of speech — it measures the invariance of the representation. The results of this first experiment are summarized in Table 4.3.

In the second set of experiments, recognizers were trained on a reverberant version of the training set, created by convolving the training utterances with the same impulse response used to create the reverberant test set and then tested on the clean and reverberant test sets. In this experiment, performance on the clean test measures the invariance of the front-end representation. The word error rate on the reverberant test may be considered a lower bound for the reverberant test for a given front end, MLP, lexicon and language model, because it is obtained with matched training data. Only the PLP front end (which was the best of the RASTA-PLP front ends on the reverberant test in the initial experiments) and the MSG features were tested. The results of this experiment are summarized in Table 4.4.

In the first experiment, in which the recognizers were trained on clean speech, the three RASTA-PLP front ends yield essentially identical performance on the clean test, while on the reverberant test the PLP features are significantly better than the log-RASTA-PLP features, and the log-RASTA-PLP features are, in turn, significantly better than the J-RASTA-PLP features. The MSG features are much worse than the PLP features on clean speech, with a word error rate that is nearly twice that of PLP; however, they are significantly better than the PLP features on the reverberant test. Their poor performance on the clean test indicates that the MSG features fail to represent certain information important for recognizing speech. The information that the MSG features do represent, though, is relatively resistant to the effects of reverberation. The results of the experiment in which recognizers were trained on reverberant speech support these conclusions. Even

Subject	total	sub.	del.	ins.
A	146	108	25	13
B	153	103	35	15
C	145	91	40	14
average	148	100	33	14
error rate	6.1%	4.1%	1.4%	0.6%

Table 4.5: Number of errors (out of 2426 words) made by human listeners on the reverberant Numbers 93 test. Percent error rates are also given for the averages. The average substitutions, deletions, and insertions do not sum to 148 due to rounding.

with matched training and testing conditions, performance on the reverberant test is quite poor, with total word error rates of 44–49%. This drop in performance occurs because temporal smearing of the speech signal increases the variability of the acoustic realizations of different phonetic segments. In the case of clean speech, the acoustics are influenced primarily by the identity of the current phonetic segment and by the identities of the preceding and succeeding phonetic segments (ignoring speaker characteristics). However, under reverberant conditions the acoustics are strongly influenced by additional phonetic segments due to temporal smearing. This effect will make recognition more difficult, especially for relatively small training sets.

4.2.4 Measuring Human Performance on the Reverberant Test

To obtain a second, independent measurement of the difficulty of the reverberant speech recognition task, the intelligibility of the reverberant test set was measured for three human listeners. The subjects, who were all native speakers of American English, had no known hearing impairments, and had considerable experience in the phonetic transcription of speech, were asked to lexically transcribe all 657 utterances in the reverberant test set. The utterances were generated using the 16-bit digital-to-analog converter in a Sun SPARC-5 workstation at a sampling rate of 8 kHz and presented over headphones at a comfortable level in a quiet office. The subjects were given a list of the thirty-six possible words in the test set, and were allowed to listen to each utterance as many times as desired. The order of the utterances was randomized to prevent the subjects' adjusting to the characteristics of the speakers in the database. To familiarize the subjects with the transcription task, they were given a practice set of ten reverberant utterances from the training set to transcribe before the actual test session commenced. During the practice session the subjects were given feedback on their transcription accuracy, but they did not receive any feedback in the test phase.

The results of the listening test are summarized in Table 4.5. The performance of each individual subject is given, as well as an average for all three subjects. The average word error rate for the three subjects is 6.1%, which is ten times more accurate than the best ASR system (trained on clean speech) on the reverberant test and over two times better than the best ASR system on the clean test. This experiment illustrates that the recognition task is a challenging, but not impossible one, for human listeners.

4.2.5 Variants on the Modulation-Filtered Spectrogram Features

Although the performance on clean speech with the new features was poor in comparison to that obtained using the PLP and RASTA-PLP features, the new features did provide a significant improvement in recognition accuracy on the reverberant test. To gain a better understanding of what signal-processing steps contribute to the improvement in reverberation and to try to improve the overall utility of the MSG features for ASR, a number of processing variations were tested. All of the tests used the same recognition architecture as the baseline study and all of the recognizers were trained on the clean training set only.

Using Higher Modulation Frequencies

The modulation filtering used in the initial version of the MSG processing is quite severe, passing only frequencies below 8 Hz. In comparison, the filter used in RASTA-PLP has a cutoff frequency of 12 Hz and the lowpass filters applied to the envelope signals in channel vocoders typically have cutoff frequencies of 20–25 Hz. To ascertain whether the higher modulation frequencies were useful in this case, two experiments were performed. First, the complex modulation filter was changed to have a passband of 8–16 Hz, and the features generated using this bandpass filter were used for recognition. Second, the features generated with both the lowpass modulation filter and the bandpass modulation filter were used together for recognition, doubling the number of features per frame. In the second

Filter	Clean test error rate				Reverberant test error rate			
Passband	total	sub.	del.	ins.	total	sub.	del.	ins.
8–16 Hz	47.1%	30.7%	11.5%	4.9%	79.1%	47.4%	26.5%	5.1%
0–8 and 8–16 Hz	29.8%	20.7%	4.9%	4.2%	70.1%	45.8%	18.4%	5.9%

Table 4.6: Word error rates on the clean and reverberant Numbers 93 test sets for MSG features generated using a bandpass modulation filter with a 8–16 Hz passband and for the standard features and bandpass features used together at the input to a single MLP.

Clean test error rate				Reverberant test error rate			
total	sub.	del.	ins.	total	sub.	del.	ins.
31.0%	21.9%	6.4%	2.7%	66.5%	44.7%	17.1%	4.7%

Table 4.7: Word error rates on the clean and reverberant Numbers 93 test sets for MSG features with spectral smoothing accomplished via truncation of the DCT of the spectral features.

experiment, the size of the hidden layer of the MLP in the recognizer was reduced to keep the total number of MLP weights approximately constant.

The results of these two experiments are summarized in Table 4.6. The recognition systems based on the bandpass features alone are significantly less accurate than those based on the original, lowpass features (Table 4.3). Using the two feature sets together does not provide a significant improvement on the clean test and leads to a significant decrease in accuracy on the reverberant test. Thus, for this specific combination of signal processing, ASR system and testing regime, the 8–16 Hz modulations do not appear to carry useful linguistic information.

Applying Cepstral Smoothing

It is generally believed that a spectral representation with critical-band frequency resolution provides more detail than is necessary for speech recognition, and that using a lower-resolution representation may improve the speaker independence of an ASR system [Kla82, Her90]. To test if such spectral smoothing would improve the accuracy of the MSG features, a discrete cosine transform was applied to the original features, and only the nine low-order coefficients, including the zero-order coefficient, were used for recognition.

The results of this experiment are summarized in Table 4.7. The performance on both tests was not significantly different from the MSG baseline recognizer. While the smoothing did not lead to any performance improvements, this experiment did demonstrate that it might be possible to reduce the number of features used for recognition without significantly impacting system performance. The cepstral transform can also be useful in an ASR system based on a Gaussian mixture acoustic model, to the extent that the transform decorrelates the features, making them a better match to the diagonal-covariance Gaussians which are typically used in mixture models.

Omitting Various Stages of the Processing

To develop a clearer picture of which steps in the MSG processing contributed to robustness in reverberation, a series of recognition experiments were performed using MSG variants that omitted different signal-processing steps. The steps that could be omitted were the normalization of each amplitude envelope signal by its average level, filtering of the envelope by the complex, lowpass modulation filter, normalization of all amplitude signals with respect to the global peak level and thresholding of levels more than 30 dB below the global peak level to -30 dB. Note that if all four steps are omitted, the signal processing generates a short-term log power spectrum with quarter-octave resolution.

The outcomes of these experiments are summarized in Table 4.8. These data illustrate a number of important points:

- The thresholding operation (T), vital for producing stability in the visual displays, is detrimental to the accuracy of an ASR system on both the clean and reverberant tests. In comparing all pairs of experiments that differ only by the presence or absence of the thresholding (experiments 0 and 3, 1 and 5, and 2 and 4), it is evident that omission of the thresholding drastically improves accuracy on the clean test and either has an insignificant effect on accuracy for the reverberant test or results in slight improvements. It is likely that the thresholding operation masks low-energy parts of the speech signal that carry important phonetic information. A comparison of experiments 0 and 1 supports this hypothesis. If the thresholding is performed and the normalization of the amplitude signals by their average levels is not performed, then the recognition accuracy on both tests is significantly degraded. The normalization

#	A	F	P	T	Clean test error rate				Reverberant test error rate			
					total	sub.	del.	ins.	total	sub.	del.	ins.
0	+	+	+	+	30.1%	21.0%	6.6%	2.5%	65.2%	41.3%	20.5%	3.4%
1	-	+	+	+	40.0%	27.5%	9.2%	3.4%	70.7%	38.2%	28.9%	3.7%
2	+	-	+	+	30.6%	21.2%	7.0%	2.4%	67.8%	42.6%	21.6%	3.6%
3	+	+	+	-	17.5%	11.8%	3.5%	2.1%	66.1%	37.7%	25.9%	2.6%
4	+	-	+	-	13.6%	9.2%	2.1%	2.2%	69.9%	40.8%	24.6%	4.4%
5	-	+	+	-	17.8%	12.3%	3.0%	2.6%	63.8%	32.3%	28.3%	3.4%
6	-	+	-	-	18.3%	12.0%	2.7%	3.6%	68.8%	30.7%	34.4%	3.8%
7	-	-	-	-	16.1%	10.5%	2.7%	2.9%	73.5%	31.8%	39.0%	2.7%

Table 4.8: Word error rates on the clean and reverberant Numbers 93 test sets for MSG variants that omit one or more of the key processing steps: (A) normalization of the amplitude envelope signals by their average levels; (F) filtering of the amplitude envelope signals with a lowpass, complex filter; (P) normalization of the amplitude signals by their global peak value; (T) thresholding of all values more than 30 dB below the global peak to -30 dB. A “+” indicates that the processing step is included in a given experiment, while a “-” indicates that the step is omitted. The results from the baseline experiment are reiterated here as experiment 0, to facilitate comparison with the other results.

by average levels tends to flatten the spectrum of the speech, reducing the differences between high-energy and low-energy bands. If this normalization is not performed, more of the speech signal is masked by the thresholding and recognition accuracy is reduced.

- The modulation filtering operation (F) is crucial for accuracy on the reverberant test, but decreases accuracy on the clean test. Comparing the experiments that differ only by the presence or absence of the filtering (experiments 0 and 2, 3 and 4, and 6 and 7), its inclusion always provides a significant improvement in accuracy on the reverberant test, but in two of the three tests it also causes a significant loss of accuracy on the clean test.

- Automatic gain control (AGC) may be beneficial or detrimental, depending on the acoustic conditions of a test set and the specific properties of the AGC. Comparing experiments 3 and 5, it appears that without the thresholding, the normalization of the amplitude envelopes by their average levels (A) has a detrimental effect for the reverberant test and no significant effect on performance for the clean test. Comparing experiments 5 and 6 shows that the normalization of the amplitude signals by the global peak level in an utterance (P) significantly improves accuracy on the reverberant test and has no significant effect on accuracy for the clean test. The best performance on the clean test is obtained in experiment 4, with a front end that includes both normalization steps, but no other processing. Utterances in the "clean" test were collected over many different telephone handsets and lines. The two normalization/AGC steps are useful for reducing the variability resulting from the spectral shaping imposed by these different channels.

Changing the Modulation Filter

The role of the modulation filtering was examined in more detail by splitting the complex filter into its real and imaginary parts and using the parts either separately or together for speech recognition. The outputs of the real and imaginary filters were compressed using a cube-root function instead of a logarithm. These variations on the filter were tested with a signal-processing system that performed neither thresholding nor normalization of the amplitude signals by their average levels. For the cube-root compressed signals, normalization with respect to the global peak was performed by finding the point in an utterance with the greatest absolute value and dividing all points in the signal by that magnitude. In the experiment where the real and imaginary parts of the original complex filter were used together, one normalization factor was used for the outputs of the real filters and a separate one used for the outputs of the imaginary filters (as illustrated in Figure 4.9). As a control, an experiment was run in which the recognition features were computed by taking the cube root of the magnitude of the output of the complex filter.

The results of these experiments are summarized in Table 4.9. Changing the compression from logarithmic to cube root significantly degrades recognition accuracy on the reverberant test, but has no significant effect on accuracy on the clean test. With cube root compression, there is no significant difference in performance on either test between using the complex filter and only its real part. Using the imaginary part of the filter gives a significant improvement in accuracy on the reverberant test and no significant change on the clean test. Using both filters in parallel, however, produces a significant improvement in accuracy for the clean test over all the other configurations, and on the reverberant test achieves an accuracy on par with that obtained using the complex filter with log compression.

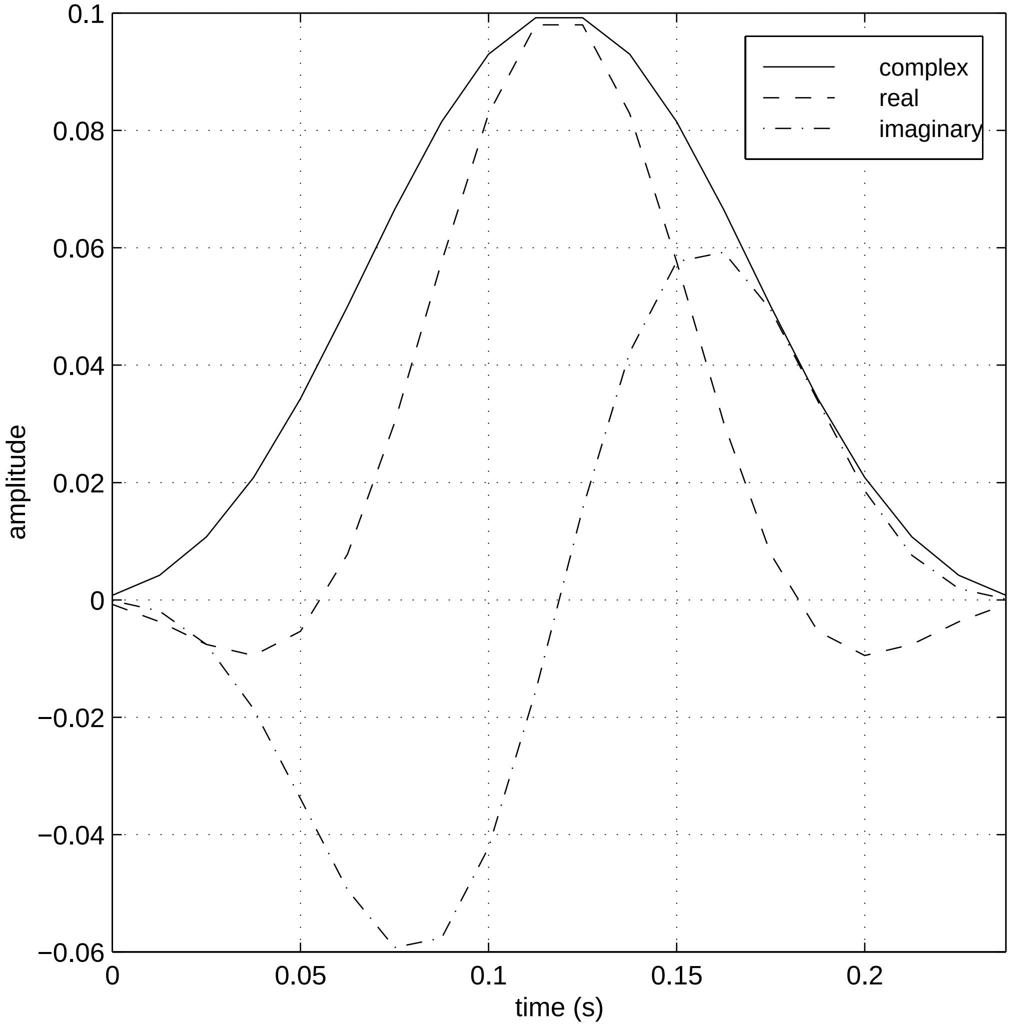


Figure 4.7: The magnitude of the impulse response of the complex filter and the impulse responses of its real and imaginary components. The complex and real filters are lowpass, with the complex filter having a broader response than the real filter. The imaginary filter is a time-local differentiator.

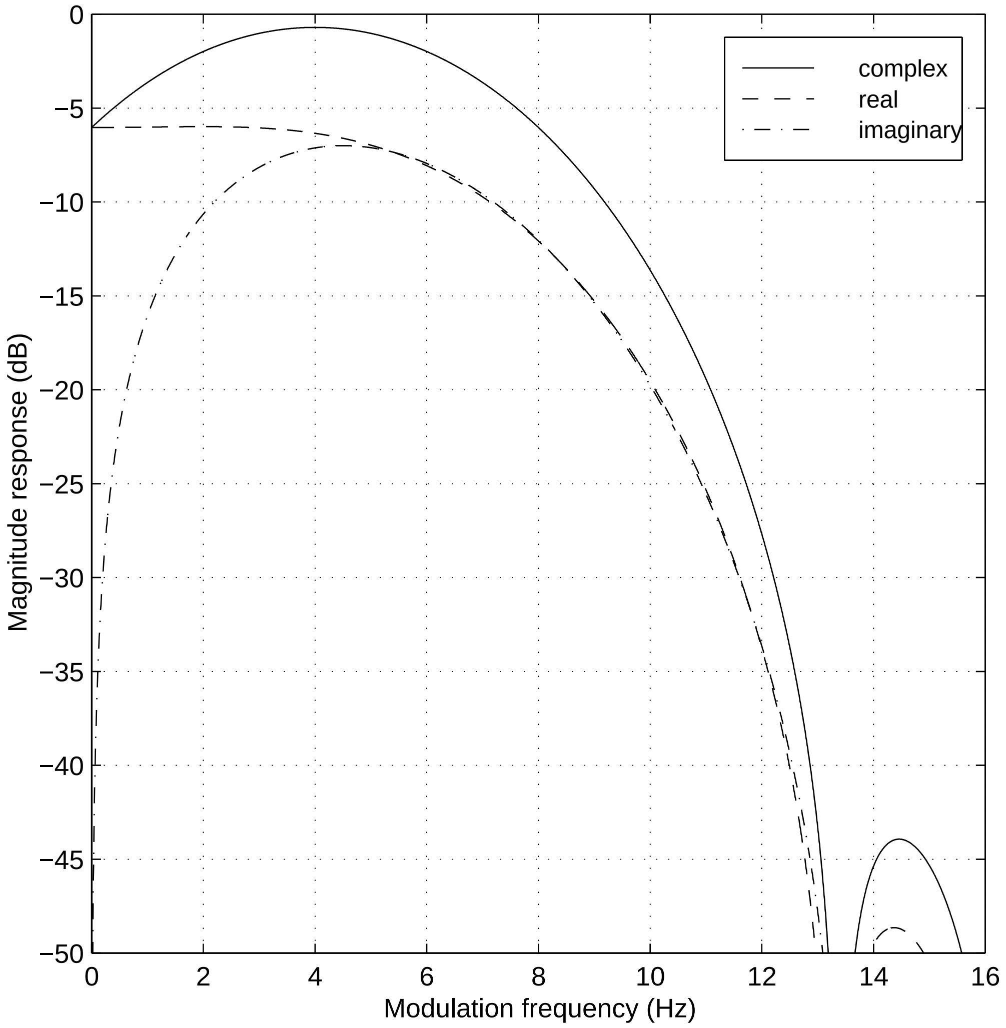


Figure 4.8: The frequency responses of the complex filter and its real and imaginary components. The complex filter is lowpass, with some degree of attenuation at 0 Hz. The real filter is strictly lowpass. The imaginary filter is bandpass with a zero at 0 Hz.

Filter	Compress.	Clean test error rate				Reverberant test error rate			
		total	sub.	del.	ins.	total	sub.	del.	ins.
complex	log	17.8%	12.3%	3.0%	2.6%	63.8%	32.3%	28.3%	3.4%
complex	cube root	17.8%	12.2%	4.0%	1.6%	67.2%	34.8%	29.5%	2.9%
real part	cube root	16.5%	11.1%	2.6%	2.7%	68.3%	33.9%	30.9%	3.4%
imaginary part	cube root	17.3%	12.2%	2.6%	2.5%	64.3%	40.9%	14.3%	9.1%
real and imaginary part	cube root	14.7%	9.8%	2.3%	2.6%	63.5%	37.9%	15.7%	9.9%

Table 4.9: Word error rates for the clean and reverberant Numbers 93 test sets obtained using different modulation filters.

These results may be explained by examining the properties of the three filters. The magnitude of the impulse response of the complex filter and the impulse responses for the real and imaginary filters are shown in Figure 4.7. The frequency responses of the three filters are shown in Figure 4.8. Using the real and imaginary components of the complex filter separately has two advantages. First, the real filter has a slightly narrower temporal response than the complex filter. Second, and more importantly, the use of the real and imaginary filters provides two distinctly different representations of the speech signal to the recognizer. The one produced by the real filter is essentially a normalized, temporally smoothed spectral representation, while the one produced by the imaginary filter is a normalized, smoothed, and differentiated spectral representation.

Summary

These initial experiments with MSG variants demonstrated how the signal processing could be changed to produce a more useful representation for ASR. Most important is the elimination of the thresholding, which leads to dramatic improvements in recognition accuracy on clean speech without significantly affecting recognition accuracy on reverberant speech. Other important changes from the original visual representation include the elimination of the normalization of the amplitude envelope signals by their average levels, the replacement of the complex FIR filter by its real and imaginary components, which are used in parallel and the replacement of the log compression of the filtered amplitude envelopes by cube-root compression. Figure 4.9 depicts the resultant signal-processing system.

4.3 Combining MSG and RASTA-PLP Features

The performance of the MSG features in combination with other features is of great interest because the use of multiple representations in combination frequently yields more accurate recognizer performance. The MSG processing, as outlined in Figure 4.9, is somewhat distinct from RASTA-PLP processing, so a recognizer that uses both representations in combination may be more accurate than a recognizer that uses only one of the two front ends. This possibility was tested in another series of recognition experiments.

4.3.1 Experimental Speech Material

These experiments, and almost all subsequent experiments, used a larger subset of the Numbers corpus [CNLD95] known as Numbers 95. Utterances included in this subset were chosen to have valid phonetic transcriptions and to not contain any words truncated by the segmentation routine used to isolate the numbers from their carrier utterances. The subset vocabulary of thirty-two different words was the same as that listed in Table 4.2, but without the words “a,” “and,” “double,” “dash,” “hyphen,” and “thousand.” The subset was divided into a training set of 3590 utterances (containing a total of 13873 words), a development test set of 1206 utterances (containing a total of 4673 words) and a final test set of 1227 utterances (containing a total of 4757 words). The final test set was not used in the following series of experiments.

Recognizers were tested on six different versions of the development test set: a clean version, a reverberant version and four noisy versions. The four noisy test sets were generated by adding pink noise from the NOISEX CD-ROM to the clean set at signal-to-noise ratios of 30 dB, 20 dB, 10 dB and 0 dB. The signal-to-noise ratio was set on an utterance-by-utterance basis and was measured over an entire utterance. The reverberant test set was generated by convolving the clean test set with a room impulse response characterized by a T_{60} of 0.5 s and a direct-to-reverberant energy ratio of 1 dB. The early-to-late energy ratio for this impulse response, counting all sound arriving within 80 ms of the arrival of the direct sound as contributing to intelligibility, is 22 dB. The impulse

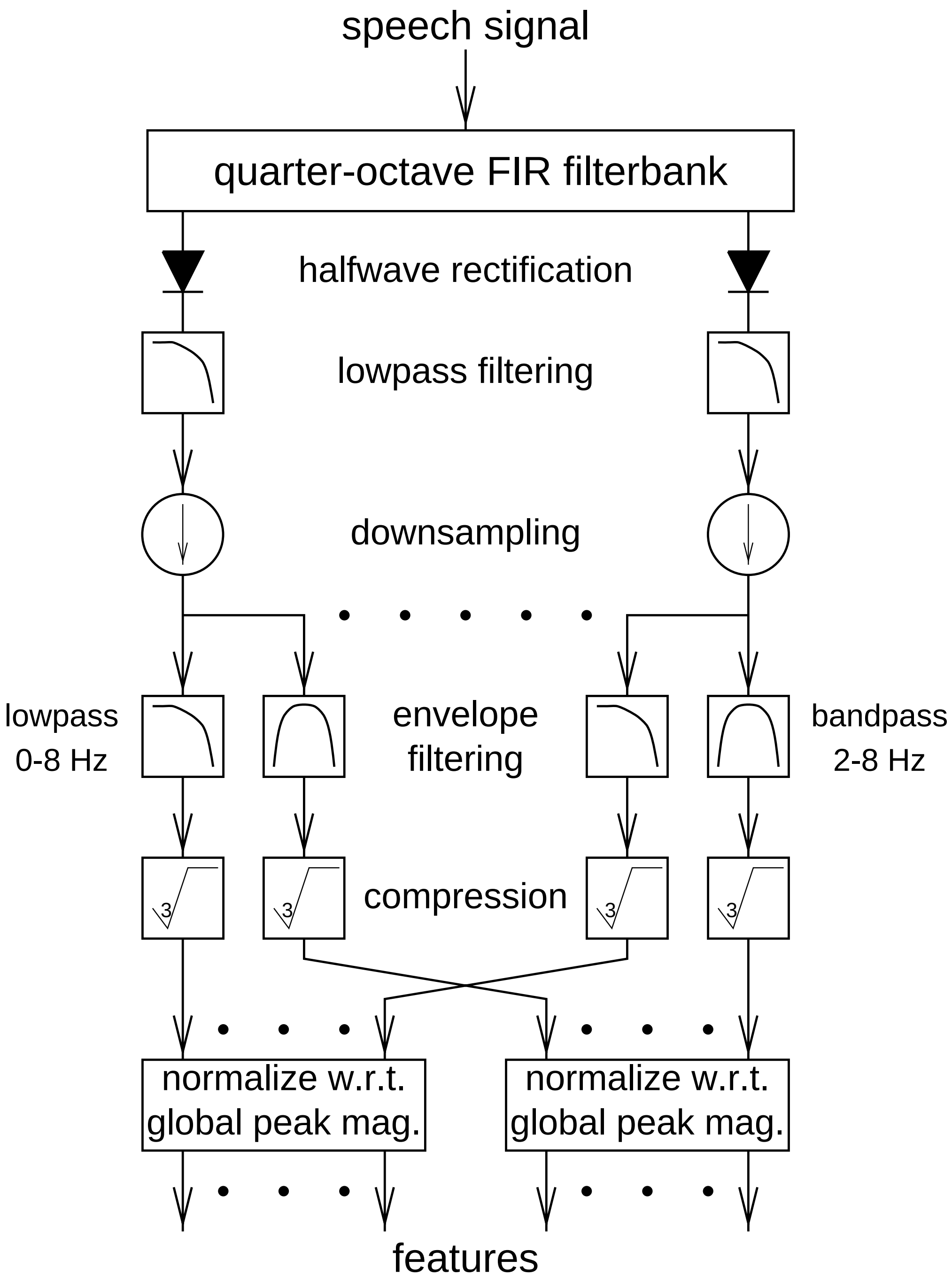


Figure 4.9: Diagram of the signal processing that produces an optimized form of the modulation-filtered spectrogram features, based on the experiments with the Numbers 93 subset.

Subject	clean	reverb.
1	0.3%	0.5%
2	0.3%	0.0%

Table 4.10: Word error rates for human listeners for 100 utterances from the clean Numbers 95 development test set and 100 different utterances from the reverberant Numbers 95 development test.

response is one of a set of impulse responses collected in the Bell Labs Varechoic chamber [MPC97, WEKM94, Ave97a].[5] The varechoic chamber is a 6.71 m × 5.94 m × 2.74 m room in which the walls, floor, and ceiling are covered by a total of 368 individually controlled panels with variable acoustic absorbance. The panels contain two perforated metal sheets that may be positioned so that the holes align with one another to expose sound-absorbing material behind the sheets, creating a highly absorbant surface. The sheets may also be positioned so that the holes in the top sheet are entirely occluded by the lower sheet, creating a highly reflective surface. Each panel may be set to the highly reflective or highly absorbant state. For the collection of the impulse responses used in this thesis, four omnidirectional microphones were placed at distances of 2 m, 2.35 m, 2.7 m, and 3.05 m from a source, and measurements were recorded from each microphone for one of three panel settings: 100% of the panels open, 43.7% open, or none open [Ave97a]. The room impulses were measured using a chirp-excited system identification program. For the current experiment as well as the bulk of the experiments in this thesis, the impulse response for the microphone located 2 m from the source with 43.7% of the panels open was used. The remaining eleven impulse responses were reserved for final tests, described in Chapter 6.

Intelligibility of the Numbers 95 Utterances

The intelligibility of the Numbers 95 utterances for human listeners in both the clean and reverberant conditions was measured. Two native speakers of American English with no known hearing impairments lexically transcribed 100 utterances from the clean development test set and 100 different utterances from the reverberant test set. The tests were counterbalanced so that the clean utterances heard by one listener were the same as

[5] I am most grateful to Jim West, Gary Elko, and Carlos Avendaño for collecting the impulse responses and making them available to me.

the reverberant utterances heard by the other listener, and vice-versa. The utterances were generated using the 16-bit digital-to-analog converter in a Sun SPARC-5 workstation at a sampling rate of 8 kHz and presented over headphones at a comfortable listening level in a quiet office. The listeners could hear each utterance up to four times. The subjects' transcriptions were scored using the same program that was used to score recognizer output. The error rates for each subject, in each condition, are shown in Table 4.10. Clearly, both conditions are trivial for human listeners.

4.3.2 Structure of the Experimental Recognizers

The automatic speech recognition system used in these experiments was also a hybrid HMM/MLP recognizer. The MLP used for phonetic probability estimation had an input context window of nine frames centered on the current input, a single hidden layer and a softmax output layer. Unless otherwise specified, the size of the hidden layer was set so that every MLP had approximately 106,000 weights. For recognizers in which two front ends are combined, the combination was accomplished by averaging the phone log likelihoods from the MLPs. The lexicon used in these experiments was a multiple-pronunciation lexicon with simple context-dependent phone duration modeling.[6] The labeling of the training data and the pronunciations and duration constraints in the lexicon were optimized using iterative embedded Viterbi training. Thirty-two context-independent phone units were used in the word models. Language modeling was done with a backoff bigram grammar trained on the utterances used for recognizer training. The language model scaling factor and word transition penalty were fixed, based on a set of pilot experiments. Recognizer training was performed only on clean speech.

The PLP, log-RASTA-PLP, J-RASTA-PLP, and MSG front ends were all tested. The PLP and RASTA-PLP feature calculations were based on an FFT power spectrum computed over a 25-ms window with a 10-ms window step. The nine lowest-order cepstral coefficients, including the zero-order coefficient, were used as features for recognition, supplemented with delta features computed via a regression over a nine-frame window centered on the current frame. The MSG features were calculated as shown in Figure 4.9. A down-sampling factor of 80 was used, so that the MSG front end generated one feature vector every 10 ms, like the RASTA-PLP front ends.

[6] Thanks to Su-Lin Wu for creating this lexicon.

Features	Test condition					
	clean	reverb.	30 dB SNR	20 dB SNR	10 dB SNR	0 dB SNR
PLP	6.4%	37.6%	28.3%	43.5%	60.7%	78.8%
log-RASTA	6.4%	26.0%	11.4%	16.3%	27.8%	51.6%
J-RASTA	6.6%	27.9%	15.6%	23.5%	35.7%	54.4%
MSG	8.5%	27.3%	14.6%	22.9%	38.7%	61.5%

Table 4.11: Word error rates on the clean, reverberant, and noisy Numbers 95 development test sets for recognizers using a single front-end representation.

4.3.3 Baseline Results

The results using a single front-end representation are summarized in Table 4.11. Except for the clean test, where all three RASTA-PLP front ends have a similar performance level, the log-RASTA front end is significantly more accurate than the other front ends on all tests. The J-RASTA front end is the runner-up, followed by the MSG front end. On the moderately reverberant test, the 30-dB-SNR noisy test and the 20-dB-SNR noisy test, the MSG and J-RASTA-PLP front ends are equally accurate. The PLP front end, which is the only front end that does not perform any temporal processing of the incoming signal, is the least accurate front end (except on the clean test, where it outperforms the MSG front end).

These results are somewhat unexpected in light of the earlier tests with the smaller Numbers 93 subset, where the optimized version of the MSG front end performed as well as the three RASTA-PLP front ends on the clean test and significantly better on the highly reverberant test (see Table 4.3 for the various RASTA-PLP results and the last line of Table 4.9 for the optimized MSG results). The difference may arise because the earlier tests were performed using a simple lexicon that had not been optimized via embedded Viterbi training, while the tests on the larger data set were performed with a lexicon that was optimized via embedded Viterbi training using a recognizer with log-RASTA-PLP features.

4.3.4 Combining Results

The results for combining two front-end representations are summarized in Table 4.12. Because the combination of two front ends, as implemented here, doubles the

Features	Test condition					
	clean	reverb.	30 dB SNR	20 dB SNR	10 dB SNR	0 dB SNR
PLP and log-RASTA	5.7%	26.9%	15.9%	26.6%	43.7%	67.3%
PLP and MSG	6.1%	29.1%	20.5%	36.1%	53.5%	71.3%
log-RASTA and MSG	5.5%	20.1%	10.4%	14.7%	23.2%	44.7%
log-RASTA, double num. MLP weights	5.9%	26.1%	10.8%	16.4%	29.7%	54.7%
MSG, double num. MLP weights	8.2%	27.9%	14.4%	22.1%	39.8%	65.3%

Table 4.12: Word error rates on the clean, reverberant, and noisy Numbers 95 development test sets for recognizers using a combination of two front-end representations and for recognizers with twice as many MLP parameters (ca. 212,000 weights) and a single front-end representation.

number of MLP parameters used in the recognizer (by doubling the number of MLPs), results for using the log-RASTA and MSG front ends alone, but with twice as many parameters in the MLP, are also shown. In these tests the combination of log-RASTA and MSG features is significantly better than combinations of PLP and log-RASTA or PLP and MSG features, except on the clean test, where all three combinations have roughly identical performance. It should also be noted that the log-RASTA and MSG combination is the only one to perform significantly better than the log-RASTA baseline for this data set.

The success of the log-RASTA and MSG combination may arise from two factors. First, both front ends incorporate modulation filtering to enhance robustness, while the PLP front end does not. Second, the log-RASTA and MSG front ends have somewhat different temporal characteristics, as illustrated in Figure 4.10. The lowpass portion of the MSG representation has the largest output in syllable nuclei, while the bandpass portion of the MSG representation shows strong responses to onsets and offsets occurring on phonetic-segment time scales. In contrast, log-RASTA-PLP shows strong onset responses, followed by a gradual decay, and moderate offset responses. The PLP and lowpass portion of the MSG representation are somewhat similar in that both respond most strongly to syllable

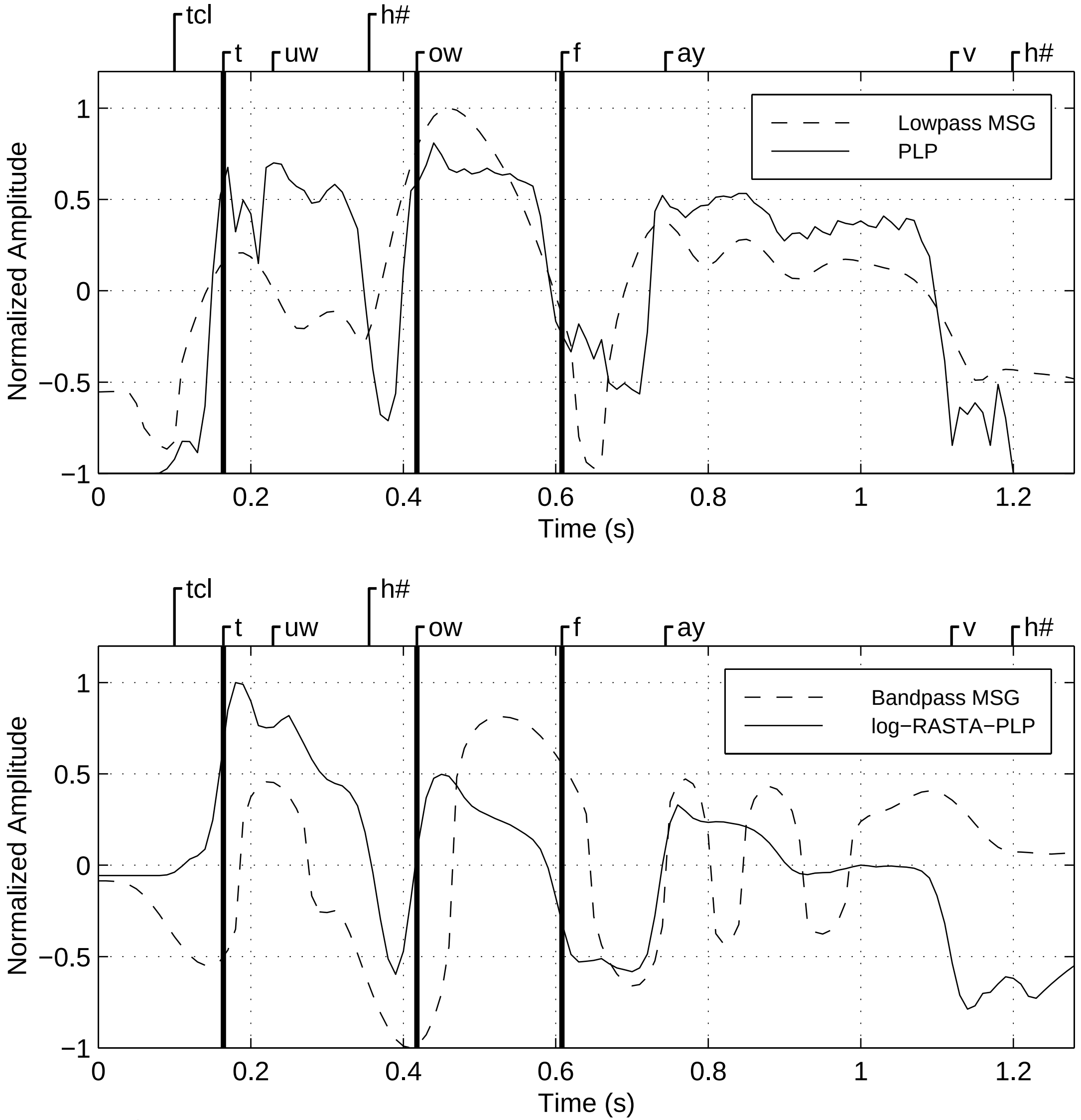


Figure 4.10: A comparison of the temporal characteristics of the lowpass MSG, bandpass MSG, PLP, and log-RASTA-PLP representations. The graphs show the temporal evolution of the output for a single frequency channel (ca. 600–700 Hz for all representations) for the clean utterance “two oh five,” collected from a female speaker over the telephone. The PLP and log-RASTA-PLP features were obtained by converting cepstral coefficients back into spectra. To facilitate comparison of the different feature trajectories, they were normalized to have means of zero and maximum magnitudes of one. The phonetic transcription of the utterance is given along the top edge of each plot, and the vertical bars mark syllable onsets.

nuclei, but the lowpass MSG features evolve much more smoothly over time.

4.4 Summary

A simple signal-processing system that implemented some of the perceptually inspired signal-processing strategies laid out in Chapter 2 — critical-band-like frequency analysis, adaptation, an emphasis of slow changes in the spectrum of the input and a crude model of masking (the thresholding applied to the final display) — could be used to generate visual displays of speech that were relatively stable across a range of acoustic distortions. The same signal-processing system proved to be a better front end for ASR in highly reverberant conditions than any of the RASTA-PLP front ends, although the new system's performance under clean conditions was quite poor, committing nearly twice as many word errors as the PLP-based recognizer.

The MSG features were as good as PLP on the clean test and surpassed the performance of the original "visual" modulation-filtered spectrogram features on the reverberant test after some modifications to the MSG processing were made. These modifications were as follows:

- elimination of the thresholding operation,
- elimination of the per-channel adaptation,
- replacement of the original log compression by cube-root compression,
- and replacement of the original complex lowpass modulation filter by two real modulation filters, one lowpass and one bandpass.

Several signal-processing steps contributed to the better performance of the MSG features reverberation. Of these, the most crucial was the modulation filtering. Adaptation was also important, provided that it was implemented correctly: per-channel normalization was detrimental to recognizer accuracy in reverberant conditions, but normalization with respect to the global peak was beneficial.

The combination of different representations also appears to be useful for improving ASR performance. Using two MSG representations with different filters as the input to the MLP-based acoustic model gives better performance than using only one representation. Also, combining MSG features and log-RASTA-PLP by averaging phone log likelihoods from two MLPs gives better performance than using a single representation across a wide range of acoustic conditions, including a reverberant test and noisy tests at several different SNRs.

Chapter 5

Optimizing the Features for Automatic Speech Recognition

Despite the relative success of the initial experiments, described in Chapter 4, a number of challenges remained. The first concerned the generality of the MSG features. When MSG features were tested on the Numbers 95 subset with a more sophisticated ASR system than the one used in the Numbers 93 tests, the performance was worse than that obtained using log-RASTA-PLP. This was true under all conditions, including the focus reverberation condition. However, a system that used the MSG features in combination with log-RASTA-PLP did yield a statistically significant improvement in performance for many conditions and outperformed all other test systems. The second challenge concerned computing time. The computation of the MSG features required an off-line processing step, namely the normalization with respect to the global peak level. Use of this off-line processing meant that the MSG computation could not be completed until an entire utterance was received. Real-time applications typically require that feature computation proceed on-line in order to minimize system response latency.

This chapter describes an extensive series of recognition experiments whose primary goals were the improvement of MSG features and the development of an on-line algorithm for computing them. The optimization experiments started from the signal processing illustrated in Figure 4.9. Different aspects of the signal processing were systematically varied, with changes leading to improvements in recognition accuracy being retained for later experiments. The experiments are therefore presented in chronological order. To see the final outcome of these experiments, refer to Section 5.12 for a description of the best MSG features for the Numbers task and to Section 6.2.1 for a description of a slightly different set of MSG features that yielded the best performance for the large-vocabulary Broadcast News task.

Virtually all the experiments described in this chapter follow the same pattern. First, different versions of the front-end signal processing were created (via a systematic alteration of the processing or its parameters). Next, an ASR system was trained on features computed by each version of the front end on the clean Numbers 95 training set. Finally, the performance of each recognizer was measured on both the clean and moderately reverberant Numbers 95 development test sets (described in Section 4.3.1). Because the recognizers were trained only on features computed from clean data, performance on the clean test material indicated how well the features described phonetic information, while performance on the reverberant test material reflected how robust the features were to reverberation.

The structure of the recognizers was kept nearly constant for all experiments. Unless otherwise specified, each recognizer

- used an MLP acoustic model that processed nine frames of input at a time and contained roughly 92,000 weights. In some experiments the total number of input features was varied and the number of hidden units changed to keep the total number of weights approximately constant.
- used the multiple-pronunciation lexicon described in Section 3.1.3.
- used the backoff bigram grammar described in Section 3.1.4.
- was trained using an embedded training procedure in which an initial recognizer was trained on a labeling of the training data produced by a log-RASTA-PLP-based Numbers 95 recognizer. The initial recognizer was then used to relabel the training data via a forced alignment procedure, and a final recognizer was trained on this relabeling of the data. The final recognizer was then tested on the clean and reverberant Numbers 95 development tests.

5.1 Modulation Filter Optimization I

The goal of the experiments described in this section was to design a set of envelope filters that provided good performance in both clean and reverberant conditions. The design of the envelope filters was briefly considered in the experiments described in Chapter 4. However, in those earlier experiments the filter designs were constrained to match FFT analysis of Hamming-windowed or Kaiser-windowed segments of the envelope signal. In the experiments described in this section, the filter design was much less constrained.

IIR envelope filters were systematically examined in this set of experiments. IIR filters can concisely realize many different frequency responses, but at the cost of having non-uniform group delay characteristics. Nonuniformities in the envelope filter's group delay could have detrimental effects on recognizer performance because the timing of envelope modulations carries linguistic information [GAS98]. This potential problem was addressed by designing the IIR filters to have relatively little group delay variance in their passbands.

The IIR filters used in this set of experiments were designed in MATLAB by Deczky's method for IIR filter design (described in [OS89]) using the CONSTR routine for constrained, nonlinear minimization.[1] Initially, a single optimization was used to satisfy all the design requirements at once, but this approach proved to be impractical because the cost function defined by Deczky's method has many local minima which tend to "trap" the optimization routine. Instead, it proved to be more effective to break the filter design into a series of discrete stages. Thus, the development of a bandpass filter began with the design of a lowpass filter with the required upper cutoff frequency and upper stopband attenuation. It continued with the design of a highpass filter with the required lower cutoff frequency and lower stopband attenuation, and concluded with the design of an allpass filter that equalized the filter group delay in the passband. Once all three elements of the filter were designed, they were concatenated into a single filter, tested to ensure that the design requirements were satisfied, converted into a cascade of second-order sections and written out in a form that could be read by the MSG software.

Two series of envelope-filtering experiments were run. The first tested lowpass envelope filters with a variable cutoff frequency. The second tested bandpass envelope filters

[1] These experiments used the CONSTR routine distributed in the MATLAB version 5 Optimization Toolbox.

Cutoff frequency (Hz)	MSG alone		Combined with log-RASTA-PLP	
	clean	reverb.	clean	reverb.
24	11.0%	33.3%	6.0%	21.6%
20	9.6%	28.5%	5.5%	19.6%
16	10.1%	27.3%	5.6%	19.4%
12	11.4%	26.6%	5.4%	20.1%
8	13.6%	28.1%	6.5%	20.0%

Table 5.1: Word error rates for lowpass MSG features on their own and in combination with log-RASTA-PLP on the clean and reverberant Numbers 95 development test set as a function of envelope-filter cutoff frequency.

with a variable lower cutoff frequency and a fixed upper cutoff frequency using the results of the lowpass filter experiments. Both the lowpass and bandpass filters were designed to have 40 dB of attenuation in the upper stopband, an upper transition bandwidth of, at most, 3 Hz, and no more than ±1 sample of group delay ripple in the passband. The bandpass filters were designed to have a magnitude response proportional to modulation frequency below their lower cutoff frequency. Thus, all bandpass filters had a zero at 0 Hz modulation frequency. Aside from the new envelope filters and the use of only a single envelope filter, the MSG processing was identical to that described in Section 4.3. The MSG features based on the different envelope filters were tested alone and in combination with log-RASTA-PLP.

The results of the lowpass experiments are summarized in Table 5.1. For the tests of the lowpass MSG features on their own, the best performance on the clean test was obtained with a cutoff frequency of 20 Hz, and the best performance on the reverberant test was obtained with a cutoff of 12 Hz. For both tests, the performance with a cutoff of 16 Hz was lower by a statistically insignificant amount. For the tests of the lowpass MSG features combined with log-RASTA-PLP, the best performance on the clean test was obtained for a cutoff of 12 Hz, and the best performance on the reverberant test was obtained with a cutoff of 16 Hz. On the clean test, the difference in performance between the 16 Hz and 12 Hz cutoffs was not statistically significant. Based on these results, an upper cutoff frequency of 16 Hz was chosen for the filters used in the subsequent bandpass experiments.

The results of the bandpass experiments are summarized in Table 5.2. When the bandpass MSG features are used on their own, the best performance on the clean test

Lower cutoff	MSG alone		Combined with log-RASTA-PLP	
frequency (Hz)	clean	reverb.	clean	reverb.
0.5	18.7%	29.9%	7.1%	19.3%
1	16.9%	26.3%	6.4%	17.8%
2	14.3%	23.5%	5.9%	16.8%
4	12.7%	21.9%	5.6%	17.7%
8	11.4%	23.6%	5.3%	18.0%
16	12.5%	23.8%	5.3%	17.7%

Table 5.2: Word error rates for bandpass MSG features on their own and in combination with log-RASTA-PLP for the clean and reverberant Numbers 95 development test set as a function of the lower cutoff frequency of the envelope filter. The upper cutoff frequency of the envelope filter was fixed at 16 Hz, based on the results of the experiments summarized in Table 5.1. Because the bandpass filters were constrained to have a magnitude response proportional to modulation frequency below their lower cutoff frequency, the bandpass filter with a 16-Hz lower cutoff frequency is a differentiator for modulation frequencies of 0–16 Hz and suppresses modulations above 16 Hz.

was obtained with a lower cutoff frequency of 8 Hz, while the best performance on the reverberant test was obtained with a lower cutoff of 4 Hz. For the tests of the bandpass MSG features combined with log-RASTA-PLP, the best performance on the clean test was obtained for lower cutoffs of 8 Hz and 16 Hz (with only insignificant decrements in performance for cutoffs of 4 Hz and 2 Hz), and the best performance on the reverberant test was obtained with a lower cutoff of 2 Hz (with only insignificant decrements in performance for cutoffs of 1 Hz, 4 Hz, 8 Hz, and 16 Hz).

These experiments did not identify a single best envelope filter for both the clean and reverberant conditions. The lowpass filters generally provide the best performance on clean tests, while the bandpass filters generally yield the best performance on reverberant tests. Because earlier tests had demonstrated that combinations of representations could give good performance, two filters were chosen for use in subsequent tests: the lowpass filter with a 16-Hz cutoff frequency (which performed consistently well on all tests) and the bandpass filter with cutoffs at 2 Hz and 16 Hz, which gave the best performance on the reverberant test in combination with log-RASTA-PLP. The impulse responses of these two filters are shown in Figure 5.1, their frequency responses are shown in Figure 5.2, and their group delay characteristics are shown in Figure 5.3. These choices are in broad agreement with the perceptual results summarized in Section 2.2.1 as well as with a set of ASR experiments that demonstrated that modulation frequencies between 2 Hz and 16 Hz are the most reliable basis for speech recognition in a noisy environment [KAHP97].

5.2 Development of an On-line Automatic Gain Control

The next problem considered was the replacement of the off-line normalization step in the computation of MSG features with an on-line automatic gain control (AGC). Inclusion of frequency-local AGC in the MSG processing was anticipated to improve recognizer robustness by reducing the effects of unknown spectral shaping of the input signal and changes in overall signal level on the speech representation. An on-line AGC was preferred because it is more compatible with real-time recognition systems, it allows the ASR system to adapt to changes in the acoustic environment that occur in the course of an individual utterance, and it enhances the representation of the dynamics in the speech signal. In the modulation-spectral domain, the enhancement of dynamics performed by on-line adaptation corresponds to a suppression of slowly-varying components of the signal (that is, components with very low modulation frequencies).

The AGC needed to work with both positive and negative inputs, as well as with the long sequences of zero input that frequently occur in telephone applications. A simple feedback AGC design used in a computational model of forward masking [KPA92] appeared promising because it does not require a significant amount of computation and because it has adaptive properties similar to those observed in the auditory system—it adapts rapidly to signal onsets, recovers more slowly following signal offsets, and its rate of adaptation is higher for large input steps than for small input steps. The AGC is illustrated in Figure 5.4. In its original form, it operated in continuous time, and only functioned for non-negative input signals. As shown in Figure 5.5, it was possible to modify the design to function in discrete time and to operate with both positive and negative inputs.

At first glance, it would appear that this AGC attempts to compute 0/0 when given a long sequence of zeros as input. However, this does not occur because of the feedback loop in the design. If the input to the AGC is $x(t)$, the output of the AGC is $y(t)$, and the

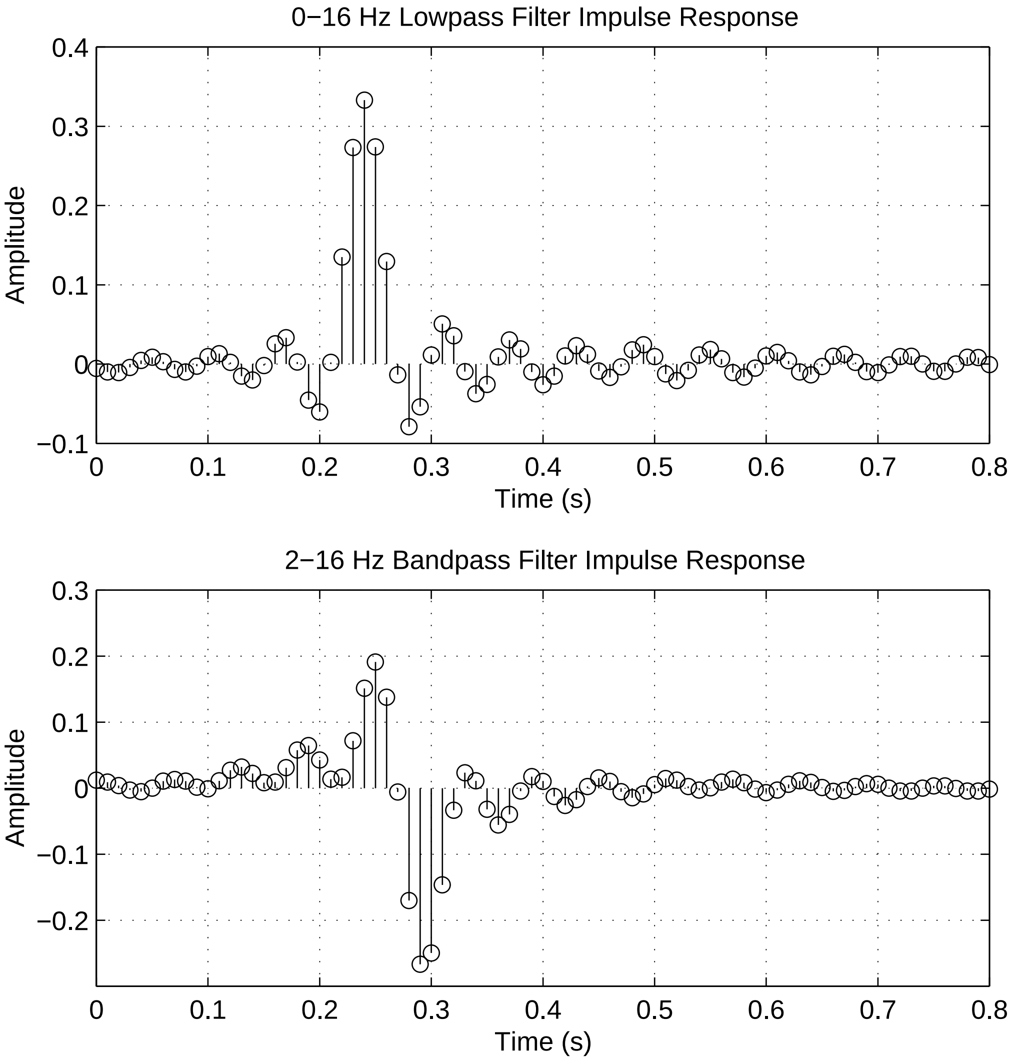


Figure 5.1: Impulse responses for the lowpass and bandpass IIR filters chosen for subsequent experiments with the MSG features.

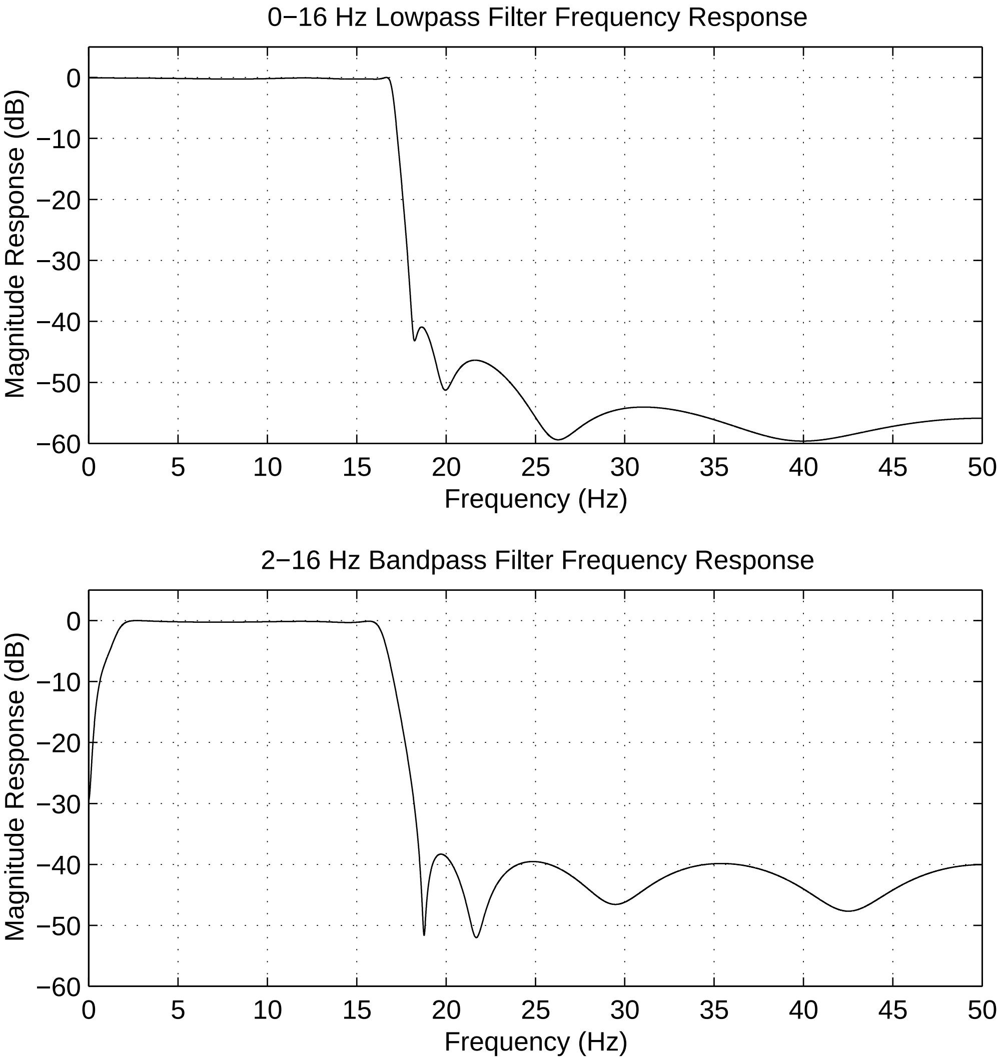


Figure 5.2: Frequency responses for the lowpass and bandpass IIR filters chosen for subsequent experiments with the MSG features.

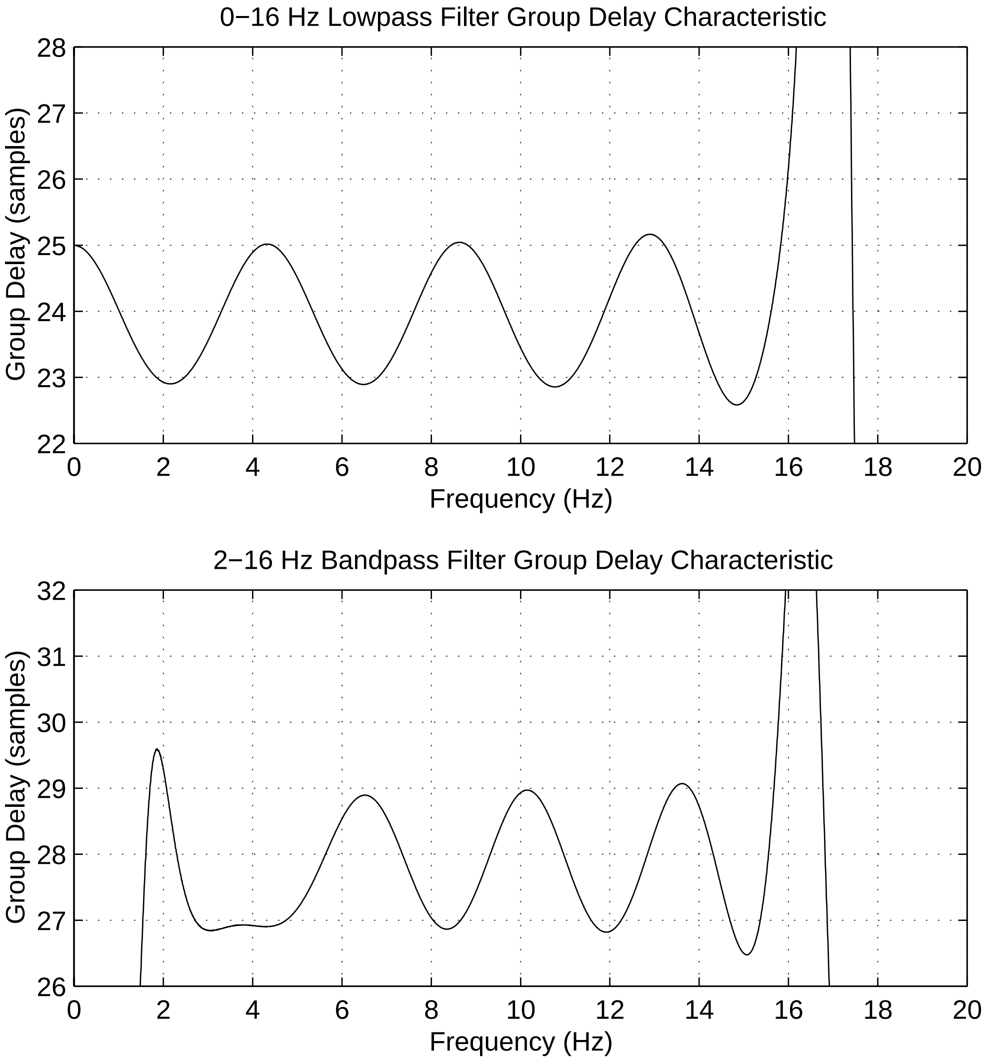


Figure 5.3: Passband group delay characteristics for the lowpass and bandpass IIR filters chosen for subsequent experiments with the MSG features.

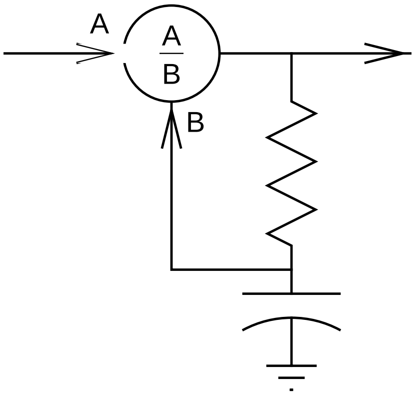


Figure 5.4: The original, continuous-time design for the feedback AGC proposed by Kohlrausch *et al.* [KPA92].

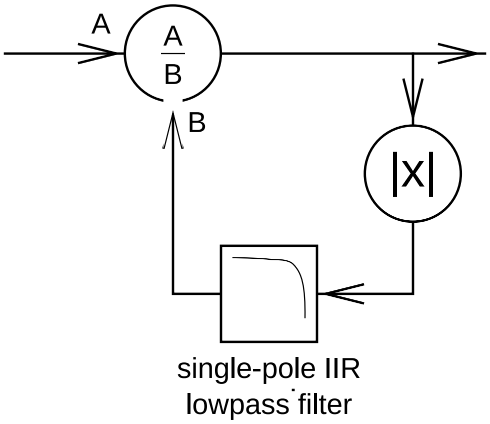


Figure 5.5: A discrete-time version of the feedback AGC unit that processes positive and negative input signals. The lowpass *RC* circuit in the continuous-time design is replaced with a single-pole lowpass filter to give a discrete-time design, and the absolute value of the divider output is fed back to permit the processing of both positive and negative input signals.

output of the lowpass filter is $g(t)$, and the transfer function for the lowpass filter is

$$H(z) = \frac{1-a}{1-az^{-1}}$$

(so that the filter has a DC gain of 1), then the AGC obeys the following two equations:

$$x(t) = y(t)g(t) \tag{5.1}$$

$$g(t) = (1-a)|y(t)| + ag(t-1) \tag{5.2}$$

Substituting the right-hand side of Equation 5.2 for $g(t)$ in Equation 5.1 gives the following expression for the output of the AGC:

$$y(t) = \begin{cases} \frac{-ag(t-1)+\sqrt{a^2g^2(t-1)+4(1-a)x(t)}}{2(1-a)} & \text{if } x(t) \geq 0 \\ \\ \frac{ag(t-1)-\sqrt{a^2g^2(t-1)-4(1-a)x(t)}}{2(1-a)} & \text{otherwise} \end{cases} \tag{5.3}$$

For a steady-state input, the magnitude of the AGC output is the square root of the magnitude of the input, and the sign of the AGC output is the sign of the input. When the input varies, the AGC output is nearly proportional to the square root of the magnitude of the input, but the constant of proportionality is greater than one. Thus, this AGC is essentially a square-root compressor with a variable gain that depends on the dynamics of the input. To prevent large transients on start-up, the AGC gain, $g(t)$, is initialized such that $g(0) = \sqrt{|x(0)|}$.

5.2.1 Experiments with a Single Feedback AGC Unit

The first experiments with this feedback AGC were designed to compare its performance to the off-line normalization. Two versions of the MSG processing were tested. In one the feedback AGC replaced the cube-root compression and off-line normalization, while in the other only the cube-root compression was replaced by a feedback AGC unit. The feedback AGC processing has two effects on the final MSG representation which might lead to better recognizer performance: it enhances the representation of onsets in the signal and, to some extent, it normalizes for unknown gains applied to the input signal. In contrast, the off-line processing only normalizes for unknown gains applied to the input signal, but it does so more effectively than the feedback AGC. Thus, the MSG features computed with only the feedback AGC could result in worse recognizer performance on the

		on-line AGC only		on-line AGC followed by off-line norm.	
Filter	AGC τ (ms)	clean	reverb.	clean	reverb.
lowpass 0–16 Hz passband	40	9.7%	42.9%	10.4%	27.4%
	80	9.7%	38.3%	10.3%	24.9%
	160	9.7%	35.9%	11.0%	24.6%
	320	10.1%	33.9%	12.3%	27.5%
	640	10.9%	33.7%	14.0%	30.3%
bandpass 2–16 Hz passband	40	13.2%	33.7%	13.0%	22.5%
	80	12.9%	32.7%	13.2%	22.7%
	160	12.8%	30.7%	13.3%	21.6%
	320	12.7%	29.5%	13.7%	21.9%
	640	11.9%	28.4%	14.4%	22.5%

Table 5.3: Word error rates for lowpass and bandpass MSG features as a function of AGC time constant, τ, for front ends using the feedback AGC as the sole gain control and for front ends that perform off-line normalization after the on-line, feedback AGC. Recall from Tables 5.1 and 5.2 that the lowpass MSG features with only the off-line normalization gave an error rate of 10.1% on the clean test and an error rate of 27.3% on the reverberant test, while the bandpass MSG features with only the off-line normalization gave an error rate of 14.6% on the clean test and an error rate of 23.3% on the reverberant test.

reverberant test even if the onset enhancement is beneficial because of the feedback AGC's lower effectiveness for gain normalization. The tests with MSG features that include both forms of AGC constitute a control — these features have both the onset enhancement of the feedback AGC and the highly effective gain normalization of the off-line normalization.

Recognition tests were performed with MSG features computed using either the lowpass envelope filter with a 16-Hz cutoff or the bandpass envelope filter with cutoffs of 2 Hz and 16 Hz. The time constant, τ, of the lowpass filter in the feedback AGC was variable, with time constants of 40 ms, 80 ms, 160 ms, 320 ms, and 640 ms tested. Aside from the changes to the AGC, the MSG computation was identical to that performed in Section 5.1. The results of these experiments are summarized in Table 5.3.

The effect of the feedback AGC time constant on recognizer accuracy was highly dependent on the test condition. For the tests with only the feedback AGC, performance decreased as the AGC time constant increased for lowpass MSG features on the clean test, but increased for the other three combinations of acoustic condition and envelope filter. For the tests incorporating both the feedback AGC and the off-line normalization, performance in both clean conditions decreased as the AGC time constant increased, while performance in reverberant conditions stayed roughly constant for the bandpass MSG features and attained a maximum at $\tau = 80$ ms or $\tau = 160$ ms for the lowpass MSG features. The variability of these results made it impossible to find a good choice for the AGC time constant.

The effect of including the off-line normalization in the MSG processing was also dependent on the test condition. For tests on the clean condition, inclusion of the off-line normalization was detrimental, while for tests on the reverberant condition inclusion of the off-line normalization greatly improved recognizer accuracy. These results suggest that a single feedback AGC unit does not provide sufficient gain normalization and that the off-line normalization is not the best AGC strategy.

The results of these experiments also indicate that the onset enhancement performed by the feedback AGC is beneficial to recognizer accuracy. Recall that lowpass MSG features with only the off-line normalization resulted in an error rate of 10.1% on the clean test and an error rate of 27.3% on the reverberant test, while the bandpass MSG features with only the off-line normalization yielded an error rate of 14.6% on the clean test and an error rate of 23.3% on the reverberant test. In most cases, the best performance obtained for a given condition (lowpass vs. bandpass MSG features and clean vs. reverberant acoustic conditions) with the on-line AGC is significantly better than the performance with only the off-line normalization.

5.2.2 Experiments with Two or Three Feedback AGC Units

Based on the results described above, MSG processing using two or three feedback AGCs in series was tested. It was expected that the additional AGC units would provide better gain normalization while also performing additional onset enhancement. The time constants of the AGC units were constrained so that units later in the chain had time constants greater than the preceding AGC units. This constraint matches the auditory system, for which it has been observed that more central regions have longer adaptation time constants than their more peripheral counterparts. None of the recognizers tested in this set of experiments included the off-line normalization, and aside from the changes in the AGC processing, the recognizers were identical to those in Section 5.2.1.

	second	first AGC τ (ms)					
filter	AGC τ (ms)	40	80	160	40	80	160
lowpass	160	9.4%	9.3%	—	29.9%	26.6%	—
0–16 Hz	320	9.1%	9.0%	9.8%	28.7%	25.6%	23.6%
passband	640	8.8%	9.2%	9.8%	27.5%	25.4%	23.9%
bandpass	160	13.1%	13.4%	—	27.0%	25.9%	—
2–16 Hz	320	12.7%	12.9%	12.6%	25.9%	23.2%	22.9%
passband	640	12.4%	12.0%	12.2%	24.3%	22.9%	21.9%
		clean tests			reverberant tests		

Table 5.4: Word error rates for lowpass and bandpass MSG features as a function of the time constants of the first and second feedback AGCs.

	second	third	first AGC τ (ms)			
filter	AGC τ (ms)	AGC τ (ms)	40	80	40	80
lowpass	160	320	10.2%	10.6%	24.6%	23.5%
0–16 Hz		640	10.0%	10.2%	23.5%	22.3%
passband	320	640	9.7%	10.3%	22.6%	22.0%
bandpass	160	320	15.7%	14.2%	24.6%	22.5%
2–16 Hz		640	13.8%	14.0%	22.9%	22.4%
passband	320	640	14.2%	12.9%	22.5%	20.6%
			clean tests		reverberant tests	

Table 5.5: Word error rates for lowpass and bandpass MSG features as a function of the time constants of the first, second, and third feedback AGCs.

The results of the experiments with two feedback AGC units are summarized in Table 5.4. For the lowpass MSG features there is no significant variation in performance for the different AGC time constants on the clean test. On the reverberant test the performance in the two cases where $\tau = 160$ ms for the first AGC is significantly better than any of the other cases. For the bandpass MSG features the performance on the clean test generally improves as the time constants of the two AGCs increase, but only a small number of the differences in performance are significant. On the reverberant test, the best performance is obtained when $\tau = 160$ ms for the first AGC and $\tau = 640$ ms for the second AGC. This result is significantly better than all other tests except for the two instances where a word error rate of 22.9% was obtained.

The results of the experiments with three feedback AGC units are summarized in

Table 5.5. The only significant improvement over the two-AGC MSG features is obtained on the reverberant test with lowpass MSG features and three AGCs having time constants of 80 ms, 320 ms, and 640 ms. This outcome indicates that two AGC units in series are sufficient.

It is difficult to relate these results directly to measurements of adaptation in the human auditory system obtained via measurements of forward masking. However, it is interesting to note that the best time constant for the first AGC is 160 ms for both the lowpass and bandpass MSG features. This is nearly identical to the time constant of the filter that, sandwiched between two nonlinearities, performs automatic gain control in RASTA processing. A correspondence between some measurements of forward masking and RASTA processing has been noted [PH94]. More generally, these results indicate that normalization of the short-time spectrum of speech with respect to an average spectrum measured over a duration corresponding to several syllables can lead to a more stable representation of speech, at least as characterized by these ASR experiments.

5.2.3 Cross-coupling the AGCs

The coupling of AGCs across frequency channels has been proposed in a number of computational auditory models (e.g., [Lyo82]). In a cross-coupled AGC system, the signal in a given channel is normalized by a gain that is estimated not only from the signal itself, but also from signals in neighboring channels. Because it is a form of lateral inhibition, an important effect of the coupling is to emphasize spectral peaks and, more generally, to preserve spectral shape information that could be eliminated by fast-acting, per-channel AGC processing. Figure 5.6 illustrates the cross-coupling of the feedback AGC units for the case where only adjacent channels are coupled.

The addition of cross-coupling complicates the AGC computation somewhat. It was necessary to place unit delays in the cross-channel coupling paths to ensure that the AGC output at a given time has a closed-form solution. Without the unit delays, the AGC computation would require some sort of general, numeric root-finding procedure. With the unit delays, the AGC computation is described by the following equations:

$$u_i(t) = ag_i(t-1) + (1-a) \sum_{d=-c_l}^{c_h} w_{i+d,i} |y_{i+d}(t-1)|$$

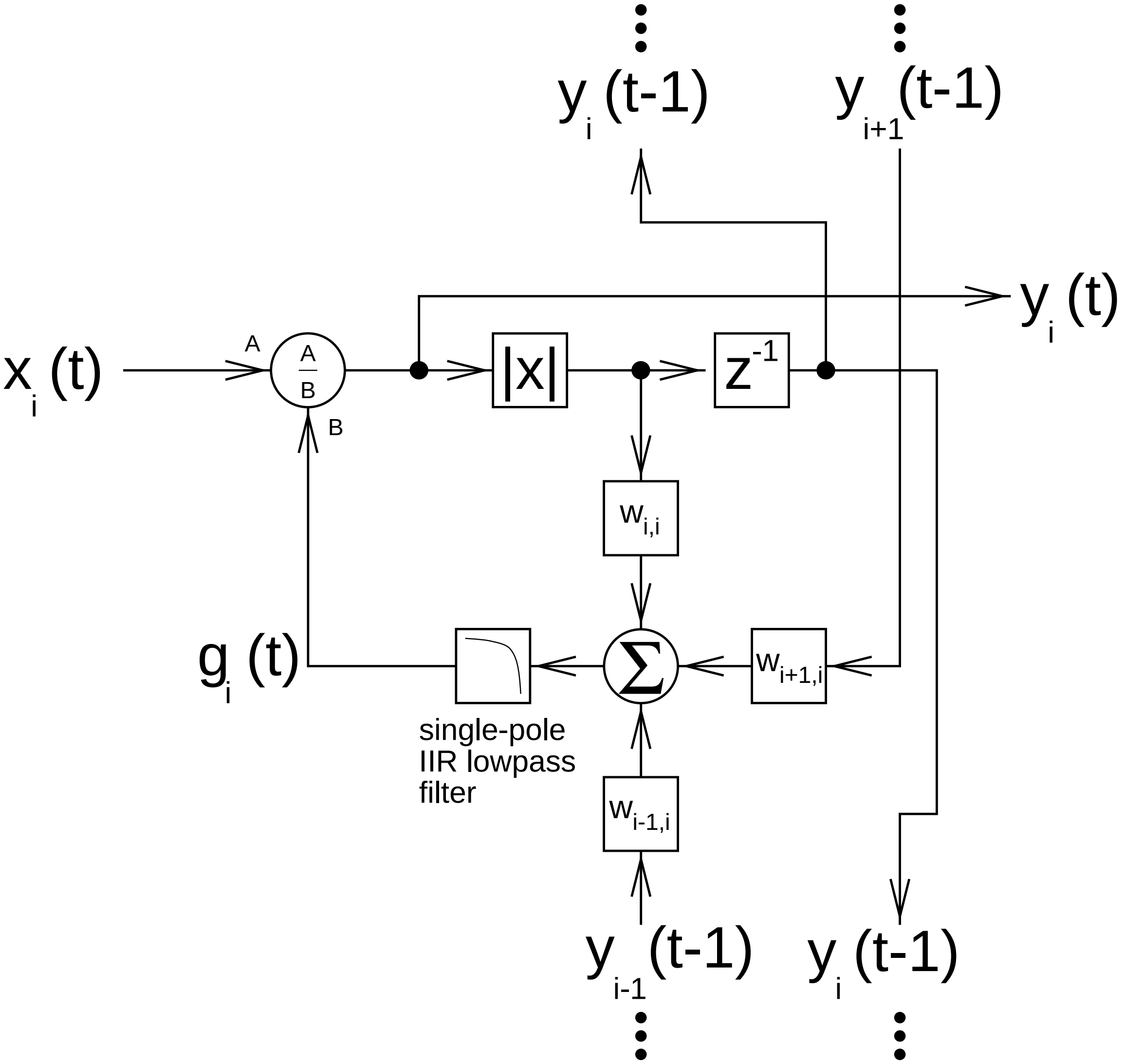


Figure 5.6: Signal processing for cross-coupled feedback AGC unit. In each channel the signal, $x_i(t)$, is normalized by a factor, $g_i(t)$, that is a temporally smoothed, weighted average of the signal level in the channel itself ($y_i(t)$) and in other channels ($y_j(t)$, where $j \neq i.$). This processing is a form of lateral inhibition (as well as automatic gain control) that serves to enhance energy peaks in time and frequency. The unit delays (the boxes labeled z^{-1}) are included to simplify the coupled AGC computation. The $w_{i,j}$ factors are the coupling weights between channels. In this figure, only coupling between neighboring channels is portrayed.

$$v_i(t) = \sqrt{u^2 + 4(1-a)w_{i,i}|x_i(t)|}$$

$$g_i(t) = \frac{u+v}{2}$$

$$y_i(t) = \begin{cases} \frac{v-u}{2(1-a)w_{i,i}} & \text{if } x_i(t) \geq 0 \\ -\frac{v-u}{2(1-a)w_{i,i}} & \text{otherwise} \end{cases}$$

where $x_i(t)$ is the input from the i-th channel, $y_i(t)$ is the output for the i-th channel, $g_i(t)$ is the normalization factor for the i-th channel, $w_{i,j}$ is the weight applied to y_j in the computation of g_i, the AGC for the i-th channel is coupled to channels $i - c_l$ through $i + c_h$, and the lowpass filters in the AGC units all have the same transfer function

$$H(z) = \frac{1-a}{1-az^{-1}}$$

The boundary conditions are handled by padding the input vectors, $x(t)$, with copies of the highest and lowest channels.

As with the uncoupled feedback AGCs, it was desirable to initialize the outputs of the cross-coupled AGCs to the steady-state response to the first input for an utterance in order to prevent start-up transients. Performing this initialization for the cross-coupled AGCs was more complicated than for the uncoupled AGCs, however, because there was no closed-form expression for the steady-state response to an input vector. At first, the cross-coupled AGCs were initialized by clamping their inputs to the initial values for an utterance and then running the AGC computation until the outputs converged to the steady-state response. While this approach was simple, it was also very slow. For many of the Numbers 95 utterances the AGC initialization required more time than the actual processing of the utterance, so the initialization of the cross-coupled AGCs was sped up using a Newton-Raphson root-finding procedure.

In the first experiment using the cross-coupled AGC processing, a single AGC block with variable time constant was used, and no other normalization was performed on the features. The coupling weights between channels were set arbitrarily to $w_{i,i} = 0.5$ and $w_{i-1,i} = w_{i+1,i} = 0.25$. Only lowpass MSG features were tested.

The results of these experiments are summarized in Table 5.6. The only significant difference between the results with the cross-coupled AGC and the results with the uncoupled AGC (summarized in Table 5.3) was that the performance of the coupled AGC with $\tau = 320$ ms on the reverberant test was significantly worse than the comparable uncoupled

AGC τ (ms)	clean	reverb.
40	10.2%	43.6%
80	9.8%	39.3%
160	9.7%	35.4%
320	10.2%	35.5%

Table 5.6: Word error rates for lowpass MSG features computed with a single cross-coupled, feedback AGC unit on the clean and reverberant Numbers 95 development test sets, as a function of AGC time constant. Experiments with $\tau = 640$ ms failed due to an overflow in the fixed-point MLP training procedure.

AGC test. A test with $\tau = 640$ ms failed due to an overflow in the fixed-point MLP training procedure. No attempt was made to work around the overflow problem because it was assumed that the performance would be inferior to that obtained with uncoupled AGCs, given the result with coupled AGCs having $\tau = 320$ ms. These results may indicate that any lateral-inhibitory processing of speech spectra should be relatively fast-acting, working over durations of time shorter than a syllable and are consistent with the observation that spectral peaks in the speech signal generally change their positions significantly over the course of a syllable.

Next, two cross-coupled AGCs were used in series to normalize the features. It was expected that this configuration would produce better performance, as had occurred in the experiments using uncoupled feedback AGCs, described in Section 5.2. The experiments described in this section are identical to those summarized in Table 5.4, with the exception of the cross-coupling in the AGCs. In these tests the first AGC had coupling weights of $w_{i,i} = 0.5$ and $w_{i-1,i} = w_{i+1,i} = 0.25$, while the second AGC had coupling weights of $w_{i,i} = 3/9$, $w_{i-1,i} = w_{i+1,i} = 2/9$, and $w_{i-2,i} = w_{i+2,i} = 1/9$. As in the earlier experiment, the selection of coupling weights was arbitrary. As summarized in Table 5.7, there was no significant improvement in performance over the best results using two uncoupled feedback AGCs.

Because no consistent benefit could be obtained by using more than two feedback AGC units in series or by using cross-coupled feedback AGC units, subsequent experiments computed MSG features using two uncoupled feedback AGC units in series. For the lowpass MSG features (calculated using the IIR lowpass filter with a 16 Hz cutoff) the first AGC had a time constant of 160 ms and the second had a time constant of 320 ms. For the

	second	first AGC τ (ms)					
filter	AGC τ (ms)	40	80	160	40	80	160
lowpass	160	9.2%	9.2%	—	30.1%	27.4%	—
0–16 Hz	320	9.2%	8.9%	8.7%	29.2%	26.7%	25.1%
passband	640	8.9%	9.1%	8.9%	29.6%	26.9%	26.0%
bandpass	160	12.6%	12.5%	—	25.5%	24.0%	—
2–16 Hz	320	12.1%	12.0%	12.3%	24.5%	23.2%	22.1%
passband	640	11.5%	11.5%	11.5%	23.7%	23.0%	21.7%
		clean tests			reverberant tests		

Table 5.7: Word error rates for lowpass and bandpass MSG features computed with two cross-coupled, feedback AGC units on the clean and reverberant Numbers 95 development test sets, as a function of the AGC time constants.

bandpass MSG features (calculated using the IIR bandpass filter with a 2–16 Hz passband) the first AGC had a time constant of 160 ms and the second had a time constant of 640 ms.

5.3 Modifying the Resolution of the Initial Frequency Analysis

The MSG features have about twice the spectral resolution of the most common speech representations used for ASR. For telephone-bandwidth speech, the MSG processing produces thirty features per frame (fifteen lowpass features and fifteen bandpass features), while PLP and RASTA-PLP processing typically produce eighteen features per frame (nine features and nine delta features). The spectral resolution of an ASR front end should not be any higher than necessary to produce a good description of the speech signal because the higher resolution can increase the complexity of the acoustic model and can make the recognizer more sensitive to speaker-dependent signal characteristics.

The experiments described in this section tested the effects of reducing the spectral resolution of the MSG features by the simplest means possible: reducing the resolution of the initial spectral analysis in the MSG processing. In these experiments, the bandwidth and spacing of the filters in the constant-Q FIR filterbank were varied independently. For both the filter bandwidth and the filter spacing, values of 1, 1/2, 1/3, and 1/4 octave were tested. The spacing between adjacent filters was constrained to be less than or equal to the

bandwidth (octaves)	spacing (octaves)	lowpass 0–16 Hz		bandpass 2–16 Hz	
		clean	reverb.	clean	reverb.
1	1	16.5%	41.3%	21.4%	38.0%
	1/2	12.2%	36.4%	16.6%	32.2%
	1/3	11.3%	33.4%	14.8%	29.8%
	1/4	11.4%	34.7%	14.1%	28.8%
1/2	1/2	11.2%	31.3%	13.8%	26.6%
	1/3	8.9%	27.8%	12.2%	24.0%
	1/4	9.0%	28.1%	12.2%	24.5%
1/3	1/3	9.4%	28.3%	12.2%	23.4%
	1/4	9.6%	26.5%	11.5%	22.8%
1/4	1/4	9.9%	24.1%	12.0%	22.1%
1/8	1/8	8.9%	22.2%	—	—

Table 5.8: Word error rates for lowpass and bandpass MSG features on the clean and reverberant Numbers 95 development test sets, as a function of the bandwidth and spacing of the filters in the initial constant-Q FIR filterbank.

filter bandwidth so that no spectral gaps would occur in the representation. Both lowpass and bandpass MSG features were tested, and except for the changes to frequency analysis, the recognizers in these experiments were identical to those used in the previous section.

The results of these experiments are summarized in Table 5.8. On the clean tests, performance for both envelope filters reached an optimal plateau for bandwidths of 1/2 octave or less and filter spacings of 1/3 octave or less. This outcome is consistent with the view that a detailed representation of the speech spectrum is not required, nor even particularly desirable, for recognition of speech.

On the reverberant tests, performance for both envelope filters improved as the filterbank bandwidth and spacing decreased, with the best performance obtained with 1/4-octave bandwidths and 1/4-octave spacings. To see if this pattern continued, a test was also run for lowpass MSG features computed with 1/8-octave bandwidths and 1/8-octave spacings. Recognizer performance improved significantly under reverberant conditions with the 1/8-octave filterbank. This result is consistent with the idea that increasing the temporal window of the initial spectral analysis (thereby increasing its spectral resolution) makes compensation for reverberation via techniques such as AGC or cepstral mean normalization possible [Ave97b].

Based on these experiments, the quarter-octave filterbank was retained. While it appeared that using a spectral resolution higher than a quarter octave could give better performance under reverberant conditions, this approach was not adopted since it has already been explored [Ave97b]. The filterbank resolution was eventually changed when tests, described in Section 5.5, demonstrated that a Bark-scale filterbank having coarser resolution than the quarter-octave filterbank in the lower frequencies yielded equally good recognizer performance.

5.4 Variations on the AGC

5.4.1 An Experiment with Off-line Feature Normalization

Recall from Section 5.2.1 that the feedback AGC processing basically computes the square root of its input signal, multiplied by a variable gain. This processing reduces the effect of an unknown gain term but does not completely eliminate it. While the experiments in Section 5.2.2 demonstrated that two feedback AGC units in series could yield better performance than off-line gain normalization, it is conceivable that further improvements of the AGC could be made. In order to test this possibility, additional, off-line normalization of the features was performed, based on statistics computed from the test data.

As described in Section 3.1.2, the speech features are normalized to have approximately zero mean and unit variance by subtracting an estimate of the mean value of the features and dividing by an estimate of the standard deviation of the features before they are input to the MLP acoustic model. These mean and standard deviation estimates are computed from the recognizer's training data; however, better recognition performance on a given test may usually be obtained by computing these estimates over the test data, especially in cases where the acoustic conditions differ between the training and test sets. Computing the means and variances from the test data is only possible in off-line applications. In this study, algorithms operating on-line are preferred because they can be applied to both on-line and off-line tasks. The use of test-set statistics for feature normalization can (to a certain extent) indicate if performance on acoustically mismatched test data can be improved by making changes to the front-end signal processing. If their use does not improve the recognition performance on acoustically mismatched test data, then the front-

Training Set Normalization	Test Set Normalization
21.8%	18.6%

Table 5.9: Word error rates on the reverberant Numbers 95 development test set for a recognizer that uses both lowpass and bandpass MSG features and normalizes the features using either the training data (the usual case) or the reverberant test data.

end signal processing may already be performing adequate normalization of the means and variances of the speech features.

The recognizer used in this experiment differed from the ones used in the experiments described in Sections 5.1 to 5.3 in that the lowpass and bandpass MSG features were both presented to a single MLP for acoustic likelihood estimation. The MLP had 270 input units (30 features per frame × 9 frames of input), 400 hidden units, and 32 output units, for a total of 120,800 weights (about 30% more weights than the other recognizers).

The recognizer's performance was measured on the reverberant test with the features being normalized using estimates of the means and standard deviations from the training set (the usual case) or from the reverberant test set. The results are summarized in Table 5.9. Using the means and standard deviations from the test data improved the performance on the reverberant test by 15% (an absolute reduction in word error rate of 3.2%). This improvement was statistically significant, and suggested that better normalization of the features in the MSG signal processing could improve recognizer performance under reverberant conditions.

5.4.2 An Alternative AGC Design

In light of the test-set normalization results, a second on-line AGC was designed that performs a more complete normalization of signal variance than the original feedback design. This alternative operates according to the following two equations:

$$y(t) = \begin{cases} g(t)x(t) & \text{if } x(t) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$g(t+1) = g(t) - ay(t) + b(1 - g(t))$$

AGC τ (ms)	clean	reverb.
40	Failed	Failed
80	15.6%	52.1%
160	15.5%	45.7%
320	15.6%	41.8%
640	16.9%	40.0%

Table 5.10: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass MSG features, as a function of AGC time constant. The $\tau = 40$ ms test failed due to an overflow in the fixed-point MLP training procedure.

That is, the output of this AGC, $y(t)$, is the input, $x(t)$, halfwave rectified and multiplied by a variable gain term $g(t)$. At each time step the gain is reduced by a factor proportional to the AGC output and is increased by a factor proportional to one minus the gain. The value b controls the recovery rate of the AGC, and the ratio a/b controls the adaptation rate. This design is loosely based on a single-reservoir model for the inner-hair-cell/auditory-nerve-fiber complex (e.g., [OS75]). Unlike the other feedback AGC, this design is a "true" AGC in the sense that the output converges to a fixed value of b/a for a fixed input level x, provided that $x \gg b$.

The first experiments with this design used a single AGC unit, with the AGC time constant being an experimental parameter. The adaptation and recovery time constants were set to be approximately equal. Only lowpass MSG features were tested because it was expected that the halfwave rectification would eliminate important data in the bandpass features. The results of these experiments are summarized in Table 5.10. The experiment with an AGC time constant of 40 ms failed because of an overflow in the fixed-point MLP training routine. Performance in the other cases was markedly worse than with a single feedback AGC unit (compare to Table 5.3).

A second series of experiments was run in which a single AGC unit was used and the adaptation and recovery time constants were varied independently, with adaptation time constants of 40 ms or 60 ms and recovery time constants of 160, 240, or 320 ms. All of the tests with the 40-ms adaptation time constant failed due to overflow in the MLP training procedure, so only the 60-ms results are shown in Table 5.11. The performance in these experiments is worse than that obtained in the experiments with equal time constants, so no further work was done with this AGC design.

AGC recovery τ (ms)	clean	reverb.
160	17.0%	49.3%
240	17.0%	47.9%
320	18.0%	47.1%

Table 5.11: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass MSG features, as a function of AGC recovery time constant. The AGC adaptation time constant was fixed at 60 ms.

5.4.3 Normalizing the Features On-Line

The test-set normalization results in Section 5.4.1 suggested another approach to improving recognizer performance in reverberation: performing an on-line normalization of the feature means and variances [CCC+96, TH97]. This processing will reduce the effects of slowly varying additive offsets and multiplicative gain terms caused, for example, by additive noise and spectral shaping. The signal processing used to perform on-line normalization is illustrated in Figure 5.7. The lowpass IIR filter in the first stage computes an estimate of the signal mean (with an exponentially decaying window into the past) which is subtracted from the signal. The lowpass IIR filter in the second stage computes an estimate of the signal's standard deviation, with the signal being normalized by that estimate. A small offset, ε, is added to the estimate of the standard deviation to preclude division by zero. For the first utterance in a set of test data, the estimates of the mean and standard deviation are initialized from estimates computed over the training data. For each following utterance, the final estimates from the previous utterance are used.

The time constants of the lowpass filters in the normalization control the duration of the processing's memory. The longer the time constant, the more reliable the estimates of the mean and variance; however, longer time constants also entail slower adaptation to changes in the acoustic environment. To determine suitable time constants for the lowpass filters, recognition experiments were run. In the experiments, the time constants of the two lowpass filters were constrained to be equal, and the offset ε to the standard deviation estimate was set to 1. As usual, lowpass and bandpass MSG features were tested separately. The results of these experiments are summarized in Table 5.12.

The performance in reverberation was significantly better with on-line feature

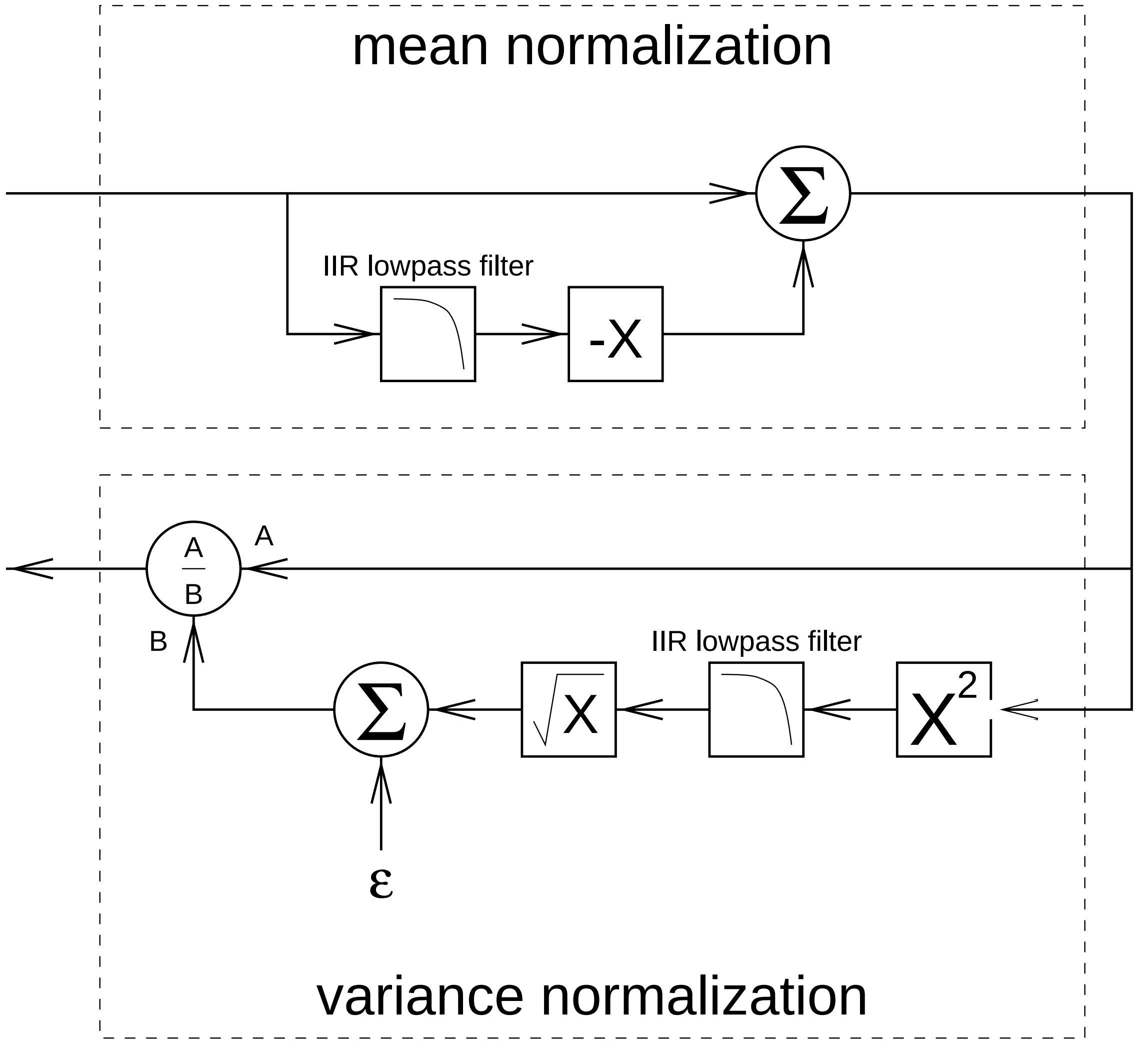


Figure 5.7: Implementation of the on-line feature normalization.

Normalization	lowpass 0–16 Hz		bandpass 2–16 Hz	
τ (ms)	clean	reverb.	clean	reverb.
250	12.2%	21.3%	14.6%	20.7%
500	11.8%	20.2%	14.1%	20.6%
1000	11.4%	19.5%	13.7%	20.1%
2000	10.9%	19.5%	13.4%	20.2%
3000	10.7%	20.2%	12.8%	19.5%
4000	10.4%	20.1%	13.7%	20.7%
none	9.8%	23.6%	12.2%	21.9%

Table 5.12: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass and bandpass MSG features with on-line normalization of the features as a function of the time constant of the on-line normalization. The comparable results without on-line normalization are listed in the row labeled "none," and are copied from Table 5.4.

normalization than without for all values of τ that were tested. On the clean test there was a significant degradation in performance with the lowpass features for $\tau < 3.0$ s and with the bandpass features for all values of τ except for $\tau = 3.0$ s. For subsequent tests, τ was set to 2.0 s because it was expected that longer time constants might not allow sufficiently rapid adaptation to changes in acoustic conditions and because the degradation in performance on the clean test was acceptably small. This setting of τ is identical to that arrived at in [TH97] using a different recognition task.

A second set of experiments was run to measure the effect of different settings of the offset ε. These experiments were run only for the lowpass MSG features, and $\tau = 2.0$ s for the on-line normalization. The results of these tests are summarized in Table 5.13. Varying ε had no significant impact on recognizer performance, so the original setting of $\varepsilon = 1$ was retained in subsequent experiments.

5.5 A Power-spectral Implementation of the Initial Frequency Analysis

Compared to the RASTA-PLP front end, MSG processing was very slow: for a given utterance it took ten times as long to compute MSG features as it did to compute PLP

ε	clean	reverb.
0.01	11.1%	19.6%
0.03	11.1%	19.3%
0.1	11.2%	19.7%
0.3	10.9%	19.2%
1.0	10.9%	19.5%

Table 5.13: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass MSG features with on-line normalization of the features as a function of the offset, ε, added to the estimate of standard deviation.

Filter Implementation	Filter Shape	Frequency Scale	lowpass 0–16 Hz		bandpass 2–16 Hz	
			clean	reverb.	clean	reverb.
direct FIR	trapezoidal	quarter-octave	10.9%	19.5%	13.4%	20.2%
power spectral	trapezoidal	quarter-octave	8.8%	19.4%	11.3%	20.1%
		Bark	9.0%	19.7%	10.9%	20.0%
	triangular	quarter-octave	9.0%	18.5%	11.0%	18.3%
		Bark	8.4%	18.5%	10.8%	18.5%

Table 5.14: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass and bandpass MSG features computed with different initial filterbanks.

or RASTA-PLP features. Profiling of the MSG code showed that the bulk of the processing time was spent in the core routines of the initial FIR filterbank. Because this code was already implemented efficiently (using an FFT-based block convolution), replacement of the FIR filterbank with a bank of filters operating on power spectra computed with the short-time Fourier transform was investigated as a means of speeding up the MSG processing. This filterbank operates in the power-spectral domain, so the square root of its output is computed to produce an amplitude spectrum at the output. Two different filter shapes, trapezoidal and triangular, and two different auditory-like frequency scales, quarter-octave and Bark, were investigated. The results of these experiments are summarized in Table 5.14.

On the clean test, the performance with features computed using the power spectral filterbank was significantly better than the performance with features computed using the FIR filterbank for all conditions. On the reverberant test, the performance with features computed using the power spectral filterbank, triangular filters, and bandpass modulation

Filter #	Bark Scale	Quarter-octave Scale
1	3680	3670
2	3100	3070
3	2620	2590
4	2200	2180
5	1850	1830
6	1550	1540
7	1300	1300
8	1070	1090
9	880	920
10	720	770
11	570	650
12	440	550
13	320	460
14	—	390
15	—	320

Table 5.15: Filter passband centers for the Bark-scale and quarter-octave filterbanks. The two filterbanks are essentially identical for frequencies above 1 kHz. Below 1 kHz the quarter-octave filterbank has finer frequency resolution.

filters was significantly better than the performance with comparable features computed using the FIR filterbank. There was no significant difference in performance between using features computed with the quarter-octave or Bark-scale filterbanks, indicating that the quarter-octave filterbank has finer resolution in the lower frequencies than is necessary for the recognition task. As can be seen from Figure 2.1, which shows filter bandwidth as a function of center frequency, and from Table 5.15, which gives the passband centers for the Bark and quarter-octave filterbanks used in this set of experiments, the Bark and quarter-octave scales are nearly identical for frequencies above 1 kHz, while for frequencies below 1 kHz the quarter-octave scale has finer frequency resolution. Because the Bark-scale filterbank produced two fewer features per frame, a Bark-scale power spectral filterbank with triangular filters was used for all subsequent experiments.[2]

5.6 Modulation Filter Optimization II

The optimizations made to the MSG feature processing in Sections 5.1 through 5.5 had an unexpected consequence: the improved lowpass MSG features were as accurate as the improved bandpass MSG features on the reverberant test (see Table 5.14). Using both the lowpass and bandpass MSG features still produced better performance, though. A system in which both feature sets were input to a single MLP acoustic model with 400 hidden units had a word error rate of 7.8% on the clean test and 17.4% on the reverberant test. The performance on the reverberant test is a great improvement over that reported in Section 4.3.3 for a comparable recognizer using an earlier version of the lowpass and bandpass MSG features. This earlier recognizer had a word error rate of 8.5% on the clean test and 27.3% on the reverberant test.

The earlier version of the MSG features used FIR envelope filters with passbands of 0–8 Hz for the lowpass features and 2–8 Hz for the bandpass features, while the newer version used IIR envelope filters with passbands of 0–16 Hz for the lowpass features and 2–16 Hz for the bandpass features. One of the motivations for using the broader envelope filters was that they were anticipated to yield better performance on clean speech; however,

[2] These results also show that more accurate auditory modeling does not necessarily produce better recognition performance. As demonstrated in Section 2.1, the Bark scale has unrealistically low resolution in the low frequencies. Nevertheless, in these experiments features computed with a Bark-scale filterbank provided performance similar to that obtained using features computed with a quarter-octave filterbank, even though the resolution of the quarter-octave filterbank is more realistic in the 300–1000 Hz range.

this expectation was not met. Thus, the design of appropriate envelope filters was re-examined, but now using the optimized choices for the other signal-processing steps in the MSG computation. In the experiments described in this section, FIR filters were tested because they are simpler to implement, may require less computation than IIR filters, and have uniform group delay characteristics (because linear-phase FIR filters are used) that better preserve envelope timing information.

5.6.1 FIR Lowpass Filters

Lowpass MSG features produced using FIR filters having different lengths and different cutoff frequencies were tested on the clean and reverberant Numbers 95 development test sets. The FIR filters had at least 40 dB of stopband rejection. The filter transition bandwidth varied with filter length, from a minimum of 3 Hz for the 49-point filters to a maximum of 10 Hz for the 13-point filters. The results of these experiments are summarized in Table 5.16.

The results of the experiments with the 49-point and 25-point filters indicated that the cutoff frequency could be as low as 12 Hz with no statistically significant, negative impact on recognizer performance. The results of the experiments with the shorter 19-point and 13-point filters demonstrated that there was no advantage to using longer filters with narrower transition bandwidths. Filters shorter than 13 points were not studied because it was not possible to design shorter filters that had the desired 40 dB of stopband rejection. It is possible that shorter filters with less stopband rejection would be equally effective, but this possibility was not explored.

5.6.2 FIR Lowpass Filters with DC Suppression

A consistent result from work on the data-driven design of envelope filters [Ave97b, HAvVT97, AH96, AvVH96, HWA95] is that the filters produced by the automatic procedures suppress DC somewhat, but are generally lowpass in form. This suggested that recognizer performance might be improved by adding some DC suppression to the filter used to produce lowpass MSG features. To test this possibility, a set of lowpass filters with a cutoff frequency of 12 Hz and variable DC suppression was generated by convolving a 14-point lowpass FIR filter with a 2-point highpass FIR filter of the form $H(z) = 1 - xz^{-1}$,

Num. Points	Cutoff Frequency (Hz)	clean	reverb.
49	24	8.0%	18.4%
	20	7.7%	18.6%
	16	8.4%	18.6%
	12	8.6%	19.1%
	8	10.7%	19.8%
	4	18.7%	27.6%
25	22	7.4%	18.7%
	20	7.3%	18.5%
	18	7.6%	18.5%
	16	8.8%	18.2%
	14	7.3%	17.8%
	12	8.0%	19.0%
	10	8.7%	19.0%
19	16	7.6%	18.2%
	14	8.0%	18.6%
	12	8.5%	19.2%
13	16	7.8%	18.0%
	14	8.1%	17.9%
	12	7.8%	18.0%

Table 5.16: Word error rates for the clean and reverberant Numbers 95 development tests using lowpass MSG features computed with FIR envelope filters as a function of filter length and cutoff frequency. The relatively poor performance on the clean test using the 25-point filter with a cutoff frequency of 16 Hz appears to be an outlier caused, perhaps, by a relatively poor initialization of the MLP weights during recognizer training.

DC Magnitude Response (dB)	clean	reverb.
-4	8.1%	17.9%
-5	8.1%	18.1%
-6	7.8%	16.9%
-7	8.1%	17.2%
-8	8.0%	16.6%
-9	8.0%	16.4%
-11	7.9%	16.2%
-12	7.8%	17.1%
-13	8.2%	16.7%
-17	8.6%	17.4%
$-\infty$	9.1%	18.9%

Table 5.17: Word error rates for the clean and reverberant Numbers 95 development tests for lowpass MSG features computed with filters having variable amounts of DC suppression.

where the value of x set the amount of DC suppression (as suggested in [NJ94]). For $x = 1$, the resulting filter has a zero at DC; for $x = 0$, the resulting filter has no DC suppression. These filters were then used to generate MSG features which were tested in the usual recognition experiments.

Table 5.17 summarizes the results of these experiments. Performance on the clean test was relatively insensitive to the amount of DC suppression, with no significant differences in performance between any of the conditions. Performance on the reverberant test was significantly better for lowpass filters with 8, 9, or 11 dB of DC suppression than for a lowpass filter with no DC suppression. This result is consistent with the studies on the data-driven design of envelope filters, and with the notion that it is the dynamically-changing portions of the speech signal that are most important for characterizing its linguistic content.

5.6.3 Lowpass and Bandpass FIR Envelope Filters

Next, the use of lowpass and bandpass MSG features together was examined, with the passbands of the lowpass and bandpass envelope filters being jointly optimized. All filters were symmetric FIR filters with 40 dB of stopband rejection and transition bandwidths of no more than 6 Hz. The filter lengths were chosen to be as short as possible while

Lowpass Cutoff (Hz)	Bandpass Lower Cutoff (Hz)	clean	reverb.
9	9	8.9%	18.9%
	8	7.9%	17.0%
	7	8.3%	17.7%
	6	7.8%	16.8%
	5	7.4%	17.3%
	4	8.0%	17.5%
	3	7.8%	17.4%
8	8	8.2%	17.5%
	7	8.6%	18.0%
	6	7.4%	16.5%
	5	7.8%	17.2%
	4	7.9%	17.6%
	3	7.4%	17.1%
7	7	8.5%	18.2%
	6	7.9%	16.5%
	5	7.7%	17.1%
	4	8.2%	17.2%
	3	7.8%	17.4%
6	6	7.9%	16.0%
	5	7.5%	16.6%
	4	7.9%	17.8%
	3	8.3%	17.1%
5	5	7.9%	16.3%
	4	8.4%	16.6%
	3	8.2%	17.6%
4	4	8.7%	17.2%
	3	9.0%	17.3%
3	3	9.1%	17.0%

Table 5.18: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using both lowpass and bandpass MSG features as a function of the cutoff frequency of the lowpass filter and the lower cutoff frequency of the bandpass filter. For these experiments the upper cutoff frequency of the bandpass filter was fixed to 12 Hz.

still meeting all the design goals. The lowpass filter did not include any DC suppression. The upper cutoff frequency of the bandpass filter was fixed at 12 Hz, and the lower cutoff frequency of the bandpass filter and the cutoff frequency of the lowpass filter were varied. The upper cutoff was set to 12 Hz because the experiments in Section 5.6.1 appeared to indicate that it could be set that low without significantly affecting recognition accuracy and because there were potential advantages to setting the cutoff to be as low as possible (for example, the filtered envelope signals could then be downsampled, reducing the computation required by the rest of the recognizer). As shown in Sections 5.8 and 6.2, the 12-Hz cutoff proved to be overly aggressive, discarding information needed for the discrimination of phone sets larger than the one used for Numbers. The MLP acoustic model used in the recognizers had 234 input units (26 features per frame × 9 frames of input), 344 hidden units, and 32 output units, for a total of ca. 92,000 weights.

The results of these experiments are summarized in Table 5.18. Although the results are somewhat variable, the best overall performance was obtained with an even partition of the modulation frequency range into a 0–6 Hz lowpass portion and a 6–12 Hz bandpass portion. It should be noted, though, that equally good performance could be obtained with a single set of lowpass MSG features computed with a lowpass filter having a 12 Hz cutoff and 8–11 dB of DC suppression.

The next experiment tested banks of three envelope filters with fixed bandwidths of 4, 5, or 6 Hz and minimal overlap covering modulation frequencies of 0–12 Hz, 0–15 Hz, and 0–18 Hz, respectively. Keeping the number of weights in the MLP acoustic model constant at around 92,000 meant that the hidden layer contained fewer units than the input layer, so a second series of experiments was run in which the number of hidden units was doubled. The results of these experiments are summarized in Table 5.19. None of the three-filter results were significantly better than the best two-filter results, so the next round of experiments focused on the two-filter features, with the filters having passbands of 0–6 Hz and 6–12 Hz.

Because the performance of MSG features generated with a single lowpass filter was improved by addition of DC suppression to the filter, it seemed reasonable to try adding DC suppression to the lowpass filter in the two-filter case as well. In a first set of experiments, a set of lowpass filters with a cutoff frequency of 6 Hz and variable degrees of DC suppression was generated by convolving a lowpass FIR filter with 2-point highpass

Filter	92,000 weights		184,000 weights	
Bandwidth (Hz)	clean	reverb.	clean	reverb.
4	8.9%	16.7%	8.2%	14.9%
5	8.4%	16.0%	7.3%	15.4%
6	7.9%	16.3%	7.2%	15.9%

Table 5.19: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and two sets of bandpass MSG features generated with a bank of three envelope filters with fixed bandwidth and minimal overlap between filters as a function of the filter bandwidth.

FIR filters having the form $H(z) = 1 - xz^{-1}$, where $0 < x \leq 1$, and the performance of MSG features generated with these lowpass filters and the 6–12 Hz bandpass filter was measured. The results of these experiments are summarized in Table 5.20.

The best overall performance was obtained for a DC suppression of 4 dB, with a word error rate of 8.1% on the clean test and a word error rate of 14.8% on the reverberant test. The performance on the reverberant test using the lowpass filter with 4 dB of DC suppression was significantly better than the performance obtained with no DC suppression. In general, the filters with DC suppression of no more than 8 dB gave good performance on both the clean and reverberant tests, although the results were somewhat variable.

The filters used in the first experiment had non-uniform group delay in their passbands, which could negatively impact recognizer performance by distorting the timing information in the features. Thus, a second set of experiments was run using linear-phase lowpass filters with variable amounts of DC suppression. The filters were generated by designing a lowpass FIR filter with a 6-Hz cutoff frequency and then manipulating the positions of the pair of real-axis zeroes in the filter. Figure 5.8 shows the pole-zero plot for the base lowpass filter. The results of these experiments are summarized in Table 5.21.

The results of these experiments were somewhat more consistent than the first set of experiments, with the best overall performance obtained for a DC suppression of 4 or 5 dB. For the next set of experiments the MSG features were generated using two linear-phase FIR filters: one lowpass filter with a cutoff frequency of 6 Hz and 5 dB of DC suppression and one bandpass filter with a passband of 6–12 Hz.

DC Magnitude Response (dB)	clean	reverb.
-1	8.2%	15.6%
-2	8.4%	16.0%
-3	8.5%	15.1%
-4	8.1%	14.8%
-5	8.4%	15.3%
-6	8.6%	14.8%
-7	8.8%	15.3%
-8	8.6%	15.2%
-9	9.1%	15.8%
-11	9.4%	16.4%
-12	9.3%	16.6%
-14	9.1%	15.9%
-17	9.3%	16.4%
-20	9.5%	17.1%
-26	10.0%	17.3%
-∞	9.7%	16.7%

Table 5.20: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features as a function of the level of DC suppression in the lowpass envelope filter. The lowpass filter cutoff frequency is 6 Hz, and the bandpass filter passband is 6–12 Hz. The lowpass filters were generated by convolution of a lowpass FIR filter with a set of highpass filters of the form $H(z) = 1 - xz^{-1}$, where $0 < x \leq 1$.

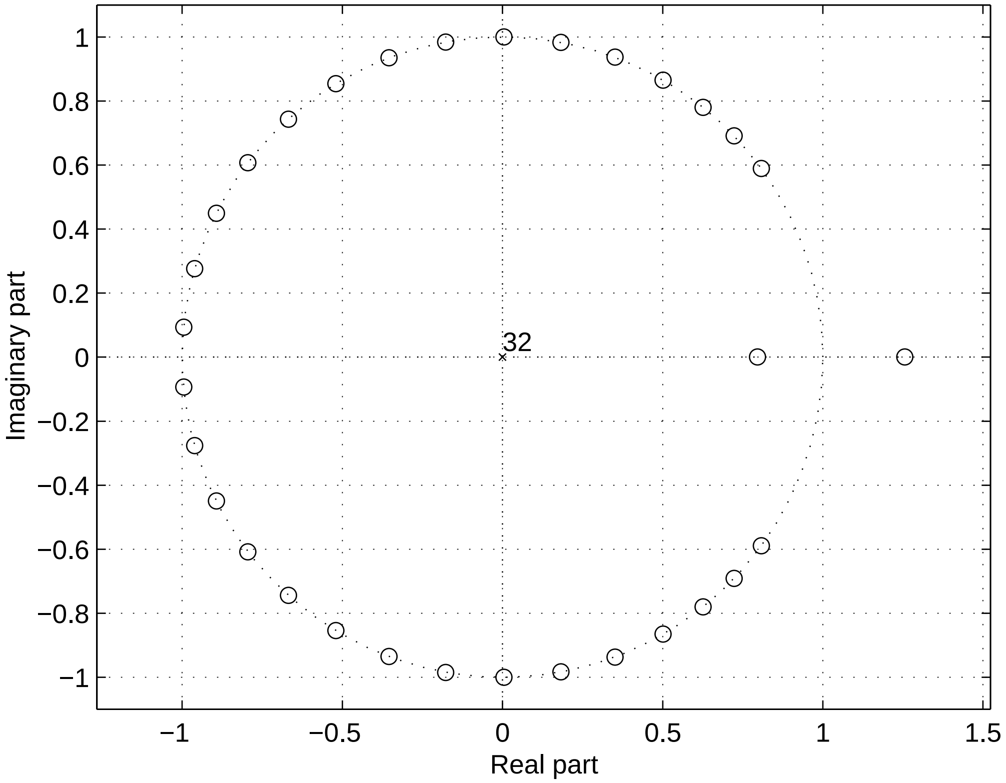


Figure 5.8: Pole and zero locations for the lowpass FIR filter from which a family of lowpass filters with variable amounts of DC suppression was derived by manipulating the positions of the pair of real-axis zeroes.

DC Magnitude Response (dB)	clean	reverb.
-1	8.0%	16.3%
-2	7.9%	15.7%
-3	8.3%	15.3%
-4	8.1%	14.9%
-5	8.5%	14.6%
-10	9.0%	15.8%
-15	9.6%	16.7%
-20	9.9%	16.7%
-25	10.2%	17.1%
-30	10.2%	17.3%
-35	10.2%	16.6%
-40	10.7%	17.6%
-∞	10.3%	17.0%

Table 5.21: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features as a function of the level of DC suppression in the lowpass envelope filter. The lowpass filter cutoff frequency is 6 Hz, and the bandpass filter passband is 6–12 Hz. The lowpass filters were generated by manipulating the locations of the pair of real-axis zeroes in a base filter.

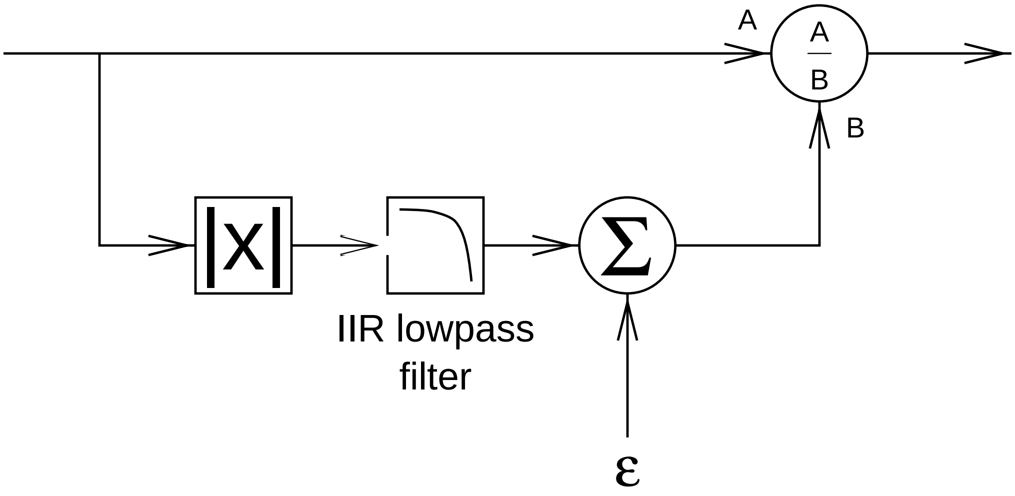


Figure 5.9: Implementation of the feedforward AGC unit.

5.7 A Feedforward AGC Design

The design of the on-line normalization processing suggested another AGC design: the feedforward design illustrated in Figure 5.9. Like the other on-line AGC designs, this feedforward design enhances signal onsets, and like the AGC design described in Section 5.4.2, it is a "true" AGC in the sense that its output converges to 1 for a constant input.

In the first series of experiments with this AGC design, it replaced the two feedback AGC units. The time constants for the AGCs applied to the lowpass and bandpass MSG features were varied independently, and the value of ε was set to 1, based on the results from the on-line normalization experiments. The results of these tests are summarized in Table 5.22.

The best performance was obtained with $\tau = 320$ or 640 ms for both the lowpass and bandpass MSG features. Performance on the reverberant test was not as good with the feedforward AGC as it was with the two feedback AGC units (see Table 5.21). Thus, a second series of experiments were run in which two feedforward AGC units were used in series, with a variable time constant for the first AGC and a fixed time constant of 320 ms for the second AGC. In these experiments the AGCs for the lowpass and bandpass features had the same time constants. The results, summarized in Table 5.23, were significantly

	Lowpass	Bandpass AGC τ (ms)				
Condition	AGC τ (ms)	40	80	160	320	640
clean	40	11.5%	10.8%	10.2%	9.4%	9.3%
	80	10.8%	10.5%	10.0%	8.9%	8.5%
	160	10.0%	9.8%	9.0%	8.6%	8.8%
	320	9.3%	9.2%	8.6%	8.0%	8.2%
	640	9.1%	9.0%	8.3%	8.1%	8.0%
reverb.	40	23.3%	22.0%	22.3%	21.2%	20.4%
	80	20.6%	19.5%	18.3%	18.0%	18.6%
	160	18.7%	17.8%	17.5%	17.7%	17.0%
	320	18.7%	17.9%	17.1%	17.0%	17.2%
	640	18.1%	17.2%	16.8%	16.6%	16.8%

Table 5.22: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features computed with a single feedforward AGC as a function of the AGC time constant, τ.

First AGC τ (ms)	clean	reverb.
40	12.5%	26.8%
80	10.9%	21.2%
160	10.1%	18.4%

Table 5.23: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features computed with two feedforward AGCs as a function of the time constant of the first AGC unit. The time constant of the second AGC unit was fixed at 320 ms.

worse with two feedforward AGCs than they were with a single feedforward AGC. Based on the results of these experiments, the two feedback AGCs were retained in subsequent experiments.

5.8 Using Broader Envelope Filters

A set of experiments that tested the efficacy of the MSG features for recognition of the large-vocabulary Broadcast News corpus were proceeding concurrently with the Numbers 95 experiments described in this chapter. While the Broadcast News experiments will be described in more detail in Chapter 6, they must be mentioned briefly here because results with Broadcast News altered the course of the Numbers 95 experiments. Recognizers for the Broadcast News task had been trained with two different sets of MSG features:

1. The best features from Section 5.5, which were computed with

   - a power-spectral Bark-scale filterbank with triangular filters,
   - IIR envelope filters with passbands of 0–16 Hz and 2–16 Hz,
   - and two feedback AGCs with $\tau_1 = 160$ ms and $\tau_2 = 320$ ms for the lowpass MSG features and with $\tau_1 = 160$ ms and $\tau_2 = 640$ ms for the bandpass MSG features,

   and

2. The best features from Section 5.6.3, which were computed in the same way as the other feature set, expect that the envelope filters were FIR filters with passbands of 0–6 Hz and 6–12 Hz, and the lowpass filter included 5 dB of DC suppression.

While the FIR-filter-based MSG features gave better performance on the Numbers 95 task, the IIR-filter-based MSG features gave significantly better performance on the Broadcast News task. It was hypothesized that the 12–16 Hz modulation frequency range, which had not appeared to be important for the Numbers 95 recognition task in the experiments in Section 5.6, were important for accurate recognition of the large-vocabulary Broadcast News task. This difference may be understood by comparing the number of phones and phonetic contexts represented in the two tasks. The Numbers task requires the recognition of just thirty-two different words based on a phone set of thirty-two elements. In contrast, the Broadcast News task requires the recognition of 65,000 different words based on a phone set of fifty-four elements. It is not surprising that the discrimination of a larger set of phones occurring in a much more diverse array of contexts requires a more (temporally) detailed description of the input. Thus, an alternate set of FIR envelope filters with passbands of 0–8 Hz and 8–16 Hz were designed and tested on the Numbers 95 tests. The lowpass filter included a variable amount of DC suppression. The results of these experiments are summarized in Table 5.24.

As in the experiments in Section 5.6.3, the best performance was obtained with a DC suppression of 5 dB. Compared to the features computed with the 0–6 Hz and 6–12 Hz

DC Magnitude Response (dB)	clean	reverb.
0	7.8%	17.0%
-5	7.5%	15.7%
-10	8.1%	15.8%
-15	8.3%	16.0%

Table 5.24: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features computed with filters having passbands of 0–8 Hz and 8–16 Hz, respectively, as a function of the level of DC suppression in the lowpass filter.

FIR filters, these features gave better performance on the clean test (by 1% absolute) and worse performance on the reverberant test (by 1.1%). As will be seen in Chapter 6, these broader filters gave much better performance on the Broadcast News task (although the best performance was obtained with the 0–16 Hz and 2–16 Hz IIR filters), so all subsequent experiments with Numbers used these filters. The impulse responses for the two filters are shown in Figure 5.10 and their frequency responses are shown in Figure 5.11. In contrast with the presentation of the IIR filters in Section 5.1, it is not necessary to plot the filter group delay characteristics here because the FIR filters are linear-phase filters that, by definition, have uniform group delay characteristics.

5.9 Verifying the AGC Time Constants

Because significant changes had been made to the MSG processing since the time constants for the feedback AGCs had been set (in Section 5.2), it seemed prudent to test different values for time constants again. In these experiments the time constants of the first and second AGCs for the lowpass and bandpass MSG features were varied independently of one another. Time constants of 80 and 160 ms were tested for the first AGC. Time constants of 160, 320, or 640 ms were tested for the second AGC. Table 5.25 summarizes the experimental results. There was very little variation in performance on the clean tests. In fact, none of the differences in performance on the clean test are statistically significant. On the reverberant tests, performance was consistently better with $\tau_1 = 160$ ms and $\tau_2 = 160$ or 320 ms for the computation of the lowpass MSG features. There was no clear, best choice

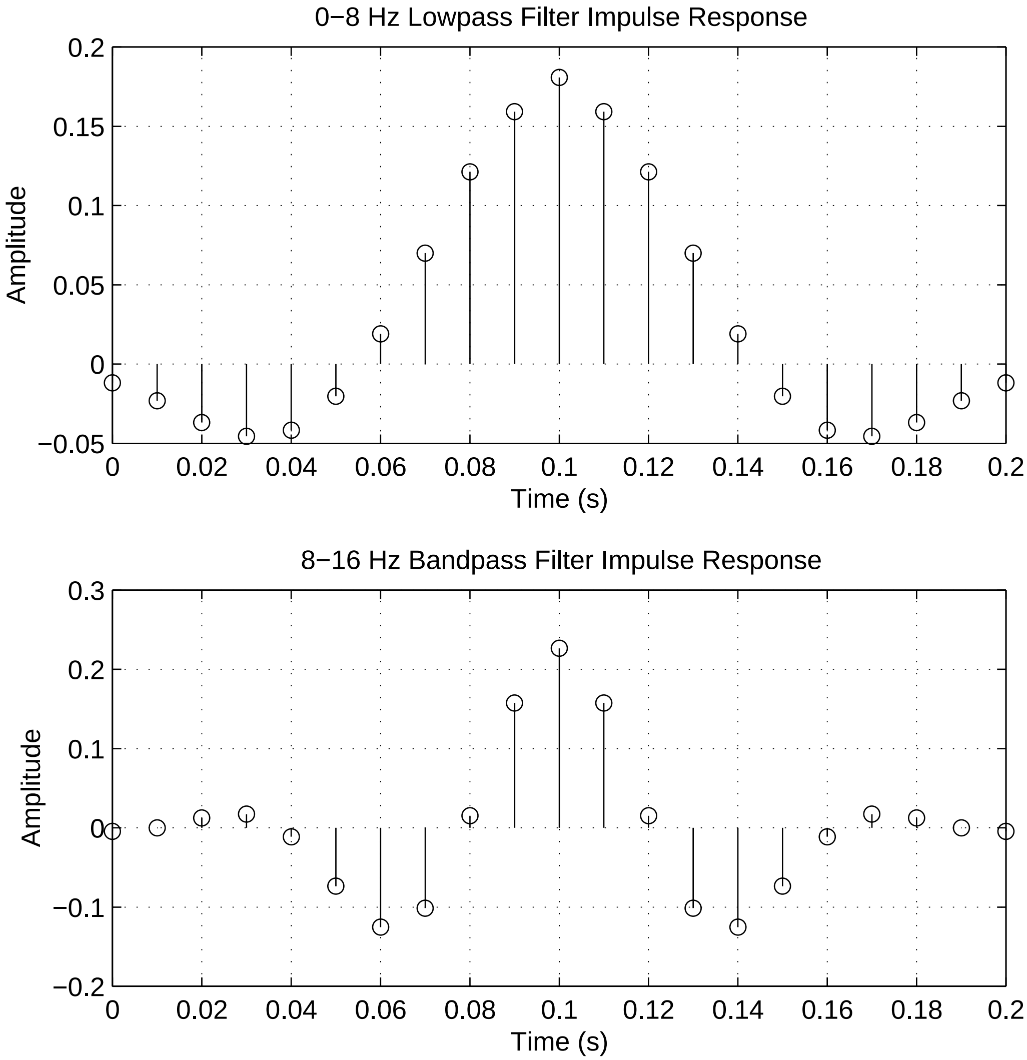


Figure 5.10: Impulse responses of the lowpass and bandpass FIR filters chosen for the final version of the MSG features used in the Numbers experiments.

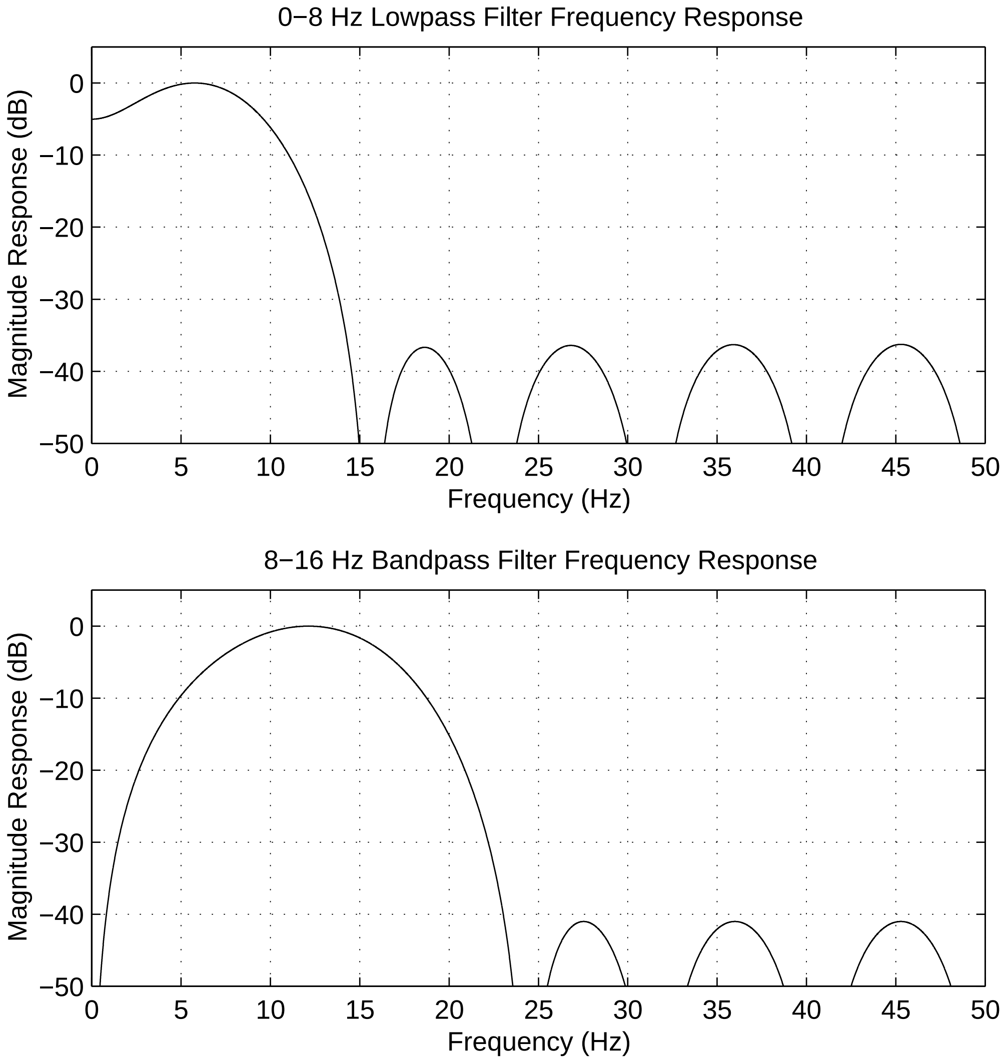


Figure 5.11: Frequency responses of the lowpass and bandpass FIR filters chosen for the final version of the MSG features used in the Numbers experiments.

	Lowpass AGC τ (ms)		Bandpass AGC τ_2 (ms)					
Condition	τ_1	τ_2	160	320	640	160	320	640
clean	80	160	7.5%	7.2%	7.1%	7.6%	7.3%	7.7%
		320	7.3%	7.1%	7.0%	7.1%	7.2%	7.0%
		640	7.1%	7.0%	7.2%	7.2%	7.1%	7.5%
	160	160	7.6%	7.3%	7.1%	7.6%	7.6%	7.7%
		320	7.8%	7.4%	7.3%	7.3%	7.5%	7.3%
		640	7.9%	7.8%	7.4%	7.6%	7.9%	7.6%
reverb.	80	160	16.7%	17.3%	16.4%	16.6%	17.4%	16.3%
		320	17.1%	16.2%	16.6%	16.7%	16.1%	16.1%
		640	17.7%	16.8%	16.9%	16.9%	16.6%	17.3%
	160	160	16.3%	15.1%	15.8%	15.8%	16.6%	15.5%
		320	16.3%	15.8%	15.7%	16.2%	15.7%	15.3%
		640	17.0%	16.3%	15.7%	16.0%	16.6%	16.1%
			80			160		
			Bandpass AGC τ_1 (ms)					

Table 5.25: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features as a function of the time constants of the feedback AGC units.

	Lowpass AGC τ (ms)		Bandpass AGC τ_2 (ms)					
Condition	τ_1	τ_2	160	320	640	160	320	640
clean	160	160	7.4%	7.3%	7.3%	7.6%	7.5%	7.3%
		320	7.3%	7.4%	7.4%	7.7%	7.4%	7.8%
reverb.	160	160	16.0%	15.4%	15.6%	15.7%	16.0%	15.2%
		320	16.2%	15.5%	16.0%	16.2%	15.3%	14.8%
			80			160		
			Bandpass AGC τ_1 (ms)					

Table 5.26: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using lowpass and bandpass MSG features as a function of the time constants of the feedback AGC units. These tests used a fixed MLP training schedule to try to reduce the variance of the results.

for the time constants for the computation of the bandpass features, however, due to the variability of the recognition results.

An examination of records from the different experiments showed that there was some correlation between the performance of a recognizer on the clean and reverberant tests and the number of epochs of training performed on the MLP acoustic model. Recall from Section 3.1.2 that MLP training stops when the performance on a cross-validation set improves by less than 0.5% in a training epoch. For the Numbers task, it is possible that this stopping criterion leads to too early a termination of MLP training because the recognizers which were trained for more epochs generally outperform recognizers which were trained for fewer epochs. To try to reduce the variability of the experimental results, the MLP training schedule was fixed to four epochs with a learning rate of 0.008 followed by six epochs with a learning rate one-half of the learning rate from the previous epoch (thus, the learning rate in the final epoch is 0.000125).

Some of the experiments from Table 5.25 were performed again using the new, fixed MLP training schedule. The results of these experiments are summarized in Table 5.26. While the variability of the results was reduced somewhat, there was still no clear, best choice of AGC time constants. For subsequent experiments, settings of $\tau_1 = 160$ ms and $\tau_2 = 320$ ms were selected for both the lowpass and bandpass features. The fixed MLP training schedule was also used in all subsequent experiments with the Numbers 95 data.

Lowpass	Bandpass Order					
Order	13	11	9	13	11	9
13	7.4%	7.4%	7.3%	15.6%	16.3%	15.3%
11	7.4%	7.5%	7.3%	15.8%	16.6%	15.5%
9	7.2%	7.2%	7.5%	15.3%	15.8%	15.7%
	clean tests			reverb. tests		

Table 5.27: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using MSG features smoothed by DCT truncation as a function of the number of DCT coefficients used for each stream. These results should be compared to those in Table 5.26 with $\tau_1 = 160$ ms and $\tau_2 = 320$ ms.

5.10 Spectral Smoothing of the Features

The spectral resolution of the MSG features, while reduced slightly by the change from a quarter-octave filterbank to a Bark-scale filterbank, was still higher than most other representations for ASR. The MSG processing produced 26 features per frame (lowpass and bandpass features streams with 13 features per frame in each stream), while comparable PLP or RASTA-PLP processing produced 18 features per frame (static and delta feature streams with 9 features per frame in each stream). As shown in Section 5.3, simply reducing the resolution of the initial frequency analysis was not a viable way to reduce the overall resolution of the MSG representation because it also reduced recognition accuracy. The experiments described in this section explored another way to reduce the spectral resolution of the MSG representation: spectral smoothing of the features following the AGC but prior to on-line normalization.

In the first set of spectral smoothing experiments, the discrete cosine transform (DCT) of the features was computed (with separate DCTs applied to the lowpass and bandpass streams), and smoothing was performed by discarding some number of the higher-order DCT coefficients. The results of these experiments are summarized in Table 5.27. There was no significant difference in performance between the results with the untransformed features (see Table 5.26) and the transformed features except for the reverberant test using the lowest 11 DCT coefficients from both the lowpass and bandpass streams, where performance was significantly worse than the performance with untransformed features. It is likely that this case is an outlier, because it is not consistent with the other test results.

From these results, it appears that the resolution of the MSG representation could be reduced to match that of the PLP and RASTA-PLP front ends using DCT truncation, with minimal impact on recognizer performance.

Although DCT truncation is an effective method for smoothing spectral features, it does have an important disadvantage in that it transforms local spectral distortions into global distortions of the features. For example, if a single channel in the spectral representation coincides with a spectral zero in the speech transmission channel, the spectral features will be unaffected, with the exception of the one that coincides with the zero. However, if the DCT of the features is computed, all of the DCT coefficients will be changed by the presence of the spectral zero in a single channel.

This transformation of local spectral distortions into global feature distortions poses a problem for some speech recognition architectures, especially multi-band approaches [BD96, HTP96, Mir98] that perform independent classification or recognition on separate frequency bands and then combine the frequency-local decisions. Thus, a second smoothing method that performs frequency-local smoothing was also tested. In this method the features in each stream are transformed by replacing adjacent pairs of spectral features with the sum and difference of the features. In matrix notation, the transformed feature vector $\mathbf{y}$ is computed as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}$$

where $\mathbf{x}$ is the vector of features for a frame and n is the number of features. Because n must be even, the bandwidth and spacing of the filters in the Bark-scale power spectral filterbank is reduced from 1.0 Bark to 0.95 Bark, producing a representation with 14 spectral channels. Thus, the transformation, which is essentially a Haar transform in the frequency domain, produces 7 sum terms and 7 difference terms. Recognition tests were run with three conditions:

Lowpass	Bandpass Transform					
Transform	s+d	sum	diff.	s+d	sum	diff.
s+d	7.6%	7.3%	8.7%	15.3%	14.8%	17.5%
sum	7.4%	8.2%	9.9%	15.3%	16.8%	18.3%
diff.	9.8%	9.5%	19.4%	16.5%	17.3%	31.3%
	clean tests			reverb. tests		

Table 5.28: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using Haar-transformed MSG features. In the **s+d** condition both the sum and difference terms from the transform were used, and thus no smoothing was performed. In the **sum** and **diff.** conditions only the sum or difference terms, respectively, were used, and thus some smoothing was performed. If no transformation is applied to the features, the word error rate is 7.5% on the clean test and 15.0% on the reverberant test.

1. using both the sum and difference terms,
2. using just the sum terms,
3. or using just the difference terms,

for each of the lowpass and bandpass streams. The results of these experiments are summarized in Table 5.28.

If no transformation was performed on the features, the word error rate on the clean test was 7.5% and the word error rate on the reverberant test was 15.0%. These results are not significantly different from those obtained with 13 spectral channels. With the transformed features, using only the difference terms from either or both the lowpass and bandpass streams led to significantly worse performance on both the clean and reverberant tests. Performance was also significantly worse on the reverberant test if only the sum terms from both streams were used. Thus, the total number of features per frame in the MSG representation could be reduced to 21 by using only the sum terms for either the lowpass or bandpass stream without affecting recognizer accuracy. This smoothing was local in frequency, and therefore a better match to some recognition algorithms. Reduction of the total number of features produced by the MSG representation is desirable because it may reduce the number of weights required in the MLP acoustic model. A final test was performed in which the lowpass stream was not processed and the only the sum terms from the Haar transform of the bandpass stream were used. With this feature set, the word error rate on the clean test was 7.4% and the word error rate on the reverberant test was 14.9%. This was the feature set selected for use in subsequent experiments.

5.11 Changing the Size of the Context Window

There is considerable evidence (reviewed in Section 2.2.3) that humans base their phonetic judgments on information integrated over segments of the speech signal roughly 200–250 ms in duration. This integration over relatively long segments may contribute to the reliability of human speech recognition by averaging out transient distortions in the speech signal and by capturing phonetic information that is distributed over syllable-like intervals by coarticulation. The automatic recognizers described so far in this thesis base their acoustic likelihood estimates on shorter segments of only 90 ms. In this section, the effect of increasing the duration of the input to the MLP acoustic model is measured.

As in the other experiments, the number of MLP weights was held constant by varying the number of hidden units as the size of the input layer changed. For these experiments the number of MLP weights was increased to 164,000 so that there would be a reasonably large number of hidden units even for the largest input size. In the first set of experiments, recognizers using the MSG features or log-RASTA-PLP were tested with MLPs taking 9, 13, 17, 21, or 25 frames of input. The log-RASTA-PLP features were normalized using exactly the same on-line normalization as the MSG features. The results of these experiments are summarized in Table 5.29. For both the MSG and RASTA features, the best overall performance was obtained with an input duration of 13 frames (130 ms). A second set of experiments, summarized in Table 5.30, were performed using PLP features with on-line normalization and MLPs using 9, 13, or 17 frames of input. Again, the best performance was obtained for an input duration of 13 frames. Based on the results of these experiments, the duration of the MLP input was increased to 13 frames in the subsequent experiments, and PLP features were compared against the MSG features instead of log-RASTA-PLP features.

The results with the MLP context window size clearly do not match human speech perception, in which phonetic judgments appear to be based on segments of the speech signal that are roughly 200–250 ms in duration. This difference may be the result of the different units of recognition employed by human listeners and the automatic recognizers

Num. Frames Context	Features			
	MSG	RASTA	MSG	RASTA
9	7.1%	6.3%	13.9%	23.3%
13	6.7%	6.3%	13.9%	22.4%
17	7.0%	6.5%	14.2%	22.3%
21	7.0%	6.5%	14.1%	22.4%
25	7.5%	7.1%	15.2%	23.0%
	clean tests		reverb. tests	

Table 5.29: Word error rates for the clean and reverberant Numbers 95 development tests for recognizers using MSG features or log-RASTA-PLP features with on-line normalization as a function of the length of the MLP input.

Num. Frames Context	Condition	
	clean	reverb.
9	5.7%	23.6%
13	5.7%	20.9%
17	6.0%	21.6%

Table 5.30: Word error rates for the clean and reverberant Numbers 95 development tests for a recognizer using PLP features with on-line normalization as a function of the length of the MLP input.

used in this thesis. It is most likely that syllable-like units that cover 110–260 ms segments of the speech signal [GHE96] play a prominent role in human speech perception. The automatic recognizers used in this work use phone-like units that cover 60–100 ms segments of the speech signal [GHE96]. The shorter input to the MLP acoustic model may be a better match to the automatic recognizer's shorter units of recognition. While it appears that coarticulation spreads information about phonetic segments over entire syllables, the automatic recognizers may not be able to use all of this information, either because there is not enough training data available or because longer recognition units may be required to exploit this information, as suggested in [Wu98].

5.12 The Final Version of the MSG Features for Numbers

The series of experiments described in this chapter ultimately led to the MSG processing illustrated in Figure 5.12. This feature extraction algorithm, which proved to be the best design for the Numbers recognition task, proceeds according to the following steps:

1. The speech signal is segmented into 25-ms frames with a 10-ms frame step, each frame is multiplied by a Hamming window, and the power spectrum for each frame is computed with an FFT.

2. The power spectrum is accumulated into critical-band-like frequency channels via convolution with a bank of fourteen overlapping, triangular filters that have bandwidths and spacings of 0.95 Bark and cover the 230–4000 Hz range. The critical-band-like power spectrum is converted into an amplitude spectrum by taking the square root of the filterbank output. The experiments leading to this choice of filterbank are described in Section 5.5.

3. The critical-band-like amplitude signals are filtered by two different FIR filters in parallel: a lowpass filter with a 0–8 Hz passband and 5 dB of DC suppression, and a bandpass filter with an 8–16 Hz passband. The experiments leading to this pair of filters are described in Section 5.8.

4. Both the lowpass and bandpass streams are processed through two feedback AGC units where the first AGC a time constant of 160 ms and the second has a time constant

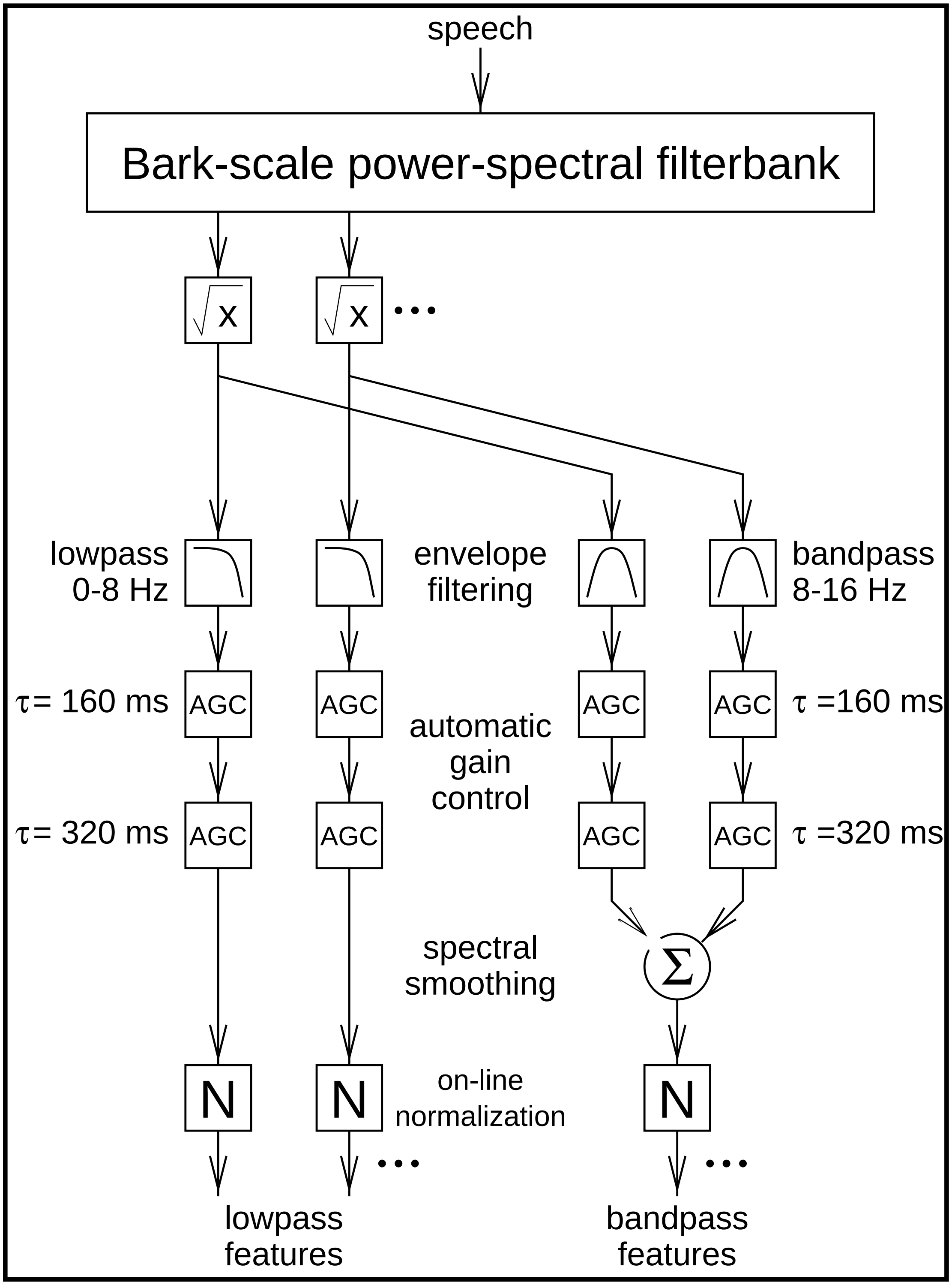


Figure 5.12: Signal processing for the final version of the modulation-filtered spectrogram (MSG) features used for Numbers recognition.

of 320 ms. See Figure 5.5 for an illustration of the AGC design and Sections 5.2 and 5.9 for descriptions of the experiments used to select the number of AGC units and their time constants.

5. The spectral resolution of the bandpass stream is halved by summing the features from adjacent channels. The experiments summarized in Section 5.3 led to this decision.

6. All features are normalized to have means of zero and variances of one using an on-line normalization procedure. The feature means and variances are estimated using single-pole lowpass filters with a time constant of 2 s. Figure 5.7 illustrates the normalization processing, and Section 5.4.3 describes the experiments in which the parameters of the normalization were selected.

5.13 Optimizing the Lexicon

All of the recognizers described so far in this chapter have used the lexicon described in Section 3.1.3: a multiple-pronunciation lexicon in which the word pronunciations and phone durations were optimized in an iterative embedded training procedure that used a recognizer based on log-RASTA-PLP features. Thus, the recognizers used a lexicon that was not matched to the features they used. As a final step in the recognizer optimization, new lexicons were created for three different recognizers and optimized via embedded training. It was expected that recognizer performance would improve somewhat if the lexicon was better matched to the features. The three recognizers were:

1. A recognizer that used the final version of the MSG features for the Numbers experiments.

2. A recognizer that used eighth-order PLP features and delta-PLP features (calculated via linear regression over a nine-frame window) normalized using the same on-line normalization procedure that was applied to the MSG features.

3. A recognizer that combined the MSG and PLP features. In this recognizer, two MLPs were trained, with one taking MSG features as input and the other taking PLP features as input. The MLPs were trained with identical training targets. During recognition and forced alignment the acoustic likelihoods from the two MLPs were

MSG	PLP	Combined
6.4%	5.8%	5.5%

Table 5.31: Word error rates on the clean Numbers 95 development test set for recognizers using optimized lexicons and either MSG features, PLP features with on-line normalization, or a combination of MSG and PLP features.

combined by averaging them. The MLPs used in the combined recognizer had half as many weights as those in the other two recognizers, so all three systems had the same total number of MLP weights.

For all three recognizers, the training process began with the hand-transcribed labels and the lexicon that included 90% of the training set pronunciations and had phone durations derived from the hand transcriptions. For each recognizer, several iterations of embedded MLP and lexicon training were performed. The performance of the recognizers was measured on the clean Numbers 95 development test set, and the best-performing recognizer was chosen.

Table 5.31 shows the performance using the three best lexicons on the clean development test set. Compared to the tests (see Tables 5.29 and 5.30) using the original lexicon, which had been optimized for log-RASTA-PLP features, performance of the MSG recognizer improved and performance of the PLP recognizer degraded, each by statistically insignificant amounts.

5.14 Summary

A comparison of Figures 4.9 and 5.12 shows that over the course of the experiments described in this chapter nearly all of the details of the MSG processing were changed:

- The implementation of the initial frequency analysis was changed from direct FIR filtering to filtering in the power-spectral domain.

- The filter shapes used for the initial frequency analysis were changed from trapezoidal to triangular and the filterbank resolution was changed from quarter-octave to 0.95 Bark.

- The envelope filters ultimately remained FIR filters, but their design changed and the passband of the bandpass filter was changed from 2–8 Hz to 8–16 Hz.
- The cube-root compression and off-line normalization of the features with respect to the global peak was replaced by a series of two on-line, feedback AGC units and an on-line normalization of the feature means and variances.
- Processing to halve the spectral resolution of the bandpass features was added.

As shown in Section 6.2, a slightly different version of these features that used broader, IIR envelope filters proved to be best for the large-vocabulary Broadcast News task, although this difference may reflect a misalignment between the features and training labels in the Broadcast News experiments rather than some fundamental difference between Broadcast News and Numbers.

A comparison of the MSG results in Table 4.11 and the performance of the best MSG features for Numbers recognition (see the conclusion of Section 5.10) shows that these changes improved the word error rate on the clean Numbers 95 development test set from 8.5% to 7.4%, for a 13% relative reduction in error rate, and they improved the word error rate on the reverberant Numbers 95 development test set from 27.3% to 14.9%, for a 45% relative reduction in error rate. Increasing the number of MLP weights by about 74%, increasing the MLP input length from 9 to 13 frames, and training a new lexicon yielded additional improvements in performance for the MSG recognizer.

Chapter 6

Testing the Generality of the Features

All of the development of the modulation-filtered spectrogram features was based on experiments with the Numbers 93 and Numbers 95 tasks, with the exception of several experiments on the Broadcast News corpus described in Section 6.2. Numbers is a small-vocabulary corpus with a relatively small set of training and testing utterances, and nearly all of the recognition tests were based on just two acoustic conditions: a clean test and a moderately reverberant test ($T_{60} = 0.5$ s and a direct-to-reverberant energy ratio of 1 dB). The small size of the Numbers lexicon and the restricted number of test conditions made it feasible to perform the very large number of recognition experiments that guided the development of the MSG features. There is a potential danger, however, that the MSG features could be overly specialized for the Numbers task and the clean and reverberant tests. To test the utility of MSG features for a broader array of tasks and acoustic conditions, they were tested on a final set of Numbers 95 test utterances in many different acoustic conditions, as well as on the large-vocabulary Broadcast News task.

6.1 Final Numbers 95 Tests

The Numbers 95 final test set is a collection of 1227 utterances that were set aside and not used in any of the previously described experiments. To verify that the MSG features are useful for acoustic conditions not represented in any of the recognition tests

MSG	PLP	combo
6.1%	5.9%	4.7%

Table 6.1: Word error rates for MSG, PLP, and combined recognizers on the clean Numbers 95 final test set.

performed during the MSG development, recognition performance was measured for the MSG, PLP, and combined recognizers from Section 5.13 on the final Numbers 95 test set in a collection of different acoustic conditions: clean, reverberation, noise, spectral shaping, and simultaneous noise and reverberation.

6.1.1 Tests Under Clean Conditions

The results of the clean tests are summarized in Table 6.1. The word error rates of the MSG and PLP recognizers on the final test set are not significantly different from one another, while the performance of the combined recognizer (which has exactly the same number of MLP weights as either of the other two recognizers) is significantly better.

6.1.2 Tests Under Reverberant Conditions

The impulse response used to generate the moderately reverberant Numbers 95 development test set was one of a set of twelve impulse responses recorded in the varechoic chamber at Bell Labs (as described in Section 4.3.1). For the final tests, each of the impulse responses was used to generate twelve different reverberant versions of the Numbers 95 final test set. The performance of the MSG, PLP, and combined recognizers on these different reverberant tests is summarized in Table 6.2. The impulse responses are described in terms of the recording conditions (percentage of open panels in the chamber and microphone position) and also in terms of reverberation time and direct-to-reverberant energy ratio. The MSG recognizer outperforms the PLP recognizer in all reverberant conditions by a significant margin, while the combined recognizer outperforms the MSG recognizer in all conditions (and by a significant margin in nine of the twelve tests). The fifth line of Table 6.2 (43% of panels open, microphone #1) gives results for the impulse response used to generate the reverberant development test used in Chapter 5.

Panels Open (%)	Mike #	T_{60} (s)	direct-to-reverberant energy ratio (dB)	MSG	PLP	combo
100	1	0.3	1	9.4%	12.0%	8.3%
	2	0.3	1	8.5%	10.6%	7.3%
	3	0.3	-1	10.2%	11.9%	8.9%
	4	0.3	-1	9.2%	11.5%	8.6%
43	1	**0.5**	**1**	**13.8%**	**22.2%**	**13.0%**
	2	0.5	-3	13.8%	20.3%	12.8%
	3	0.5	-2	15.5%	20.2%	13.2%
	4	0.5	-5	16.7%	22.7%	15.1%
0	1	0.9	-5	38.2%	55.3%	35.9%
	2	0.9	-7	38.9%	53.3%	35.7%
	3	0.9	-7	45.4%	58.3%	41.3%
	4	0.9	-9	45.2%	57.3%	40.4%

Table 6.2: Word error rates for MSG, PLP, and combined recognizers on the reverberant Numbers 95 final test sets. The boldface entry in the table indicates results with the impulse response used to generate the reverberant test set that was used in the recognition experiments described in Chapter 5. All other results in the table are for impulse responses not previously tested.

6.1.3 Tests Under Noisy Conditions

Noisy versions of the final test set were generated by adding three different noise samples to the test utterances at SNRs of 30, 20, 10, and 0 dB, with the SNR being measured on an utterance-by-utterance basis. The three different noises were:

Car noise recorded over a cellular telephone in a 1978 Volvo 244 running at 55 miles/hour on a freeway with the windows closed [MH92, HM94],

Babble noise from the NOISEX CD-ROM [VS93], downsampled to 8 kHz, and

Channel noise from a high-frequency radio channel on the NOISEX CD-ROM, downsampled to 8 kHz.

The normalized average power spectra for the three noises are shown in Figure 6.1. The noises have somewhat different spectral shapes, with the car and babble noises having lowpass shapes and the HF channel noise being relatively flat between 0.2 and 3 kHz. The noises also have different temporal characteristics, as can be seen from the modulation spectra for the three noises shown in Figure 6.2. The babble noise (not surprisingly) has a temporal structure similar to that of speech, but with less temporal variation than speech from a single speaker. In contrast, the car and HF channel noises are relatively stationary.

The results of the noisy recognition tests are summarized in Table 6.3. In all but one condition the MSG recognizer outperforms the PLP recognizer, and in the one condition where the PLP recognizer is better, the margin is not statistically significant. In nine of the noise conditions the MSG recognizer is significantly more accurate than the PLP recognizer. In all but one condition the combined recognizer is better than the MSG recognizer, and in the one condition where the MSG recognizer is better the margin is not statistically significant. In eight of the tests the combined recognizer is significantly more accurate than the MSG recognizer.

Although the improvement of recognizer robustness to additive noise was not a major goal in the development of the MSG features, they nonetheless proved to be somewhat noise-robust. This is likely a consequence of the use of relatively general signal-processing strategies in the MSG features that exploit the specific temporal properties of speech, and thus render the MSG representation resistant to a range of more slowly-varying forms of interference.

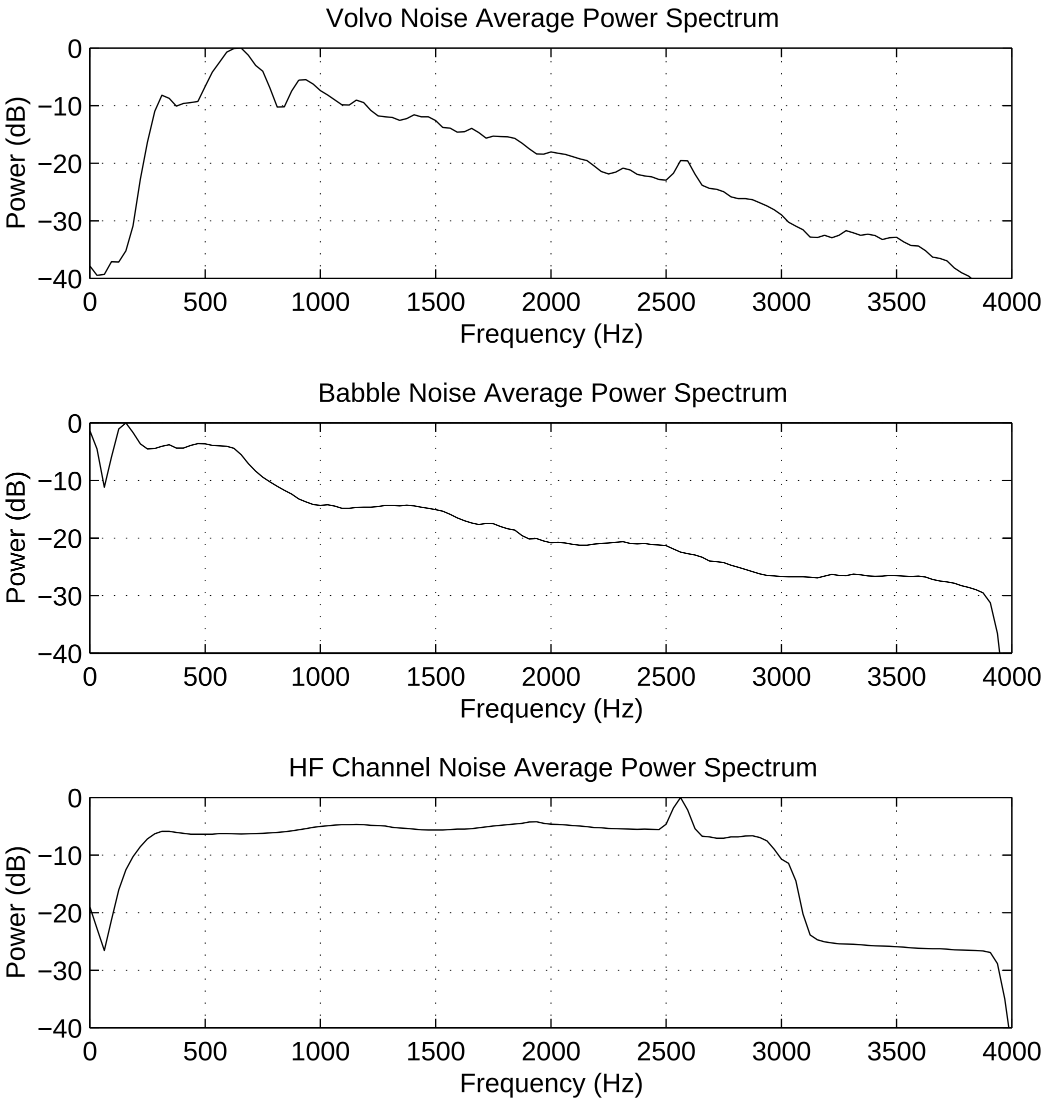
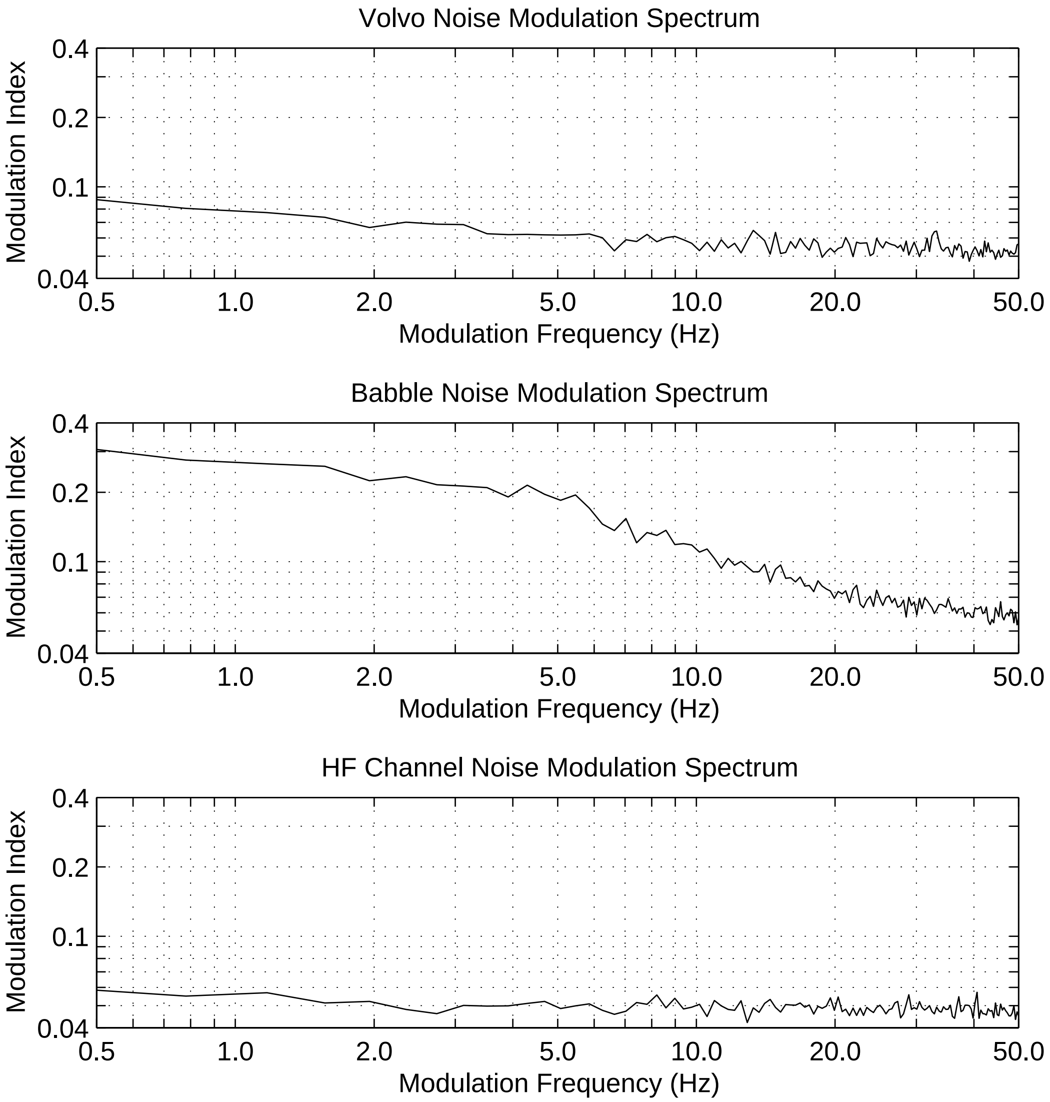


Figure 6.1: Normalized average power spectra for the three noise sample used to create the noisy Numbers 95 final test sets.



Volvo Noise Modulation Spectrum

Babble Noise Modulation Spectrum

HF Channel Noise Modulation Spectrum

Figure 6.2: Modulation spectra for the three noises used to create the noisy Numbers 95 final test sets.

Noise	SNR (dB)	MSG	PLP	combo
Volvo	30	6.2%	6.7%	4.9%
	20	7.2%	10.5%	7.0%
	10	13.1%	24.6%	13.2%
	0	42.5%	63.3%	39.6%
Babble	30	6.7%	6.2%	5.0%
	20	7.8%	9.1%	6.4%
	10	17.5%	21.7%	13.5%
	0	57.4%	59.3%	46.2%
HF Channel	30	6.5%	7.1%	5.5%
	20	8.4%	11.6%	7.7%
	10	16.2%	26.0%	16.0%
	0	48.1%	63.7%	42.2%

Table 6.3: Word error rates for MSG, PLP, and combined recognizers on the noisy Numbers 95 final test sets.

6.1.4 Tests with Spectral Shaping

Next, the performance of the recognizers was measured for data having unknown spectral shaping (i.e., convolutional distortion with an impulse response whose energy falls mostly or completely within the temporal window of the recognizer's spectral analysis stage). Robustness to spectral shaping is desirable in telephone applications. Versions of the Numbers 95 final test set with unknown spectral shaping were created by convolving the test utterances with one of four different linear-phase FIR filters:

A differentiator that imposed a +6 dB/octave spectral tilt on the utterances.

An integrator that imposed a -6 dB/octave spectral tilt on the utterances. The integrator had a flat frequency response for frequencies of 200 Hz and below.

A "Numbers 95 to Broadcast News" filter that was designed by computing the average power spectrum of the clean Numbers 95 training set and the average power spectrum of a collection of telephone-bandwidth Broadcast News utterances. This filter was designed to have a frequency response equal to the average Broadcast News spectrum divided by the average Numbers 95 spectrum, using a least-squares design procedure.

Filter	MSG	PLP	combo
differentiator	6.1%	6.5%	5.0%
integrator	6.1%	6.3%	4.9%
Num95 to BN	6.4%	7.0%	4.9%
random	7.5%	10.5%	6.5%

Table 6.4: Word error rates for MSG, PLP, and combined recognizers on the spectrally shaped Numbers 95 final test sets.

A random filter that was designed using a least-squares procedure to match a set of 64 (frequency, gain) pairs where the frequency points were equally spaced over the 0–4 kHz range and the gains were randomly selected from a uniform distribution over the range -10 dB to +10 dB.

The frequency responses of the four filters are illustrated in Figure 6.3.

Recognition results for the MSG, PLP, and combined recognizers for the four different spectral shaping conditions are given in Table 6.4. The MSG recognizer is more accurate than the PLP recognizer in all conditions, but by a significant margin only in the random filter case. The combined recognizer is significantly more accurate than the MSG recognizer in all conditions.

6.1.5 Tests Under Noisy Reverberant Conditions

A final set of recognition tests measured the performance of the three recognizers in the presence of simultaneous noise and reverberation. The NOISEX babble noise was used in these tests. Only three of the reverberant conditions, namely the microphone #4 impulse responses, were used because the reverberant tests showed that the percentage of open panels had the greatest effect on recognizer accuracy, while the effect of microphone position was much smaller. The NOISEX babble noise already included reverberation, so each noisy and reverberant test set was generated by convolving the test utterances with the appropriate room impulse response and then adding the babble noise at an SNR of 30, 20, 10, or 0 dB. As in the noisy tests, the SNR was measured on an utterance-by-utterance basis.

The results of these tests are summarized in Table 6.5. The MSG recognizer was

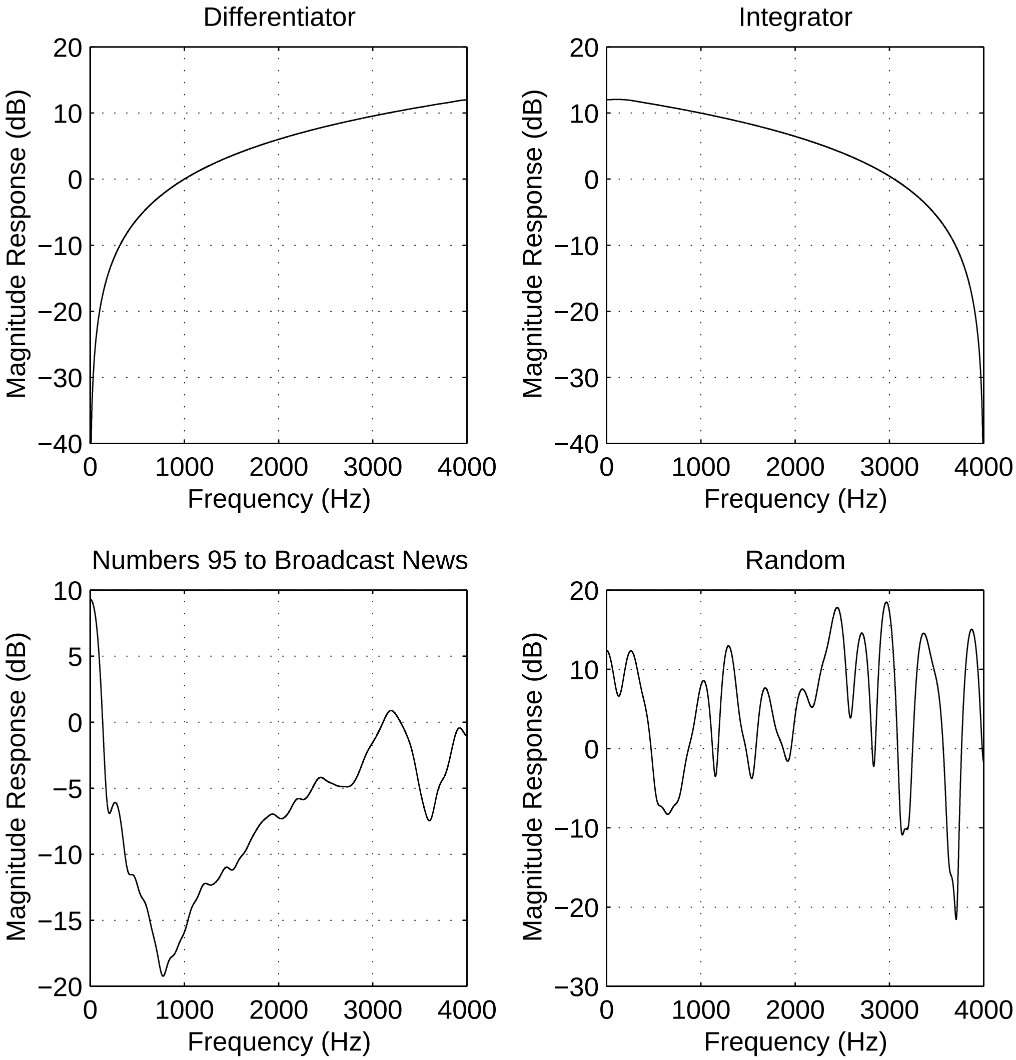


Figure 6.3: Frequency responses for the four filters used to create the spectrally shaped Numbers 95 final test sets.

T_{60} (s)	SNR (dB)	MSG	PLP	combo
0.3	30	9.6%	12.7%	8.6%
	20	12.0%	16.5%	10.3%
	10	24.3%	31.0%	19.9%
	0	69.4%	66.2%	56.0%
0.5	30	17.0%	23.7%	15.2%
	20	19.5%	27.1%	16.4%
	10	35.0%	39.8%	26.4%
	0	79.0%	73.7%	63.4%
0.9	30	45.8%	56.9%	39.3%
	20	47.5%	58.8%	41.7%
	10	63.4%	67.4%	51.4%
	0	95.9%	86.5%	81.5%

Table 6.5: Word error rates for MSG, PLP, and combined recognizers on the noisy and reverberant Numbers 95 final test sets.

significantly more accurate than the PLP recognizer for all three reverberant conditions, provided that the SNR was above 0 dB. The PLP recognizer was significantly better than the MSG recognizer in all three 0 dB conditions, and the combined recognizer was significantly better than either other recognizer in all tests.

6.1.6 Summary

The MSG features are robust to a broad set of acoustic conditions which were not tested during their development, at least for the small-vocabulary Numbers 95 task. The MSG recognizer outperformed the PLP recognizer in almost all tests, and in the tests in which the PLP recognizer was better, the difference in performance between the PLP and MSG recognizers was not statistically significant (except for the tests under noisy reverberant conditions with 0 dB SNR). This general robustness is derived from the temporal signal-processing strategies implemented in the MSG representation that exploit the specific structure of the speech signal. These tests also convincingly illustrate the potential of combining multiple representations in the recognition process. In nearly all of the conditions examined, the combined recognizer, which has the same number of MLP weights as the PLP or MSG recognizer, had the highest accuracy.

6.2 Tests with Broadcast News

Broadcast News [Ste97] is a standard task for the evaluation of large-vocabulary ASR systems. The speech material in the Broadcast News corpus is collected from radio and television news programs broadcast in the United States. These programs include, for example, NPR's *All Things Considered*, the BBC's *The World Today*, and ABC's *World News Tonight*. The corpus contains speech from a diverse set of individuals, including native speakers of American and British English as well as many non-native speakers. The material includes both scripted and conversational speaking styles. The acoustic environments represented in the corpus include high-quality studio recordings, telephone speech and speech in the presence of background noise or music. All of the utterances were recorded at a 16 kHz sampling rate with 16-bit quantization. Recognizers for the Broadcast News task typically have lexicons of 65,000 words.

For the 1998 Broadcast News evaluation, researchers from the Connectionist Speech Recognition group at Cambridge University (UK), Sheffield University (UK), and ICSI collaborated on a recognition system. This system, known as SPRACH (Speech Recognition Algorithms for Connectionist Hybrids), was based on Cambridge University's ABBOT recognition system [CR98b], a hybrid ASR system that uses recurrent neural networks (RNNs) as acoustic models. The RNNs used PLP features as their front-end speech representation.

One of ICSI's contributions to the SPRACH recognizer was an MLP acoustic model whose acoustic likelihoods are combined with those from the RNNs. The MLP was trained on Broadcast News utterances that had been downsampled to 8 kHz because it was expected that this would improve the performance of the SPRACH recognizer on the telephone and bandlimited utterances in Broadcast News. PLP features were considered as a possible front-end representation for the MLP, as were three different versions of the MSG features. 12th-order PLP features, including the zero-order cepstral coefficient, were used. They were computed from 32-ms frames with a 16-ms frame step (to match the frame and step size used with the RNNs), and were normalized in an off-line procedure to have means of zero and variances of one on an utterance-by-utterance basis.

Because the training of the Broadcast News recognizers required a considerable amount of time, it was impractical to perform a large number of experiments to find the best MSG feature set (as had been done in Chapter 5). However, because the design of the MSG features in Chapter 5 was based on a relatively small data set, it seemed prudent to test a small number of different MSG representations on Broadcast News instead of testing only the version that performed best on Numbers. Like the PLP features, the MSG features were also calculated from 32-ms frames with a 16-ms frame step. The initial spectral analysis was performed using a power-spectral filterbank containing fourteen overlapping, triangular filters with bandwidths and spacings of 1 Bark, covering the 160–4000 Hz range. The MSG features were normalized off-line on a per-utterance basis instead of using the on-line normalization of means and variances developed for the Numbers 95 experiments. The three MSG feature sets differed in the envelope filters and AGC time constants used in their computation. Because the experiments in Chapter 5 and the experiments with Broadcast News were performed concurrently, the MSG features tested on Broadcast News represent "snapshots" from different points in the optimization of the MSG features. The versions tested were the following:

- The **MSG1** features were computed using IIR envelope filters with passbands of 0–16 Hz and 2–16 Hz. The lowpass features were processed with two feedback AGC units in series, with the first AGC having a time constant of 160 ms and the second having a time constant of 320 ms. The bandpass features were also processed with two feedback AGC units in series, with the first AGC having a time constant of 160 ms and the second having a time constant of 640 ms. This was the best set of features for Numbers as of the end of the experiments described in Section 5.2.

- The **MSG2** features were computed using FIR envelope filters with passbands of 0–6 Hz and 6–12 Hz. Both the lowpass and bandpass features were processed with two feedback AGC units in series, with the first AGC having a time constant of 160 ms and the second having a time constant of 320 ms. This set of features was the best for Numbers as of the end of the experiments described in Section 5.6.

- The **MSG3** features were identical to the MSG2 features, except that the MSG3 features were computed using FIR envelope filters with passbands of 0–8 Hz and 8–16 Hz. This set of features was developed after experiments on Broadcast News with the MSG2 features indicated that the envelope filters selected in Section 5.6 might be too narrow.

Features	MLP Alone	Combined with RNNs
PLP	36.7%	31.1%
MSG1	39.4%	29.9%
MSG2	43.8%	31.9%
MSG3	41.0%	31.0%

Table 6.6: Word error rates for Broadcast News recognition using PLP features and three different versions of the MSG features, either alone or in combination with a set of four recurrent neural network acoustic models trained on PLP features. More recent experiments [EM98] have shown that the difference in performance between the PLP and MSG1 features on the MLP-only condition decreases as the number of MLP weights and amount of MLP training data increase. Thus, for MLPs with 4000 hidden units trained on 74 hours of data (vs. MLPs with 2000 hidden units trained on 37 hours for the experiments summarized above) the word error rate obtained using PLP features was 33.7% and the word error rate obtained using MSG1 features was 35.3%.

The filters for all three sets of the MSG features were versions of filters used in the Numbers 95 experiments, redesigned for a 62.5 Hz sampling rate.

The four different MLP acoustic models were trained on a phonetic labeling of the Broadcast News data generated via forced alignment with the RNNs.[1] No embedded training of the MLPs was performed due to the limited time available for these experiments. The PLP net had a 117-unit input layer (9 frames of input × 13 features per frame), a 2000-unit hidden layer, and a 54-unit output layer. The MSG nets had 252-unit input layers (9 frames of input × 28 features per frame), 2000-unit hidden layers, and a 54-unit output layers. The four MLPs were tested alone and in combination with acoustic likelihoods from the RNNs. The results of these tests are summarized in Table 6.6.

In the experiments using only the MLP acoustic models, the PLP features yielded the best performance by a significant margin, and the MSG2 features resulted in the worst performance by a significant margin. Using MSG1 features gave better performance than using MSG3 features, but not by a statistically significant margin. While these results seem to indicate that the PLP features are better for the Broadcast News task than the

[1] All of the Broadcast News experiments reported in this thesis were performed by Dan Ellis and Adam Janin. My contributions to the Broadcast News work were a feature-extraction program for generating MSG features and a set of envelope filters designed to work at a sampling rate of 62.5 Hz. I am grateful to Dan and Adam for doing all the hard work in these experiments!

MSG features, it is important to note that the MLP training labels were produced by a PLP-based acoustic model and no embedded training of the MLPs was performed (due to the long training times required). Thus, this test is not an entirely fair one. In combination with the acoustic likelihoods from the RNNs, the MSG1 features resulted in the best performance and the MSG2 features performed the most poorly. In combination with the RNNs, the differences in performance between the MSG1, MSG3, and PLP features were not statistically significant.

The MSG1 features were used in the SPRACH recognizer. The MSG1 features were chosen instead of the PLP features on the basis of the results of the combined recognition experiment. Although the performance obtained using the MSG1 features was not significantly better than that obtained using PLP features (for the size of test set used in these experiments), it was anticipated that using one of the MSG feature sets would give better performance than the PLP features, in combination with the RNNs. Recall that combinations of classifiers or recognizers work best when the models being combined have different error patterns. One way to ensure that the errors are different is to give the models different input representations. It was also expected that embedded training of the acoustic models, which was used in the training of the final SPRACH system for the 1998 Broadcast News evaluation, would improve the performance of the MSG-based acoustic model.

The MSG1 features were selected over the MSG2 and MSG3 features because they gave better performance both alone and in combination with the RNNs. The relatively poor performance of the MSG2 features in both conditions may indicate that the 12–16 Hz modulations, which are filtered out in the computation of the MSG2 features, carry important information for recognition of the phonetically richer Broadcast News data (even if they do not appear to be especially important for recognition of Numbers, as shown in Section 5.6). In both sets of experiments the performance difference between the MSG1 and MSG3 features was not statistically significant for the size of the test set used. The difference was consistent across two experimental conditions, however, so the MSG1 features were chosen. The use of only PLP-derived training labels may be a confounding factor in the comparison of the MSG1 and MSG3 features because the MSG1 features, which are computed with the broadest-bandwidth envelope filters, are the most likely MSG feature set to align well with the PLP-derived labels. The question of which MSG feature set, MSG1 or MSG3, is most useful for large-vocabulary ASR cannot be convincingly resolved without further

	Overall	Degraded Acoustics (F4)	Telephone (F2)
1997	27.2%	37.3%	37.7%
1998 (A)	21.7%	15.5%	32.4%
(B)	20.0%	23.0%	28.4%

Table 6.7: Selected results from the 1997 and 1998 Broadcast News evaluations. The 1997 results [CR98a] are from the ABBOT system, while the 1998 results are from the SPRACH system. Two data sets were used in the 1998 evaluation: one (set A) contained data from 1996 broadcasts, while the other (set B) contained data from 1998 broadcasts. Overall performance is shown, as well as performance on two of the focus conditions–degraded acoustics (F4) and telephone speech (F2).

experimentation.

Selected results from the ABBOT system for the 1997 Broadcast News evaluation and the SPRACH system for the 1998 Broadcast News evaluation are summarized in Table 6.7. Comparison of these results is complicated because there are many differences between the two systems and because the evaluations were performed on different data sets. Nevertheless, these results provide some indication that the inclusion of the MSG-based acoustic model in the SPRACH system contributed to its superior performance. The strongest evidence for this view is the improvement from 1997 to 1998 on the catch-all degraded acoustics condition (F4), which comprise 28% of the data in both 1998 evaluation sets. Average performance for the two 1998 sets on this focus condition was 48% better than the 1997 system. One of the goals in the SPRACH collaboration was to improve the system's performance on the telephone condition (F2) by adding an acoustic model trained on 8-kHz-sampled data. Indeed, an average improvement of 20% was obtained on the telephone speech over the 1997 system; however, this yielded only a small improvement in overall system performance because the telephone data comprised only 3.4% of the data in 1998 evaluation sets.

6.2.1 The Final Version of the MSG Features for Broadcast News

The MSG features used in the SPRACH recognizer (the MSG1 features) differ in certain ways from the best MSG features for Numbers recognition (illustrated in Figure 5.12 and described in Section 5.12). The most important difference is in the envelope filtering,

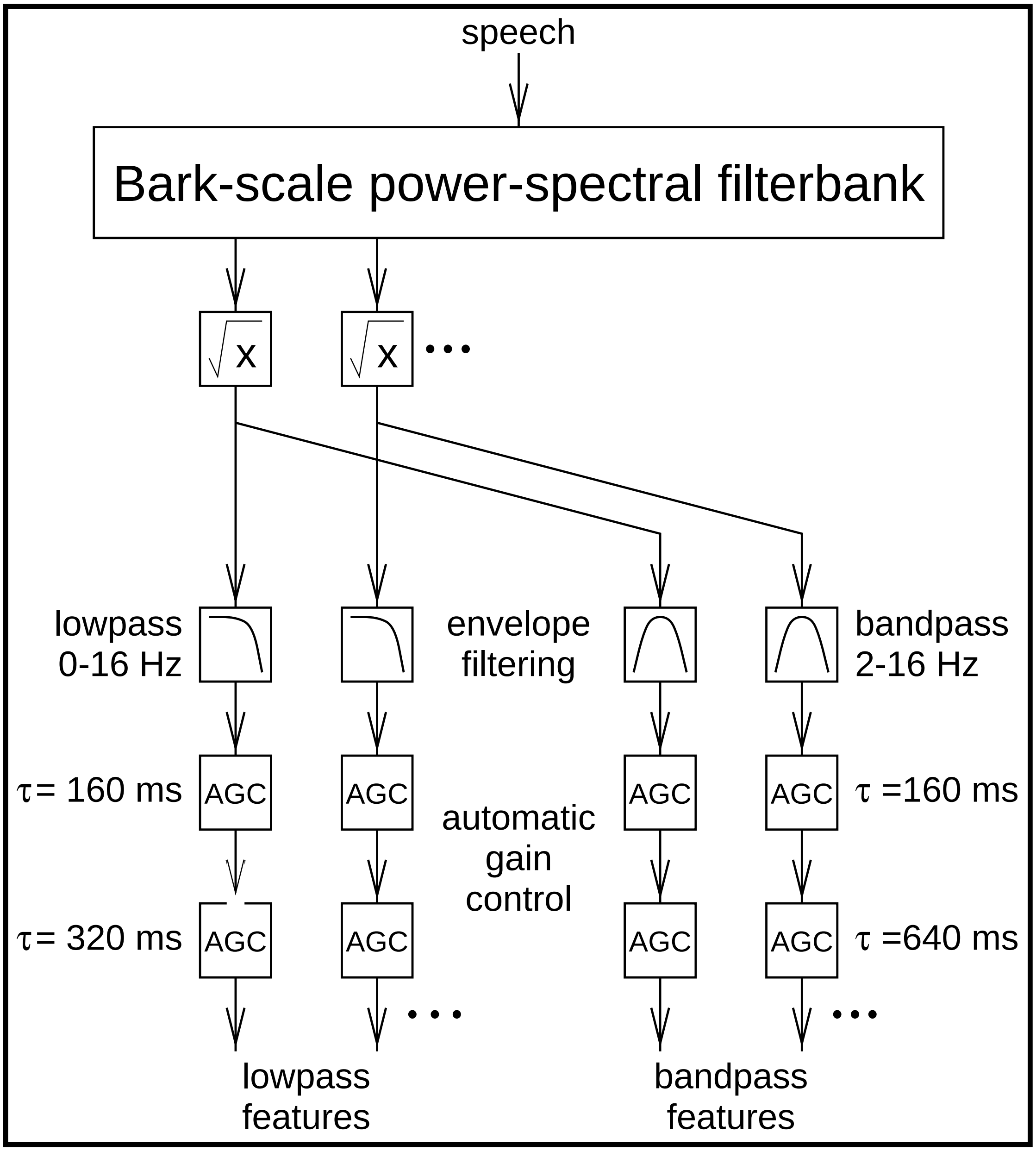


Figure 6.4: Signal processing for the best version of the modulation-filtered spectrogram (MSG) features for Broadcast News recognition. The key difference between these features and the best features for Numbers recognition is the design of the envelope filters. While these features are computed with two envelope filters having passbands of 0–16 and 2–16 Hz, the best features for Numbers recognition are computed with two envelope filters having passbands of 0–8 and 8–16 Hz. It is likely that the AGC computation could be made completely uniform by setting the time constant of the second AGC in the bandpass feature processing to 320 ms with little, if any, effect on recognition accuracy.

where much broader filters are used to compute the features for Broadcast News. The best features for Broadcast News are computed via the following steps (illustrated in Figure 6.4):

1. The speech signal is segmented into 32-ms frames with a 16-ms frame step, each frame is multiplied by a Hamming window, and the power spectrum for each frame is computed. The greater frame length and step were chosen for compatibility with the RNN acoustic models (which used 32-ms frames and a 16-ms frame step).

2. The power spectrum is accumulated into critical-band-like frequency channels via convolution with a bank of fourteen overlapping, triangular filters that have bandwidths and spacings of 1.0 Bark and cover the 160–4000-Hz range. The critical-band-like power spectrum is converted into an amplitude spectrum by taking the square root of the filterbank output. In the computation of MSG features for Numbers, bandwidths and spacings of 0.95 Bark were used to generate an even number of frequency bands (simplifying implementation of the spectral smoothing of the bandpass features) that span a slightly smaller range of frequencies. Because the Numbers data was collected over the telephone, it was expected that there would be little useful information below ca. 300 Hz. In contrast, much of the Broadcast News data came from studio recordings which were expected to contain useful information down to somewhat lower frequencies.

3. The critical-band-like amplitude signals are filtered by two IIR filters in parallel: a lowpass filter with a 0–16 Hz passband and and a bandpass filter with an 2–16 Hz passband. This is the most important difference between the Numbers features and the Broadcast News features. The broader filters were used for Broadcast News because they gave better recognition performance than the narrower filters used in the final Numbers experiments. It is not clear if this discrepancy is due to the inherent differences between the two recognition tasks, or if it is due to the broader filters' producing features that are better aligned with the PLP-derived training labels used in the Broadcast News experiments.

4. Both the lowpass and bandpass streams are processed through two feedback AGC units (illustrated in Figure 5.5). In the lowpass stream the first AGC has a time constant of 160 ms and the second has a time constant of 320 ms, while in the bandpass stream the first AGC has a time constant of 160 ms and the second has a time constant

of 640 ms. It is likely that the AGC processing of the bandpass features could be made identical to the processing of the lowpass features, but this possibility was not explored due to limitations of time.

No on-line normalization of the feature means and variances was performed for the Broadcast News features because the SPRACH recognition system included an off-line, per-utterance normalization step.

Chapter 7

Conclusions

This thesis began with the premise that the robustness of ASR systems to acoustic interference in general, and room reverberation in particular, could be improved by including signal-processing and information-processing strategies that are employed in human auditory processing in an ASR system. The strategies examined included

- critical-band-like frequency resolution,
- the emphasis of slow changes in the spectral structure of the speech signal in the representation,
- automatic gain control,
- the integration of phonetic information over syllabic durations, and
- the use of multiple signal representations in the recognition process.

A simple signal-processing system that included the first three of these ideas (the signal-processing strategies) was developed and tuned to produce visual displays of the speech signal in a spectrographic format that were stable in the presence of additive noise and room reverberation. This representation, the modulation-filtered spectrogram (MSG), was then optimized for use as an ASR front end.

The optimization was performed in a series of experiments in which different parts of the signal processing were systematically varied, and the performance of ASR systems using the different variant front ends was measured under clean and reverberant conditions.

Changes to the signal processing that yielded performance improvements were retained in later experiments. A relatively small recognition task, the recognition of continuous numbers spoken over the telephone, was used to make it feasible to perform many recognition experiments, the vast majority of which included the embedded training of the MLP acoustic model.

This work concluded with two sets of recognition tests designed to determine whether or not the MSG features are useful under a wide variety of acoustic conditions and on large-vocabulary recognition tasks. The first tests measured the performance of three different recognition systems on a collection of final test utterances from the Numbers corpus under many different acoustic conditions: clean, additive noise, reverberation, spectral shaping, as well as simultaneous noise and reverberation. The three test systems were a system using only MSG features, a system using only PLP features, and a system using MSG and PLP features in combination (with the combination performed by averaging acoustic log likelihoods from separate MSG-based and PLP-based acoustic models on a frame-by-frame basis). The three test systems had identical numbers of MLP weights, used the same language model, and each had a lexicon optimized using the same embedded training process. These final Numbers tests had two clear results. First, in comparison to the PLP recognizer, the MSG recognizer was more robust in almost all of the acoustic conditions examined. Second, the combined recognizer was more robust than the MSG recognizer in almost all of the acoustic conditions tested.

The second set of tests compared the performance of the PLP features and three different versions of the MSG features on the large-vocabulary Broadcast News corpus. The features were tested alone and in combination with acoustic likelihoods from a recurrent neural network acoustic model developed by the Connectionist Speech Recognition group at Cambridge University (UK). In the tests in which the PLP or MSG features were used on their own, the PLP-based recognizer outperformed the MSG recognizers by a significant margin. It is not clear, however, if this difference in performance is due to differences between the two representations *per se*, or whether this result is attributable to mismatches between the MSG features and the training labels (which were generated via forced alignment with a recognizer trained only on PLP features). On the combined tests, however, one of the MSG features sets gave better performance than PLP, although not by a statistically significant margin.

In purely practical terms, this research can be considered successful because:

- Using the MSG features reduced the error rate on the moderately reverberant condition by as much as 30% (versus the baseline PLP system). Combining the MSG and PLP features reduced the error rate on the moderately reverberant condition by as much as 42%.

- The MSG recognizer and the combined PLP and MSG recognizer both performed well under a wide variety of acoustic conditions, including conditions which were not tested during the development of the MSG features.

- MSG features are useful not only for the restricted Numbers task on which they were developed, but also for more general, large-vocabulary tasks such as Broadcast News (if at least some of the difference between the performance with only PLP features and only MSG features can be attributed to misalignment between the training labels and MSG features). The MSG features, in combination with PLP features, appear to be particularly useful for improving recognizer performance in degraded acoustic conditions.

A number of broader lessons may be drawn from this work as well:

- **Signal processing strategies that exploit the specific temporal properties of speech are an effective means for improving ASR robustness.** The envelope filtering and AGC processing considered in this work are both examples of such strategies. The early experiments, in which different steps in the MSG processing were omitted, clearly demonstrated that envelope filtering is crucial for good recognizer performance in reverberation. Later experiments with different envelope filters showed that recognizer performance on both clean and reverberant tests was nearly optimal when the filters passed only the 0–16 Hz modulation frequencies. Experiments comparing off-line with on-line AGC processing showed that better performance could be obtained with the on-line processing, which reduced the effects of both constant and slowly varying gains, than with the off-line processing, which eliminated only constant gain factors. These strategies proved effective not only under reverberant conditions, but also under other conditions such as additive noise and unknown spectral shaping. Moreover, they were useful not only for the small-vocabulary Numbers task, but also for improving the performance of a large-vocabulary recognizer under degraded acoustic conditions.

- **Additional experience with the application of temporal processing techniques to ASR and better theoretical understanding of these techniques are needed for temporal processing to become a standard approach to robust ASR.** The various experiments with different AGC designs most clearly demonstrate this need. There is very little information available to guide the design of such AGCs. Not enough is known about the effects of an auditory AGC on speech perception or about the encoding of phonetic information in the speech signal to aid in the selection of different AGC designs or in the setting of their parameters. Thus, the selection of an AGC design and the setting of its parameters is currently done empirically. An empirical search can require many time-consuming recognition experiments, and there is no guarantee that a design which is useful for a given task will also be useful for another.

  The design of envelope filters for ASR front ends has a somewhat firmer foundation because numerous perceptual studies and several ASR studies (including this one) have shown the importance of modulations at frequencies below 16 Hz. It is not as clear, however, whether or how the 0–16 Hz modulation frequency range should be divided, or how very slow modulations should be treated. These issues must also, at present, be decided empirically, and there is no guarantee that results from one task will generalize to another.

  The analyses of Nadeu and his colleagues [NJ94, NPLJ97] showing that highpass envelope filtering may be understood as an equalization of energy in the modulation frequency domain and that lowpass envelope filtering suppresses modulation frequencies that are not accurately characterized by front-end signal processing are a promising step towards a better understanding of temporal processing.

- **Including both modulation filtering (in the amplitude domain) and AGC processing as separate steps in the front-end signal processing leads to better recognition performance.** The bandpass filtering of log power spectral trajectories performed by log-RASTA-PLP is a form of AGC, so log-RASTA-PLP compensates for unknown spectral shaping of the speech signal. J-RASTA-PLP jointly compensates for additive noise and spectral shaping by performing bandpass filtering in a linear-logarithmic power spectral domain, with the value of the J parameter controlling the tradeoff between compensation for additive noise and compensation for spectral shaping. The MSG signal processing performs both modulation filtering and AGC, but unlike J-RASTA-PLP processing, the modulation filtering and AGC are separate steps. This separation permits the independent optimization of the filtering and AGC steps, which leads to better recognition performance.

- **Combining recognition systems is a powerful and general method for improving ASR robustness.** In the final Numbers 95 tests, the combined MSG and PLP recognizer outperformed the recognizers using only a single representation in nearly all conditions, and frequently by a statistically significant margin. In the Broadcast News tests, combination of the MSG and PLP representations also led to more accurate recognition, especially under degraded acoustic conditions. These results were obtained using only two signal representations and a simple combination technique: unweighted averaging of acoustic log likelihoods on a frame-by-frame basis. It is likely that even better performance could be obtained using additional signal representations and more sophisticated and adaptive combination methods. Recent studies of human speech perception indicate that the accuracy of human speech recognition may rely on the adaptive integration of information across different spectral regions [AG98, GAS98].

As a practical technology, automatic speech recognition is still in its infancy. It is currently useful for restricted tasks such as document dictation in quiet offices and limited interactions over telephone lines. Comparison with human capabilities suggest that considerable progress on many fronts must still be made before the ideal of natural, unrestricted verbal interaction with computers is reached. By showing how signal-processing and information-processing strategies based on human auditory processing can be applied to make automatic recognizers more robust to reverberation and other forms of acoustic interference, this thesis brings this ideal slightly closer to reality.

Bibliography

[AB79] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[ABI$^+$95] Krste Asanović, James Beck, Bertrand Irissou, Brian E. D. Kingsbury, Nelson Morgan, and John Wawrzynek. The T0 vector microprocessor. In *Proceedings of Hot Chips VII*, 1995.

[AG98] Takayuki Arai and Steven Greenberg. Speech intelligibility in the presence of cross-channel spectral asynchrony. In *ICASSP 98. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 933–936. IEEE, 1998.

[AH96] Carlos Avendaño and Hynek Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 889–892, 1996.

[AKB$^+$96] Krste Asanović, Brian E. D. Kingsbury, James Beck, Bertrand Irissou, and John Wawrzynek. T0: A single-chip vector microprocessor with reconfigurable pipelines. In H. Grünbacher, editor, *Proceedings of the 22nd European Solid-State Circuits Conference*, pages 344–347, 1996.

[All94] Jont B. Allen. How do humans process and recognize speech? *Journal of the Acoustical Society of America*, 2(4):567–577, October 1994.

[APHA96] Takayuki Arai, Misha Pavel, Hynek Hermansky, and Carlos Avendaño. Intelligibility of speech with filtered time trajectories of spectral envelopes. In

ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing, pages 2490–2493, 1996.

[ASNS89] Y. Ariki, Mizuta S, M. Nagata, and T. Sakai. Spoken-word recognition using dynamic features analyzed by two-dimensional cepstrum. *IEE Proceedings*, 136 I(2):133–140, April 1989.

[Ata74] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, June 1974.

[Ave97a] Carlos Avendaño, 1997. personal communication.

[Ave97b] Carlos Avendaño. *Temporal Processing of Speech in a Time-Feature Space.* PhD thesis, Oregon Graduate Institute, 1997.

[AvVH96] Carlos Avendaño, Sarel van Vuuren, and Hynek Hermansky. Data based filter design for RASTA-like channel normalization in ASR. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 2087–2090, 1996.

[BBdSM86] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP 86. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 49–52. IEEE, 1986.

[BD96] Hervé Bourlard and Stéphane Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 426–429, 1996.

[BM94] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[Bol79] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(2):113–120, April 1979.

[Bri90] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. F. Soulié and J. Hérault, editors, *Neurocomputing, Algorithms, Architectures and Applications*, volume F 68 of *NATO ASI Series*, pages 227–236. Springer-Verlag, 1990.

[BW89] H. Bourlard and C. J. Wellekens. Links between Markov models and multilayer perceptrons. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 1. Proceedings of the 1988 Conference.*, pages 502–510, San Mateo, CA, USA, 1989. Morgan Kaufmann.

[Cal83] William H. Calvin. A stone's throw and it launch window: Timing precision and its implications for language and hominid brains. *Journal of Theoretical Biology*, 104(1):121–135, September 1983.

[CCC$^+$96] G. D. Cook, J. D. Christie, P. R. Clarkson, M. M. Hochberg, B. T. Logan, A. J. Robinson, and C. W. Seymour. Real-time recognition of broadcast radio speech. In *ICASSP 96. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 141–144. IEEE, 1996.

[CEST77] Walter L. Cullinan, Elaine Erdos, Ronald Schaefer, and Mary Ellen Tekieli. Perception of temporal order of vowels and consonant-vowel syllables. *Journal of Speech and Hearing Research*, 20(4):742–751, December 1977.

[CG91] Kenneth W. Church and William A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5(1):19–54, January 1991.

[CGC94] Martin Cooke, Phil Green, and Malcom Crawford. Handling missing data in speech recognition. In *ICSLP 94. Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 1555–1558, 1994.

[CKA94] J. Cohen, T. Kamm, and A. Andreou. An experiment in systematic speaker variability. In *Final Day Review, DOD Speech Workshop on Robust Speech Recognition*, 1994.

[CNLD95] R. A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. In *Proceedings of Eurospeech 1995*, pages 821–824, 1995.

[CR98a] G. D. Cook and A. J. Robinson. The 1997 Abbot system for the transcription of broadcast news. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[CR98b] Gary Cook and Tony Robinson. Transcribing Broadcast News with the 1997 ABBOT system. In *ICASSP 98. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 917–920. IEEE, 1998.

[CW94] Magdalene H. Chalikia and Richard M. Warren. Spectral fissioning in phonemic transformations. *Perception and Psychophysics*, 55(2):218–226, February 1994.

[DE88] E. A. DeYoe and D. C. Van Essen. Concurrent processing streams in monkey visual cortex. *Trends in Neuroscience*, 11:219–226, 1988.

[DFP94] Rob Drullman, Joost M. Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, February 1994.

[DHC99] B. Delgutte, B. M. Hammond, and P. A. Cariani. Neural coding of the temporal envelope of speech. In S. Greenberg and W. A. Ainsworth, editors, *Listening to Speech: An Auditory Perspective.* Oxford University Press, New York, 1999.

[DM80] Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366, August 1980.

[Dud39] Homer Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11(2):169–177, October 1939.

[EB82] Kjell Elenius and Mats Blomberg. Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system.

In *ICASSP 82. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 535–538. IEEE, 1982.

[EM98] Dan Ellis and Nelson Morgan. Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. Submitted to the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1998.

[FAF+77] A. J. Fourcin, W. A. Ainsworth, G. C. M. Fant, H. Fujisaki, W. J. Hess, J. N. Holmes, F. Itakura, M. R. Schroeder, and H. W. Strube. Speech processing by man and machine — group report. In Theodore H. Bullock, editor, *Recognition of Complex Acoustic Signals: Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals, Berlin 1976, September 27 to October 2*, number 5 in Life Sciences Research Reports, pages 307–351, Berlin, Germany, 1977. Dahlem Konferenzen.

[FE91] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, January/February 1991.

[Fle40] Harvey Fletcher. Auditory patterns. *Review of Modern Physics*, 12:47–65, 1940.

[Fur81] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, April 1981.

[Fur86a] Sadaoki Furui. On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America*, 80(4):1016–1025, October 1986.

[Fur86b] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

[GAS98] Steven Greenberg, Takayuki Arai, and Rosaria Silipo. Speech intelligibility derived from exceedingly sparse spectral information. In *ICSLP 98. Proceed-*

ings of the 1998 International Conference on Spoken Language Processing, 1998.

[GHE96] Steven Greenberg, Joy Hollenback, and Dan Ellis. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages S24–S27, 1996.

[Ghi86] Oded Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech and Language*, 1(2):109–130, December 1986.

[GK97] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP 97. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1647–1650. IEEE, 1997.

[GL94] Jean-Luc Gauvain and Chin-Hui Lee. Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–299, April 1994.

[GO95] Bernhard H. Gaese and Joachim Ostwald. Temporal coding of amplitude and frequency modulation in the rat auditory cortex. *European Journal of Neuroscience*, 7(3):438–450, March 1995.

[Gon95] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291, April 1995.

[GOS96] D. Giuliani, M. Omologo, and P. Svaizer. Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 1329–1332, 1996.

[Gre61] Donald D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustical Society of America*, 33(10):1344–1356, October 1961.

[Gre90] Donald D. Greenwood. A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America*, 87(6):2592–2605, June 1990.

[Gre95] Steven Greenberg. personal communication, 1995.

[GY92] M. J. F. Gales and S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *ICASSP 92. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 233–236. IEEE, 1992.

[HA90] Brian A. Hanson and Ted H. Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech. In *ICASSP 90. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 857–860. IEEE, 1990.

[HAvVT97] Hynek Hermansky, Carlos Avendaño, Sarel van Vuuren, and Sangita Tibrewala. Recent advances in addressing sources of non-linguistic information. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 103–106. ESCA, 1997.

[Hei89] Walter Heiligenberg. Coding and processing of electrosensory information in gymnotiform fish. *Journal of Experimental Biology*, 146:255–275, September 1989.

[Her90] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[Her97] Hynek Hermansky. Should recognizers have ears? In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 1–10. ESCA, 1997.

[HHP88] Eva B. Holmberg, Robert E. Hillman, and Joseph S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America*, 84(2):511–529, 1988.

[Hir88] H. G. Hirsch. Automatic speech recognition in rooms. In J. L. Lacoume, A. Chehikian, N. Martin, and J. Malbos, editors, *Signal Processing IV: Theories and Applications. Proceedings of EUSIPCO-88. Fourth European Signal Processing Conference*, volume 3, pages 1177–1180. Elsevier Science Publishers, B.V., 1988.

[Hir92] H. G. Hirsch. Robust speech recognition in noisy and reverberant environments. In P. Laface and R. De Mori, editors, *Speech Recognition and Understanding. Recent Advances, Trends and Applications. Proceedings of the NATO Advanced Study Institute.*, volume F 75 of *NATO ASI Series*, pages 101–106, Berlin, Germany, 1992. Springer-Verlag.

[HM94] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[HMBK91] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proceedings of Eurospeech 1991*, pages 1367–1370, 1991.

[HMR91] H. G. Hirsch, P. Meyer, and H. W. Ruehl. Improved speech recognition using high-pass filtering of subband envelopes. In *Proceedings of Eurospeech 1991*, pages 413–146, 1991.

[HS72] T. Houtgast and H. J. M. Steeneken. Envelope spectrum and intelligibility of speech in enclosures. In *Proceedings of the IEEE Conference on Speech Communication and Processing*, pages 392–395, 1972.

[HS73] T. Houtgast and H. J. M. Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28(1):66–73, January 1973.

[HTP96] Hynek Hermansky, Sangita Tibrewala, and Misha Pavel. Towards ASR on partially corrupted speech. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 462–465, 1996.

[Hug75] A. W. F. Huggins. Temporally segmented speech. *Perception and Psychophysics*, 18(2):149–157, 1975.

[HWA95] Hynek Hermansky, Eric A. Wan, and Carlos Avendaño. Speech enhancement based on temporal processing. In *ICASSP 95. Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, pages 405–408. IEEE, 1995.

[HWHK65] Arthur S. House, Carl E. Williams, Michael H. L. Hecker, and K. D. Kryter. Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37(1):158–166, January 1965.

[Jor95] Michael I. Jordan. Why the logistic function? A tutorial discussion on probabilities and neural networks. Computational Cognitive Science 9503, Massachusetts Institute of Technology, 1995.

[KAHP97] Noboru Kanadera, Takayuki Arai, Hynek Hermansky, and Misha Pavel. On the importance of various modulation frequencies for speech recognition. In *Proceedings of Eurospeech 1997*, pages 1079–1082, 1997.

[KBM96] Yochai Konig, Hervé Bourlard, and Nelson Morgan. REMAP: Recursive estimation and maximization of a posteriori probabilities-application to transition-based connectionist speech recognition. In D. S. Touretzky, M. C. Moser, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference.*, pages 388–394, Cambridge, MA, USA, 1996. MIT Press.

[Kla82] Dennis H. Klatt. Speech processing strategies based on auditory models. In Rolf Carlson and Björn Granström, editors, *The Representation of Speech in the Peripheral Auditory System*, pages 181–196. Elsevier Biomedical Press, Amsterdam, The Netherlands, 1982.

[KM97] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP-97. Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1259–1262. IEEE, 1997.

[KMG97] Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Improving ASR performance for reverberant speech. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 87–90. ESCA, 1997.

[KMG98] Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1–3):117–132, August 1998.

[KPA92] Armin Kohlrausch, Dirk Püschel, and Henning Alphei. Temporal resolution and modulation analysis in models of the auditory system. In M. E. H. Schouten, editor, *The Auditory Processing of Speech: From Sounds to Words*, pages 85–98. Walter de Gruyter and Co., Berlin, Germany, 1992.

[LC95] Qiguang Lin and Chiwei Che. Normalizing the vocal tract length for speaker independent speech recognition. *IEEE Signal Processing Letters*, 2(11):201–203, November 1995.

[Lib82] M. C. Liberman. The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. *Journal of the Acoustical Society of America*, 72(5):1441–1449, November 1982.

[Lip97] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, July 1997.

[Lom11] Etienne Lombard. Le signe de l'élévation de la voix. *Ann. Mal. Oreil. Larynx*, 37:101–119, 1911. Cited in [LT71].

[LS82] Thomas Langhans and Hans Werner Strube. Speech enhancement by non-linear multiband envelope filtering. In *ICASSP 82. Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, pages 156–159. IEEE, 1982.

[LT71] Harlan Lane and Bernard Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14:677–709, 1971.

[Lyo82] Richard F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *ICASSP 82. Proceedings of the International Con-*

ference on Acoustics, Speech, and Signal Processing, pages 1282–1285. IEEE, 1982.

[Mar76] D. Marr. Early processing of visual information. *Philosophical Transactions of the Royal Society of London*, B275(942):483–519, 1976.

[Mas72] Dominic W. Massaro. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79(2):124–145, March 1972.

[Mas74] Dominic W. Massaro. Perceptual units in speech recognition. *Journal of Experimental Psychology*, 102(2):199–208, February 1974.

[MCG98] Andrew C. Morris, Martin P. Cooke, and Phil D. Green. Some solutions to the missing feature problem in data classification, with applications to noise robust ASR. In *ICASSP 98. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 737–740. IEEE, 1998.

[MD67] John P. Moncur and Donald Dirks. Binaural and monaural speech intelligibility in reverberation. *Journal of Speech and Hearing Research*, 10(2):186–195, June 1967.

[MG83] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, September 1983.

[MH80] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London*, B207(1167):187–217, 1980.

[MH92] Nelson Morgan and Hynek Hermansky. RASTA extensions: Robustness to additive and convolutional noise. In *Proceedings of the ESCA Workshop on Speech Processing in Adverse Environments*, pages 115–118, 1992.

[Mil96] Ben Milner. Inclusion of temporal information into features for speech recognition. In *ICSLP 96. Proceedings of the 1996 International Conference on Spoken Language Processing*, pages 256–259, 1996.

[Mir98] Naghmeh Nikki Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, University of California, Berkeley, CA, 1998.

[ML50] George A. Miller and J. C. R. Licklider. The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, 22(2):167–173, March 1950.

[Mou85] J. Mourjopoulos. On the variation and invertibility of room impulse response functions. *Journal of Sound and Vibration*, 102(2):217–228, September 1985.

[MPC97] Dennis R. Morgan, Vipul N. Parikh, and Cecil H. Coker. Automated evaluation of acoustic talker direction finder algorithms in the varechoic chamber. *Journal of the Acoustical Society of America*, 102(5):2786–2792, November 1997.

[MSE93] Andrew Morris, Jean-Luc Schwartz, and Pierre Escudier. An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram. *Computer Speech and Language*, 7(2):121–136, April 1993.

[NA79] Stephen T. Neely and Jont B. Allen. Invertibility of a room impulse response. *Journal of the Acoustical Society of America*, 66(1):165–169, July 1979.

[ND84] Anna K. Nábělek and Amy M. Donahue. Perception of consonants in reverberation by native and non-native listeners. *Journal of the Acoustical Society of America*, 75(2):632–634, February 1984.

[NJ94] Climent Nadeu and Biing-Hwang Juang. Filtering of spectral parameters for speech recognition. In *ICSLP 94. Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 1927–1930, 1994.

[NP74] Anna K. Nábělek and J. M. Pickett. Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing impaired listeners. *Journal of Speech and Hearing Research*, 17(4):724–739, December 1974.

[NPLJ97] Climent Nadeu, Pau Pachès-Leal, and Biing-Hwang Juang. Filtering the time sequences of spectral parameters for speech recognition. *Speech Communication*, 22(4):315–332, September 1997.

[NR82] Anna K. Nábělek and Pauline K. Robinson. Monaural and binaural speech perception in reverberation for listeners of various ages. *Journal of the Acoustical Society of America*, 71(5):1242–1248, May 1982.

[OS75] Yosiro Oono and Yasumasa Sujaku. A model for automatic gain control observed in the firing of primary auditory neurons. *Abstracts of IECE Transactions*, 58(6):61–62, June 1975.

[OS89] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Signal Processing. Prentice Hall, Englewood Cliffs, NJ, USA, 1989.

[OSTGS68] Alan V. Oppenheim, Ronald W. Schafer, and Jr. Thomas G. Stockham. Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56(8):1264–1291, August 1968.

[Pat76] Roy D. Patterson. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59(3):640–654, March 1976.

[PH94] Misha Pavel and Hynek Hermansky. Temporal maksing in automatic speech recognition. *Journal of the Acoustical Society of America*, 95(2):2876, May 1994.

[RB94] Stuart Rosen and Richard J. Baker. Characterising auditory filter nonlinearity. *Hearing Research*, 73(2):231–243, March 1994.

[RL91] M. D. Richard and R. P. Lippmann. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3(4):461–483, Winter 1991.

[Sab22] Wallace Clement Sabine. *Collected Papers on Acoustics.* Harvard University Press, Cambridge, MA, USA, 1922.

[San94] Sumeet Sandhu. A comparative study of mel cepstra and EIH for phone classification under adverse conditions. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1994.

[SBMK93] Caroline L. Smith, Catherine P. Browman, Richard S. McGowan, and Bruce Kay. Extracting dynamic parameters from speech movement data. *Journal of the Acoustical Society of America*, 93(3):1580–1588, March 1993.

[Sch81] M. R. Schroeder. Modulation transfer functions: Definition and measurement. *Acustica*, 49(3):179–182, November 1981.

[Sch85] Richard Schulman. Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America*, 85(1):295–312, 1985.

[Sch89] Martin F. Schlang. An auditory-based approach for echo compensation with modulation filtering. In *Proceedings of Eurospeech 1989*, pages 661–664, 1989.

[SG95] Sumeet Sandhu and Oded Ghitza. A comparative study of mel cepstra and EIH for phone classification under adverse conditions. In *ICASSP 95. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 409–412. IEEE, 1995.

[SH80] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 67(1):318–326, January 1980.

[SH82] H. J. M. Steeneken and T. Houtgast. Evaluation of a physical method for estimating speech intelligibility in auditoria. In *ICASSP 82. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1452–1454. IEEE, 1982.

[Ste97] R. M. Stern. Specification of the 1996 Hub 4 Broadcast News evaluation. In *DARPA Speech Recognition Workshop*, pages 7–10, 1997.

[SU88] Christoph E. Schreiner and John V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hearing Research*, 32(1):49–63, January 1988.

[TH97] Sangita Tibrewala and Hynek Hermansky. Multi-band and adaptation approaches to robust speech recognition. In *Proceedings of Eurospeech 1997*, pages 2619–2622, 1997.

[THCG70] Ian B. Thomas, Peter B. Hill, Francis S. Carroll, and Bienvenido Garcia. Temporal order in the perception of vowels. *Journal of the Acoustical Society of America*, 48(4):1010–1013, October 1970.

[Tho97] David L. Thomson. Ten case studies of the effect of field conditions on speech recognition systems. In S. Furui, B.-H. Juang, and W. Chou, editors, *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 511–518, New York, NY, USA, 1997. IEEE, IEEE.

[tKFP92] Mariken ter Keurs, Joost M. Festen, and Reinier Plomp. Effect of spectral envelope smearing on speech reception. I. *Journal of the Acoustical Society of America*, 91(5):2872–2880, May 1992.

[TMK84] T. Takahashi, A. Moiseff, and M. Konishi. Time and intensity cues are processed independently in the auditory system of the owl. *The Journal of Neuroscience*, 4(7):1781–1786, July 1984.

[Tod94] Neil P. McAngus Todd. The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research*, 23:25–70, 1994.

[vDAP87] J. N. van Dijkhuizen, P. C. Anema, and R. Plomp. The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences. *Journal of the Acoustical Society of America*, 81(2):465–469, February 1987.

[VS93] Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, July 1993.

[WAK$^+$95] John Wawrzynek, Krste Asanović, Brian E. D. Kingsbury, James Beck, David Johnson, and Nelson Morgan. SPERT-II: A vector microprocessor system and its application to large problems in backpropagation training. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 619–625, Cambridge, MA, USA, 1995. The MIT Press.

[WAK$^+$96] John Wawrzynek, Krste Asanović, Brian Kingsbury, David Johnson, James Beck, and Nelson Morgan. SPERT-II: A vector microprocessor system. *IEEE Computer*, 29(3):79–86, March 1996.

[Wat91] Anthony J. Watkins. Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, 90(6):2942–2955, December 1991.

[WEKM94] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald. The new varechoic chamber at AT&T bell labs. In *Proceedings of the Wallace Clement Sabine Centennial Symposium*, pages 343–346, Woodbury, NY, USA, 1994. Acoustical Society of America. Cited in [MPC97].

[WHC96] Richard M. Warren, Eric W. Healy, and Magdalene H. Chalikia. The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *Journal of the Acoustical Society of America*, 100(4):2452–2461, October 1996.

[WKMG98a] Su-Lin Wu, Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *ICASSP 98. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 721–724. IEEE, 1998.

[WKMG98b] Su-Lin Wu, Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. In *ICSLP 98. Proceedings of the 1998 International Conference on Spoken Language Processing*, 1998.

[Wu98] Su-Lin Wu. *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*. PhD thesis, University of California, Berkeley, CA, 1998.

[ZFS57] E. Zwicker, G. Flottorp, and S. S. Stevens. Critical band width in loudness summation. *Journal of the Acoustical Society of America*, 29(5):548–557, May 1957.

[Zwi61] E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248, February 1961.