A Model of Dynamic Auditory Perception and Its Application to Robust Word Recognition

Brian Strope and Abeer Alwan, *Member, IEEE*

Abstract— **This paper describes two mechanisms that augment the common automatic speech recognition (ASR) front end and provide adaptation and isolation of local spectral peaks. A dynamic model consisting of a linear filterbank with a novel additive logarithmic adaptation stage after each filter output is proposed. An extensive series of perceptual forward masking experiments, together with previously reported forward masking data, determine the model's dynamic parameters. Once parameterized, the simple exponential dynamic mechanism predicts the nature of forward masking data from several studies across wide ranging frequencies, input levels, and probe delay times. An initial evaluation of the dynamic model together with a local peak isolation mechanism as a front end for dynamic time warp (DTW) and hidden Markov model (HMM) word recognition systems shows an improvement in robustness to background noise when compared to Mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), and relative spectra (RASTA) based front ends.**

Index Terms—**Dynamic auditory perception, forward masking, robust speech recognition.**

I. Introduction

MOST MODERN automatic speech recognition (ASR) systems model speech as a nonstationary stochastic process by statistically characterizing a sequence of spectral estimations [1]. The common technique for spectral estimation includes an approximation of auditory filtering, a compressive nonlinearity (usually the logarithm), and decorrelation of the spectral estimation through an approximate Karhunen–Loève (KL) transform (the discrete cosine transform). These steps represent rough approximations of the most fundamental aspects of auditory processing: frequency selectivity and magnitude compression. In the last five to ten years, the frequency selectivity for ASR front-ends has migrated from a linear to a perceptually based frequency scale [2]. This progress, toward a better auditory model for ASR, has improved robustness [3].

A large discrepancy remains between current auditory models and the approximations used in ASR front ends. Recent efforts to incorporate more sophisticated auditory models with ASR systems, however, have shown little to no improvements, typically at a severe increase in computational costs [3]. The challenges are to determine what auditory functionality missing from the current front end would be useful for improving recognition robustness and to design efficient mechanisms which reproduce that functionality.

This paper focuses on two aspects of audition not included in current representations: short-term adaptation and sensitivity to the frequency position of local spectral peaks. For each, a mechanism with low computational complexity is described, which adds to the common front end and provides a representation that is more robust to background noise. The dynamic mechanism is parameterized by psychophysical data described here and in the literature [4]. The peak isolation mechanism is a simple modification of a previous cepstral liftering technique [5]. Emphasizing dynamic local peaks is shown to be more robust than emphasizing either dynamics or local peaks.

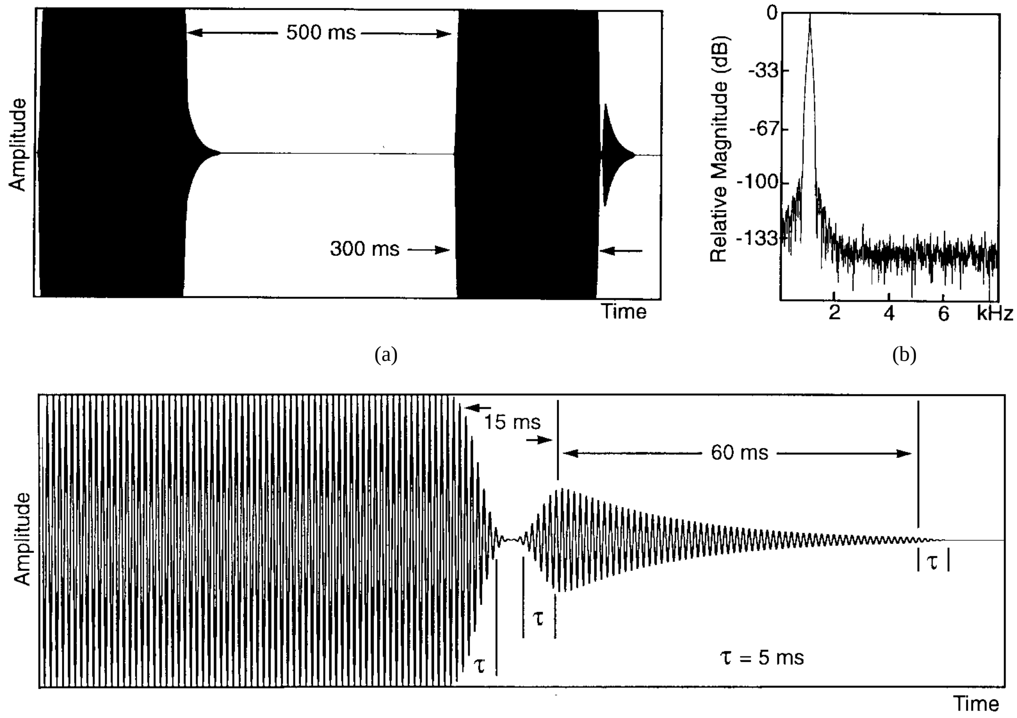
To incorporate a dynamic mechanism within a front end, a method of quantifying auditory adaptation must first be identified. There is considerable physiological and psychophysical evidence of dynamic audition. Short-term adaptation, usually defined as a decreasing response after the onset of a constant stimulus, has been measured in individual auditory nerve firings [6]. The neural response to a stimulus is also reduced during the recovery period following adaptation to a prior stimulus [7]. Here the general term *adaptation* is used for both dynamic processes (short-term adaptation and post-adaptation recovery), and its direction is explicitly specified when significant. *Attack* refers to the decreasing response following stimulus onset, while *release* and *recovery* both refer to the increasing response following stimulus offset. Motility of outer hair cells, the likely source of an active cochlear response, also adapts with time constants which may be significant when quantifying short-term adaptation [8]. Finally, neural responses to onsets and abrupt spectral changes are substantial [9], providing a physiological substrate for the sensitivity of human speech perception to onsets and dynamic spectral cues [10]. Although recognition systems typically statistically characterize the evolution of relatively static spectral segments, the auditory system responds most strongly to dynamic segments. This response strength is a consequence of adaptation. What remains is to quantify the adaptation, and to design a mechanism that reproduces it.

The task is similar to observing evidence of frequency selectivity and requiring a specification (critical bandwidths) and a mechanism for its realization (a filterbank). Following the example of using static masking data to quantify frequency selectivity [11], adaptation was quantified from a series of dynamic, forward-masking experiments. The adaptation mechanism designed is a modified form of automatic gain control (AGC), which adds an exponentially adapting linear offset to logarithmic intensity. Just as the current triangular

Manuscript received September 6, 1995; revised November 12, 1996. This work was supported in part by NIDCD Grant DC02033 and by the NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

The authors are with the Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: bps@ucla.edu; alwan@icsl.ucla.edu).

Publisher Item Identifier S 1063-6676(97)06390-6.



(a)

(b)

(c)

Fig. 1. Forward-masking stimuli. (a) Large time-scale view of a single 2AFC trial. (b) Fourier transform of the probe signal (128 ms rectangular window). (c) Smaller time-scale view of the probe following the masker by 15 ms.

filters used in the common ASR front end are first-order approximations of auditory frequency selectivity, the simple dynamic mechanism provides a first-order approximation of auditory adaptation. The strategy is to parameterize simple dynamic mechanisms from forward masking thresholds to provide a better approximation of the auditory response to dynamic stimuli.

Dynamic auditory models [12]–[16] are often physiologically based computational models that characterize only a relatively low level of the complete auditory system, or resort to some speculation either about higher level processing and/or about appropriate dynamic parameters. Because these systems usually require processing time-domain signals for each auditory filter (often $\approx$ 100 filters) at the full sampling rate, they imply a large computational burden, making them difficult to use in engineering applications [3]. Also, successfully separating and quantifying measurable functionality (e.g., frequency selectivity, or short-term adaptation), which may be distributed across several related physiological processes, is not a simple task. Some researchers [17], [18] propose novel computationally efficient techniques, targeted at automatic speech recognition, which emphasize spectral dynamics with varying perceptual accuracy and recognition improvements. The approach here differs from most detailed physiological models in that it “closes the loop” with observations of top-level functionality. Because the relatively simple model of frequency selectivity followed by additive adaptation is consistent with underlying physiological processes, the resulting quantified nonlinear model provides useful approximations of the perception of (nonstationary) speech.

II. Forward Masking

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

To build the dynamic model, data describing *sinusoidal* forward masking were desirable. The most complete data of pure-tone forward masking experiments is from [19]. Although [19] includes a wide range of frequencies and masker levels, the longest probe delay measured is 40 ms, short of the duration necessary for complete adaptation. To obtain recovery parameters, a set of pure-tone forward-masking experiments that included probe delays from 15 to 120 ms across wide ranging frequencies and masker levels was performed. Short-delay pure-tone forward-masking data, from [4], as a function of masker duration, were used to quantify attack parameters.

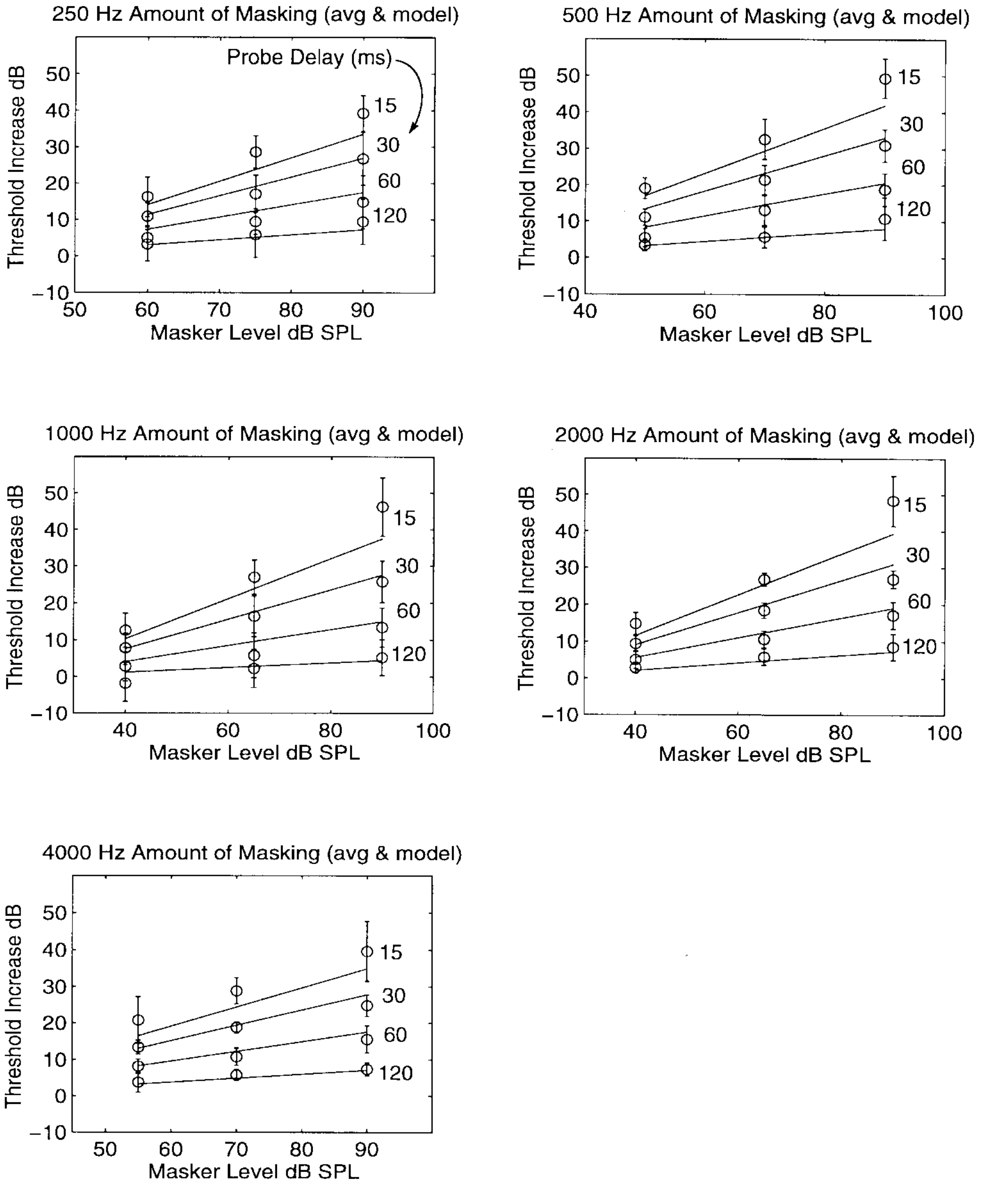


Fig. 2. Average forward masking data (circles), and standard deviation (error bars), together with the model fit (lines) as a function of masker level across five octaves, with probe delay of 15, 30, 60, and 120 ms as parameter.

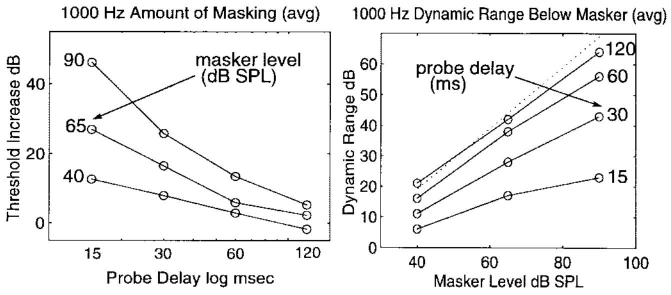
A. Experiments

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

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

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

To reduce the spectral splatter of transitions, the entire stimulus was filtered through a linear-phase, finite impulse response (FIR) filter, with a bandwidth of one critical band [21]. In the Fig. 1 example, the frequency is 1 kHz [Fig. 1(b)], the delay from masker to probe is 15 ms [Fig. 1(c)], and (measured at the envelope peak) the probe is 8 dB less intense



(a) (b)

Fig. 3. Average forward masking data at 1 kHz: (a) as a function of the log delay with contours for constant masker levels and (b) as the dynamic range below masker as a function of the masker level with contours for constant probe delays. The dotted line reflects the probe threshold in quiet.

than the masker. The stimulus is shown after the critical band filter.

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

3) Methods: For each condition, the level of the probe was adaptively varied to find its threshold. An adaptive "transformed up-down" procedure [22] determined the 79% correct point, defined as the threshold for the 2AFC task. The initial adaptation step size of 4 dB was reduced to 2 dB and 1 dB after the first and third reversals. The initial probe was clearly audible. The experiment continued for nine reversals. The probe levels at the last six reversals were averaged to determine the threshold. Thresholds were averaged across the five subjects to obtain the values used for parameterizing the model.

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

B. Results

Fig. 2 summarizes the average threshold increase (circles) across the five subjects as a function of masker level with probe delay as a parameter. The solid lines in Fig. 2 indicate the model's fit to the forward masking data. The derivation of the model is described in the following sections.

C. Modeling Implications

The amount of forward masking (in dB) decays as a straight line as a function of the logarithm of the probe delay (first described in [20]). A straight line with respect to logarithmic probe delay can be approximated by an exponential with respect to linear probe delay. This suggests additive exponential adaptation in decibels.

Fig. 3(a) plots the threshold increase as a function of probe delay, and Fig. 3(b) shows the effective dynamic range below masker, defined as the difference between the masker and probe threshold levels, as a function of masker level. Fig. 3(a) shows that the rate of decay of the forward masking (shown on a log time scale) increases with an increasing amount of masking. These data may suggest different adaptation rates for different masker intensities, or complexity beyond a simple exponential adaptation of dB level. Such complexity is not necessary. The adapting mechanism derived below has a greater initial distance to target after a more intense masker offset. Exponential processes decay more quickly over the same amount of time when the output is further from the final static target. Therefore, a simple exponential dynamic mechanism can predict a faster rate of decay of forward masking with more intense maskers.

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

III. From Experimental Results to Model Parameters

In the perceptual model, a dynamic adaptation stage follows each output of a linear filterbank. At every time sample, each adaptation stage slowly adjusts an internal offset to move its output incrementally closer to an input/output (I/O) target, specified on a log/log scale.

The dynamic adaptation stages are referred to as automatic gain control (AGC). However, it is significant that the AGC is implemented as an adapting *additive* offset to the log energy of the signal, and not as an adapting multiplicative gain. There are at least two points that appear to require additive, and not multiplicative, adaptation. First, the measured incremental neural response to a second onset after partial adaptation to a first is not proportional to an adapted amount of multiplicative gain [6]. Second, AGC that adjusts a multiplicative gain proportional to the linear distance to the I/O target does not predict a higher rate of decay of forward masking for greater amounts of masking.

A. AGC: I/O Curves, Attack, and Release Times

Time constants describing the rate of adaptation for the dynamic mechanisms are defined here as the time required for the *logarithmic* distance to target to reduce by a factor of $1/e$. Different time constants are used for attack (decreasing offset), and release (increasing offset). Over short durations, the AGC stage has little time to adapt, and is therefore nearly linear. I/O graphs do not include a time axis, so to discuss the temporal evolution of the system, we describe trajectories

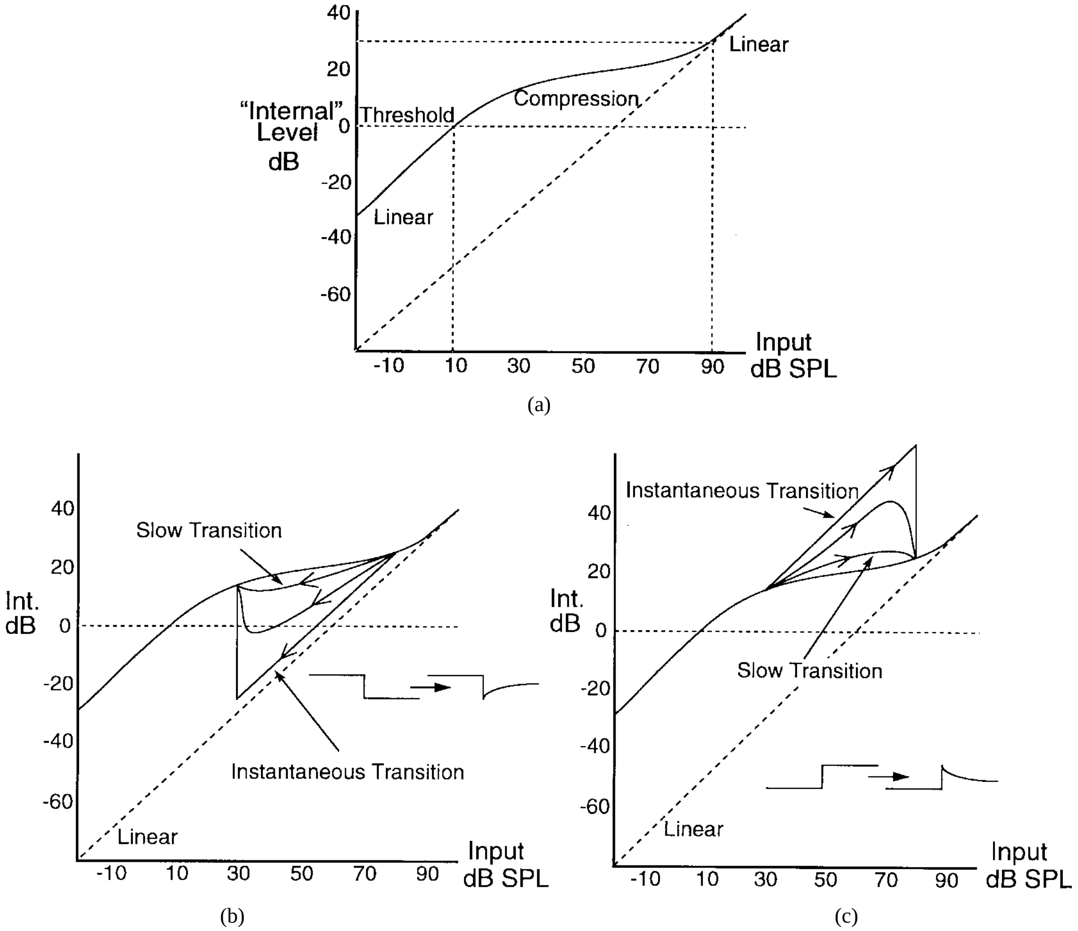


Fig. 4. (a) Prototypical I/O curve for a single channel in the dynamic model. Schematic output trajectories corresponding to a level change at three different rates for (b) decreasing inputs from 80 to 30 dB SPL, and (c) increasing inputs from 30 to 80 dB SPL.

that characterize the motion of the instantaneous I/O point on the I/O graph. When the input changes abruptly, the output initially tracks the input, moving in nearly a 45° line. Over long durations with static inputs, the output asymptotically approaches the I/O target.

Fig. 4(a) shows a prototypical I/O curve for a single channel in the dynamic model. At low levels, the I/O function is nearly linear, over normal levels it is compressive, and at extremely high levels it is again linear. The general shape of the prototypical I/O curve was motivated by the saturating response of the basilar membrane [23]. For each adaptation stage, a fixed internal threshold, corresponding to the static audibility threshold, is imposed at the compression threshold. Similarly, the compression region ends, and the model again becomes linear, at a high level of equal loudness (near 90 dB SPL), which varies with the center frequency of the adaptation stage. By carefully choosing the threshold and I/O curve for each adaptation stage, the AGC sections map a specified static input range as a function of center-frequency into a normalized internal level consistent with constant loudness contours.

Fig. 4(b) and (c) schematically show the response of the model to decreasing and increasing inputs, respectively. When the input changes abruptly, the trajectory on the I/O curve moves along a 45° angle, and then slowly settles to the target on the I/O curve. When the input changes slowly, the output trajectory follows the I/O curve more closely. The model predicts forward masking when output trajectories momentarily fall below the internal threshold, as in Fig. 4(b).

B. Derivation of Model Parameters

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

To simplify the model and this derivation, a constant I/O slope is imposed across the compressive region. Fig. 5 describes the geometries necessary to measure the model's prediction of the forward-masking threshold with long maskers as a function of masker level and probe delay. Before the masker offset, the output trajectory reaches the target on the I/O curve (point A in Fig. 5). As the masker shuts off abruptly, the output trajectory instantly falls along the diagonal (from A to B). Once the trajectory is below the compressive region, the distance to target is constant, and the model adapts by slowly increasing toward maximum additive offset (from B toward C). At some point during this adaptation (point C), the onset of the probe causes an abrupt transition from below threshold back up along a new diagonal (from C to D). If the probe level is intense enough to place the trajectory above threshold

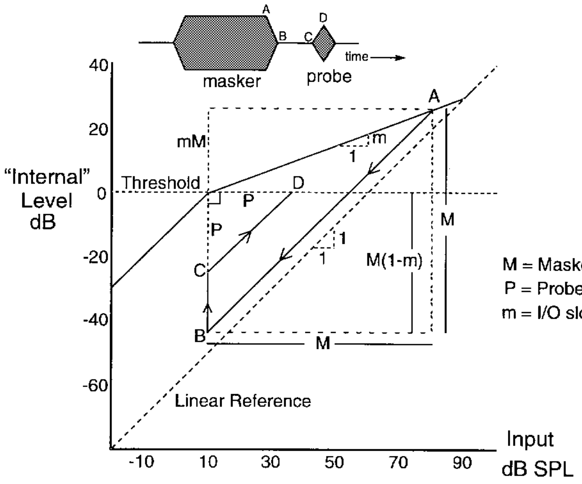


Fig. 5. Geometry to derive recovery (upward adaptation) parameters from forward-masking thresholds.

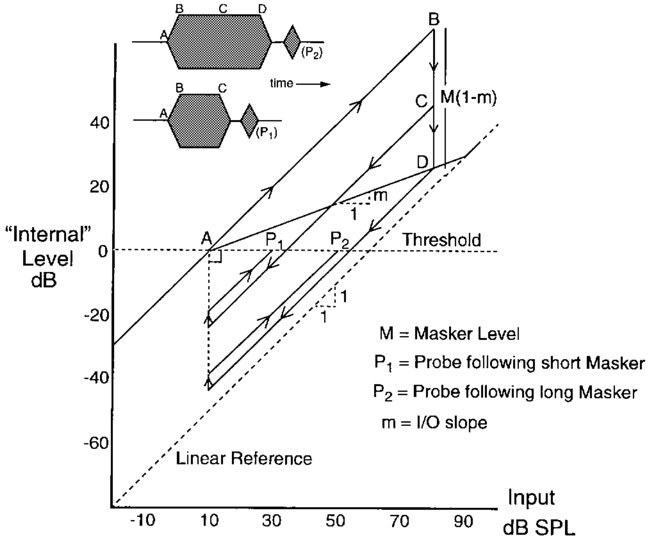


Fig. 6. Geometry to derive attack (downward adaptation) parameters from forward-masking thresholds as a function of masker duration.

(at the instant of the probe onset) the probe is audible. If the internal level just reaches threshold, the model predicts a forward masking threshold (at point D).

Incremental adaptation of the model is implemented using a (nonconstant coefficient) first-order difference equation leading to an exponential decay of the logarithmic distance to target. From the geometry in Fig. 5, probe level at threshold P as a function of masker level M, discrete-time probe delay n, I/O slope m, and incremental adaptation a, is

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

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

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

Therefore, the data points at the shortest delay [Fig. 3(b)] provide an approximation for the I/O slope parameter m. An iterative procedure was used to minimize the total mean squared error (MSE) between the model predictions of the probe thresholds and the average forward masking data for all data points at each center frequency, as a function of the two model parameters m and a. The total MSE is relatively insensitive to the I/O slope, m, compared to the adaptation parameter a. Therefore, the initial estimate of m from the short-delay conditions was averaged with the value that minimizes total MSE, to determine a final m estimate. A second MSE minimization as a function of only a determined the final a estimate.

Just as forward-masking data as a function of probe delay are used to characterize recovery, the change of forward masking with the duration of the masker is used to characterize attack. Short-duration maskers reduce the time for downward adaptation, which decreases the amount of adaptation, and in turn, reduces the time to recovery. Geometries necessary to derive attack (downward adaptation) parameters are described in Fig. 6. Before the onset of the masker, the model reaches the static threshold (at point A in Fig. 6). At the abrupt masker onset, the output trajectory translates diagonally upward (from A to B) and then slowly drops toward the I/O target as the model adapts (from B to C to D). If the duration of the masker is short relative to the downward time constant, the trajectory will not reach the I/O target by the time of the abrupt masker offset (point C). In response to the masker offset, the output trajectory corresponding to the short masker moves diagonally (from point C), crossing the internal threshold at a lower point than the trajectory corresponding to the longer masker (from point D). After brief recovery during a short probe delay, the model predicts less forward masking from the short-duration masker.

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

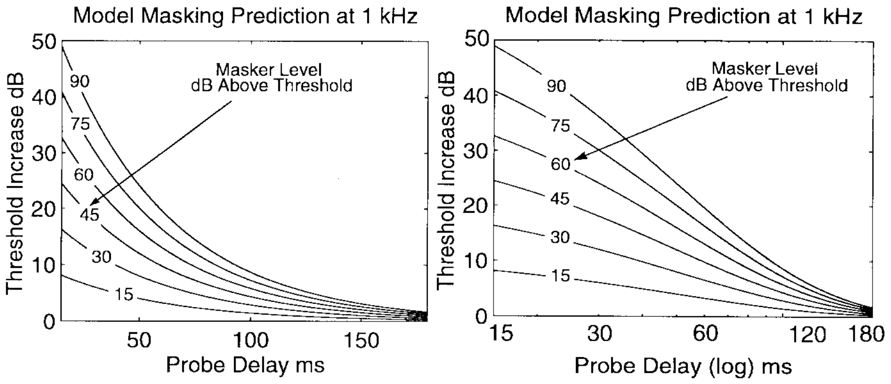
$$P = M(1-m)(1-b^{nd})a^{nu}.$$

The probe threshold difference, ΔP, between short and long masker durations is

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

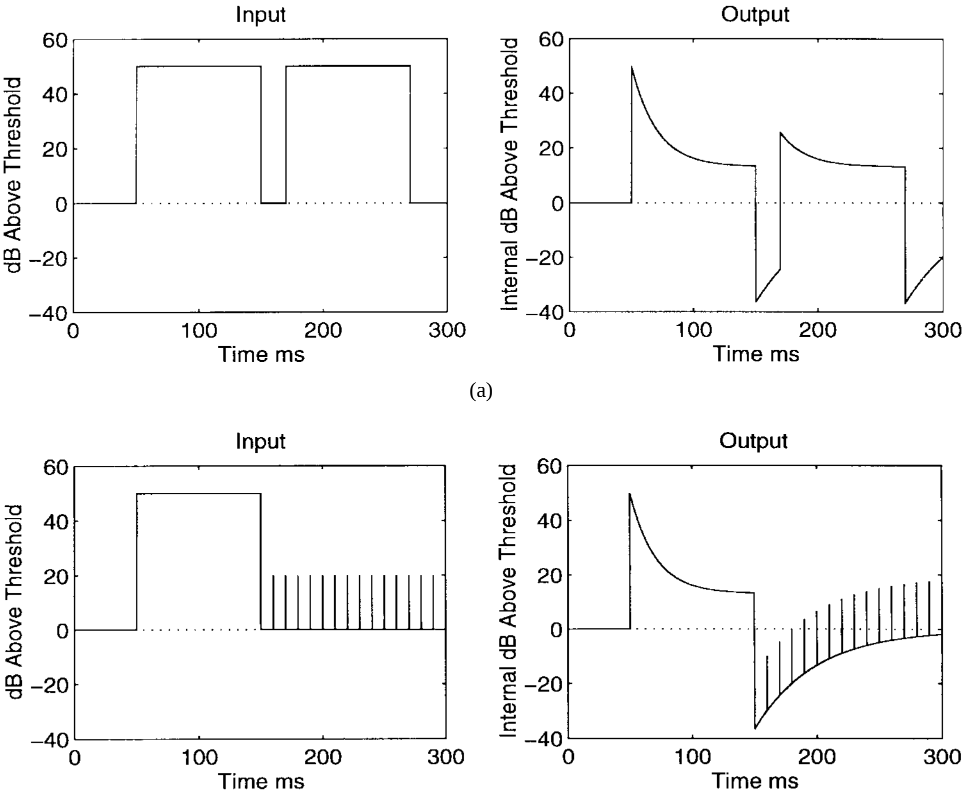
This probe threshold difference equation was solved for the model parameter b, and then its value was estimated from the differences reported in [4], using the m and a parameters derived above. Table I summarizes the model parameters and adaptation time constants across frequencies. The a and b terms are with respect to a 100 Hz spectral sampling rate. Adaptation stages with center frequencies between measured points use a weighted average of neighboring parameters. Attack time constants are approximately three to four times shorter than release time constants. These times, and more accurately their ratio, approximate those derived from physiological data [13].

Fig. 7 shows the model's prediction of the decay of masking at 1 kHz. Note that the decay rate of forward masking is greater with more intense maskers, and that the decay is nearly



(a) (b)

Fig. 7. Model's prediction of the decay of forward masking as a function of masker level at 1 kHz with (a) linear time reference and (b) logarithmic time reference.



(a)

(b)

Fig. 8. Adaptation to, and recovery after, a pulse. (a) Response to the second pulse is diminished. (b) Impulses, corresponding to onsets, are initially masked (similar to figures in [13]).

linear with logarithmic time. Fig. 8 shows two examples of the model's behavior at 1 kHz. Fig. 8(a) shows the response to two consecutive pulses. The model adapts in response to the onset of the first pulse, and the response to the onset of the second pulse "rides on top of" the recovery from adaptation. Fig. 8(b) shows forward-masking examples. The model starts adapting at the onset of the long pulse, and then recovers after its offset. Lower-intensity impulses following the long pulse, corresponding to potential probe onset points, again ride on top of the model's recovery from adaptation to the pulse. The responses to the impulses are initially below threshold (masked) and with time, rise above threshold.

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

First, the exponential derivation assumes the 15 ms delay between the masker and probe is silence. This assumption provides the maximum possible distance to target during the

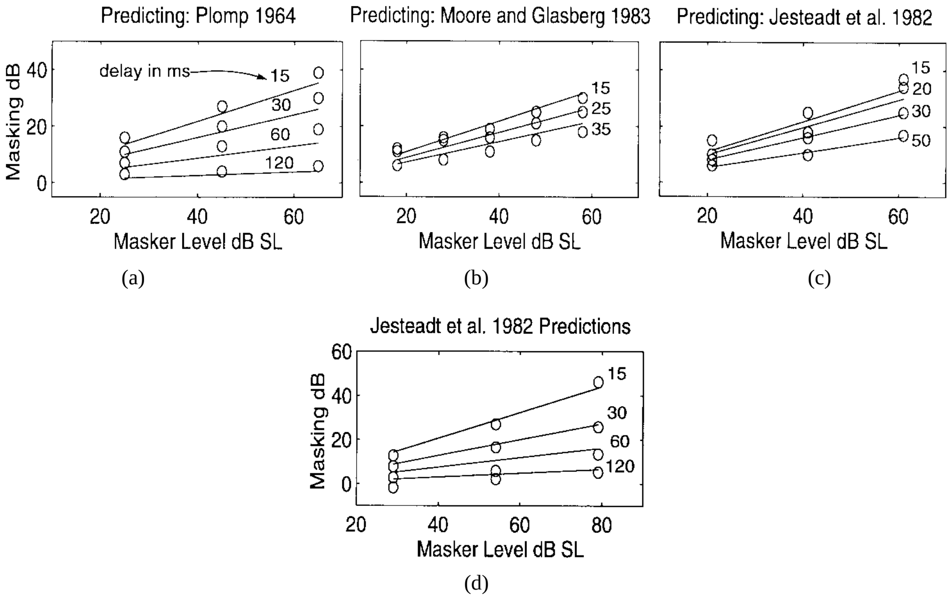


Fig. 9. Using the model to predict other forward masking data. (a) Wideband masker and probe [20]. (b) Wideband masker, sinusoidal probe at 1 kHz [24]. (c) Sinusoidal masker and probe at 1 kHz [19]. (d) Equation provided in [19] predicting the present data.

15 ms, the maximum amount of recovery, and the lowest prediction of forward masking. In fact, the stimuli had 5 ms of offset, 5 ms of silence, and 5 ms of onset during this interval. Any nonsilence during the 15 ms delay decreases the distance to target, reduces the amount of recovery, and increases the estimation of forward masking. Ignoring the finite onsets and offsets reduces the model's predictions of the amount of forward masking at short delays.

Second, in this derivation, forward masking is assumed to occur when insufficient auditory recovery keeps the response to the probe below threshold. However, at shorter (near zero) delays, with extremely similar maskers and probes, the probe may only be audible as a change in level at the end of the masker [24], and not as a separate event. Even though the response to the probe is above threshold, the subject may not distinguish the probe from the masker, and therefore not detect the probe. Because the derivation requires the model's response to the probe to be below threshold to be masked, it underestimates the amount of forward masking, especially at short delays with intense maskers.

C. Predicting Other Data

Fig. 9 (a)–(c) shows the model's predictions of previous forward masking data. Fig. 9(a) shows the model's prediction of average data with wideband stimuli [20]. These data provide relatively complete measurements of forward masking across level and delay. In the results shown in Fig. 2, there is only slight variation of forward masking with frequency. Because the adapting response of the model to wideband stimuli approximates the response at middle frequencies, the wideband data were predicted using the model parameters derived from the 1 kHz data. Although the model underestimates these data, the trends are consistent.

Figs. 9(b) and 9(c) show the predictions for wideband and pure-tone maskers of 1 kHz pure tones, respectively [19], [24]. These measurements were made at relatively short delays. Authors have historically disagreed on how to specify delay in a forward-masking experiment [20]. In this paper, delay is measured between the envelope peaks, while [19] used zero-voltage points, and [24] chose half-voltage points between the masker and probe *offset*. The present study used 5 ms ramps, [19] used 10 ms, and [24] used 5 ms for the masker and 10 ms for the probe. To compensate for these differences, 2.5 ms is subtracted from the delay reported in [24], and 10 ms is added to the numbers in [19]. The masker level in the 1 kHz band for the wideband masker is determined by the energy in the critical band [21] centered at 1 kHz. Although comparisons are only possible at relatively short delays, the model overestimates the amount of masking by wideband noises, and underestimates masking by pure tones. Once parameterized, however, the simple dynamic mechanism approximates dynamic psychophysical responses.

TABLE I
MODEL PARAMETERS RELATIVE TO A 100 Hz SPECTRAL RATE

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

Fig. 9(d) shows the prediction of data from this study by an equation proposed in [19]

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

P and M are the levels of the probe and masker above threshold, and the constants $a, b,$ and c are chosen to fit the average forward-masking data at 1 kHz in [19]. Even though the parameters in this equation were chosen from a data set that did not include measurements at the longer delays used in this study, it provides an excellent prediction of the present data.

D. Other Models Predicting Forward Masking

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

Forward, backward, and forward/backward masking combinations have been predicted with great precision assuming a relatively standard model of filtering, rectification, power-law compression, temporal integration and a decision device [25]. In its original derivation, however, there was no mechanism to account for the level-dependence of forward masking. Either the temporal window shape [25] or the power-law compression [26] may vary with level. The decision device required an unusually high minimum detectable temporal amplitude variation of 6 dB, which may not extend well to a general processing scheme. Finally, if forward masking is entirely a consequence of temporal integration, physiological measurements of adaptation are ignored, and there is no mechanism that explains physiological and perceptual sensitivity to onsets and transitions.

Other researchers have proposed models using adaptation mechanisms to explain forward masking [27]–[29]. The first of these [27] uses a modified version of a previous model [30] that includes filtering, envelope detection, power-law compression, rapid and short-term adaptation, and long-term integration. The long-term integrator is bypassed in forward-masking tasks. Immediately following a stimulus, the model assumes that there is no rapid onset component in response to a probe, that this component recovers exponentially with time, and that the relative level of this component is used to determine forward masking. The model is somewhere between a complete processing mechanism and an equation summarizing psychophysical responses, and therefore, is also difficult to incorporate into ASR systems. The exponential recovery of the rapid onset component has similarities to the exponential adaptation used in the dynamic mechanism described in this paper.

More recently, other researchers have developed a general auditory model that, together with an optimal decision device, predicts well a wide variety of psychophysical data [28], [29]. In each auditory channel, the model uses linear filtering, half-wave rectification, and lowpass filtering, followed by five adaptation stages. The output is correlated with templates that store the model's response to other (masker-only) conditions to predict masking thresholds, thereby imposing a relatively complex postprocessing mechanism to predict the data. The model provides a dynamic spectral representation of speech that is likely to improve recognition robustness; potential application improvements may warrant the significant computational complexity.

IV. Peak Isolation

Both speech perception and the response of individual auditory nerves are extremely sensitive to the frequency position of local spectral peaks. There are several mechanisms and corresponding modeling approaches that may explain this sensitivity. Physiologically motivated by the local fan-out of the neural connection to outer hair cells, [14] suggests cross-coupling AGC stages to improve static spectral contrast, providing functionality similar to the higher level lateral inhibitory network in [31]. Significant effort [15], [16], [32] also focuses on modeling how the auditory system derives, and makes use of, redundant temporal microstructure. Auditory nerves with center frequencies as far as an octave away from a local spectral peak can synchronize their response to the frequency of the peak, providing a composite neural representation dominated by that frequency [33]. Similarly, perceptual discrimination of vowels is more sensitive to the frequency location of spectral peaks than to other aspects of the spectral shape [34]. These data suggest that the auditory system may derive a noise-robust representation by attending to the frequency locations of local spectral peaks.

The dynamic model was therefore also evaluated with a novel processing technique, based on raised-sine cepstral liftering [5] together with explicit peak normalization, which isolates local spectral peaks. Raised-sine cepstral liftering is weighting the cepstral vector by the first half-period of a raised-sine function.

The cepstral vector is an expansion of the even log spectrum in terms of cosine basis functions. The c_0 term specifies the log-spectrum average, the c_1 term approximates the log-spectrum tilt, etc., and high cepstral terms represent quickly varying ripples across the log spectrum. Weighting the cepstral vector specifies the relative emphasis of different types of log-spectrum variations. A raised-sine lifter deemphasizes slow changes with frequency, often associated with overall level and vocal driving-function characteristics, as well as fast changes that may reflect numerical artifacts [5].

It is helpful to view the effects of cepstral liftering in the log spectral domain. Fig. 10(a) starts with the log spectrum, from a vowel $(/i/)$, implied by a truncated cepstral vector. Fig. 10(b) shows the log spectrum implied after raised-sine cepstral liftering. The average level as well as slow (and fast) variations with frequency are deemphasized, leaving components that change at a moderate rate with frequency. This process emphasizes both spectral peaks and valleys.

The valleys are removed by half-wave rectifying the log spectral estimate implied after raised-sine liftering, and a final vector is obtained by transforming back to the cepstral domain. Because the half-wave rectifier is nonlinear, explicit transformation from cepstrum to log spectrum (processing through the rectifier) and then transformation back to cepstrum are required. The raised-sine lifter also affects the magnitude of the peaks. Therefore, before transforming back to the

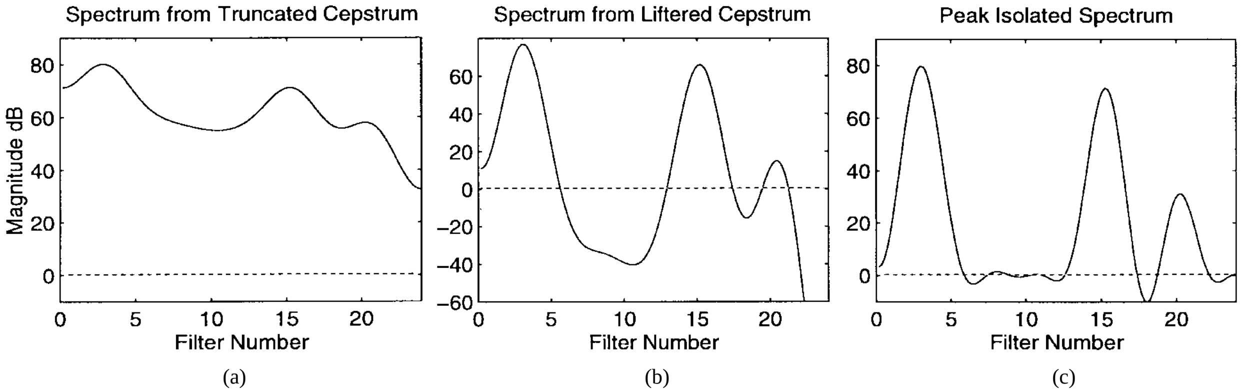


Fig. 10. Peak isolation processing. Log spectrum of the vowel $/i/$ after (a) cepstral truncation, (b) raised-sine cepstral liftering, and (c) half-wave rectification and peak normalization.

cepstrum, peaks are scaled to the level measured in the original log spectrum. The final peak-isolated estimation is shown in Fig. 10(c).

V. Robust Recognition Evaluation

The model was evaluated as the front end for two word recognition systems. The first is a talker-dependent dynamic time warp (DTW) system, and the second uses talker-independent hidden Markov models (HMM). The DTW system provides an initial assessment of the model and the HMM evaluations are better approximations of potential ASR applications. The vocabulary for all systems is limited to the ten digits. Finally, a comparison with RASTA-based processing is included. The inputs to all recognition experiments are corrupted with additive noise shaped to match the long-term average speech spectrum [35]. Test words are embedded in (noisy) silence, so that the recognizers are required to both isolate and recognize the discrete words.

Two versions of the dynamic model were implemented: a full-rate system and a downsampled version. The full-rate system uses rounded exponential filter shapes [36], and then adapts the envelope of each filter output at the full sampling rate. The downsampled system obtains Mel-scale power spectrum estimations every 10 ms by weighting and adding power spectrum points from a fast Fourier transform (FFT), and then adapts these outputs at the downsampled rate. On an HP715 workstation, the downsampled system runs at 0.43 × real time, while the full-rate implementation requires 9.4 × real time. The recognition evaluations below used the downsampled implementation.

Three basic front ends are compared: linear prediction cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), and mel-frequency cepstral coefficients with adaptation (MFCCA). Each front end computes a spectral estimation every 10 ms using overlapping 30 ms Hamming windows. LPCC are computed in two stages [1]: 12th-order, autocorrelation-based linear prediction provides an all-pole vocal-tract transfer function. Real cepstral coefficients are then recursively computed for this minimum-phase estimation. MFCC are computed in three stages [2]. The power spectrum is computed using a zero-padded FFT. To estimate the energy at the output of each approximate auditory filter, power spectrum outputs are weighted by a triangular filter shape and then summed. The filters have a half-power bandwidth of 100 Hz up to center frequencies of 1 kHz, and a bandwidth of 0.1 times the center frequency above 1 kHz. A DCT converts the spectral estimation obtained from the logarithmic energy across filters into a final cepstral vector. Before the DCT, the logarithmic filter energies of MFCC are also processed through the dynamic stages derived in Section III to obtain the adapting spectral estimation vector MFCCA. A 13-element cepstral vector and its temporal derivative (approximated by the slope of a linear fit to seven cepstral points) are obtained for each front end, but the undifferentiated spectral level term (c_0) is ignored during recognition.

For the initial DTW evaluation, the peak isolation mechanism was applied only to the MFCCA to obtain MFCCAP. For subsequent HMM evaluations, all front ends were compared with and without peak isolation.

Fig. 11 shows spectral representations of the digits “nine six one three” from MFCC, MFCCA, and MFCCAP. The dynamic model emphasizes spectral changes in time, while peak isolation enhances spectral contrast in frequency. Together, these mechanisms highlight the spectro-temporal representation of changing frequency peaks. The second half of this picture shows representations at 5 dB signal-to-noise ratio (SNR). Onsets, transitions, and changing local spectral peaks may remain as robust cues for recognizing speech in a noisy background.

A. DTW Evaluation

An initial evaluation with a simple dynamic programming-based isolated word recognition system [1] and a single talker was performed. A system was constructed that used an Itakura path constraint [1], and a Euclidean local distance metric excluding the undifferentiated c_0 term. Clean templates were isolated from surrounding silence, but test tokens were not. As more noise is added, word isolation, or endpoint detection, becomes more difficult. To asses the robustness of the system, it is therefore, unrealistic to assume the temporal placement of the speech within the background noise is known. Instead, dynamic programming is used to find the speech within the noise. At each time slice in the test token, a new path starts at the beginning of the template and an accumulated distance

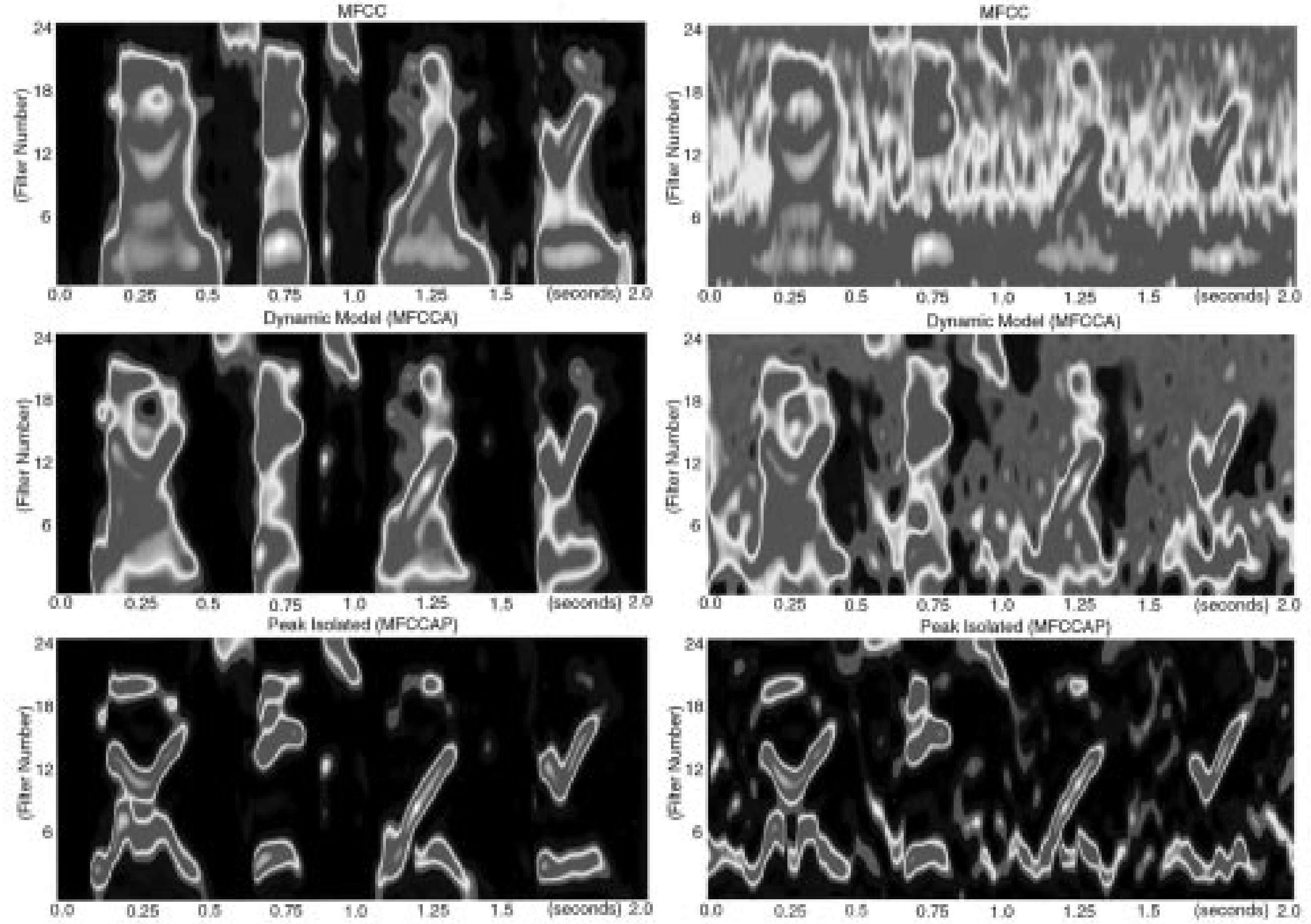


Fig. 11. Spectrogram representations of the words "nine six one three" from MFCC analysis, the dynamic model MFCCA, and the dynamic model with peak isolation MFCCAP, at greater than 40 dB SNR and at 5 dB SNR.

propagates through the end of the template. Accumulated distances usually reach a minimum at the end of the speech in the test token, marking the best alignment for that test token/template pair, without explicit endpoint detection. The accumulated distances are divided by their path lengths to normalize for templates of different duration. The minimum normalized distance specifies the distance to each template, and the minimum template distance determines the word recognized.

The data were digitally recorded from a single talker in a sound-isolated room using a close-talking microphone. Fig. 12 shows the degradation of recognition performance in background noise across the four front ends: LPCC, MFCC, dynamic model (MFCCA), and the dynamic model with peak isolation (MFCCAP). Consistent with [3], the MFCC is more robust than LPCC. However, both the dynamic model MFCCA and the dynamic model with spectral peak isolation MFCCAP are significantly more robust to background noise than MFCC.

B. HMM Evaluations

Using the male talkers in the TI-46 data base and the HTK-Toolkit, a series of talker-independent HMM-based robust digit isolation and recognition evaluations were also conducted. The TI-46 database is hand segmented so that words are placed in the center of each file. Before adding background noise to these files, random amounts of silence were added before and after each token. Two sets of evaluations were performed. The first used only clean data for training while the second trained both clean and noisy models.

For all models, six-states per word, simple left-to-right state transitions, continuous Gaussian densities, diagonal covariances, and fixed global variances were used. Mean feature vectors and transition probabilities for each state were trained as described below, but variances were set to the global variance estimated over all tokens in the training set. This technique is useful with limited training data and when the testing environment is significantly different from the training environment [3].

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

When the test environment differs from the training environment, recognition performance deteriorates. A common

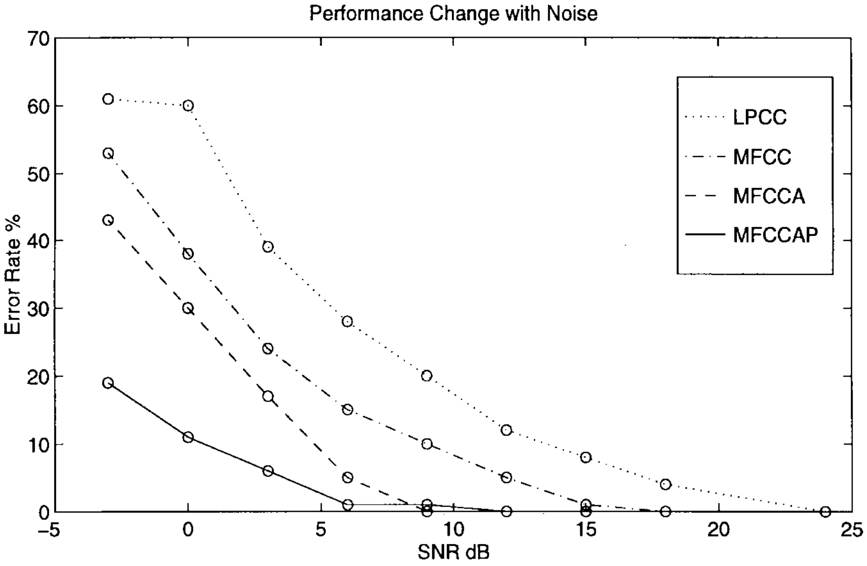


Fig. 12. Talker-dependent DTW recognition performance in noise.

approach to address this issue is to train models using noisy data [1]. One set of clean models was built, as described above, and then a second set of "noisy models" was built using training data at an SNR of 12 dB. Both sets of models were used for recognition; the model with the highest probability (from either set) determined the word recognized. To train the noisy models, stationary background noise was added, and then forced-Viterbi alignment with the corresponding clean model was used to isolate the noisy speech from the background. The same Viterbi and forward-backward training algorithms, used for training clean models, were used to train noisy models from the isolated noisy words.

For Viterbi alignment in training and recognition, silence models were used together with a "grammar" of silence-word-silence. In a fixed-variance system, the silence models were simply the long-term moving average of the front end's response to the background noise. As the SNR changed, the silence model's mean updated to the new background noise.

Fig. 13(a) shows the increasing error rate at lower SNR's for the different front ends. Each front end was evaluated with and without the peak isolation mechanism. The dynamic model MFCCA by itself shows no improvement over standard MFCC, however, adaptation improves the robustness of MFCCP, and MFCCAP remains the most robust front end. Isolating peaks is helpful, but isolating changing peaks is perhaps more helpful.

There are at least two reasons to expect the performance of the dynamic model to degrade when using HMM-based recognition in a noisy environment. The dynamic model provides a context-dependent response that may increase differences between onset responses in clean and noisy environments. However, MFCCA improves DTW performance over MFCC. The difference may be that in the DTW system, templates are continuously varying over the utterance. The HMM system requires discretizing the variation over the utterance into a finite number of states. The nonstationary response of the dynamic model (as seen in Fig. 11) may not be as well-suited to segmentally stationary statistical characterization as the MFCC representations; intrasegment changes are reduced to averages.

Fig. 13(b) shows the evaluation using models of both clean and noisy data. Performance across all front ends improves, and MFCCAP continues to provide the most robust representation. This evaluation removes some of the context-dependent mismatch between training and testing.

Fig. 13(c) and (d) compare the performance using the dynamic mechanism and the perceptually motivated RASTA technique [18]. RASTA involves filtering the logarithmic temporal trajectories (log energy temporal excitation patterns) with a bandpass filter that has a sharp zero at DC. By deemphasizing slow and fast changes with time, RASTA also provides an adapting response. Both front ends were evaluated with and without the peak-isolation algorithm. Fig. 13(c) shows the performance with clean models, and Fig. 13(d) compares the performance with clean and noisy models. Our dynamic mechanism is more robust for these tasks. In this comparison, the RASTA technique was applied directly to the logarithmic filter energies, without the perceptual linear prediction (PLP) processing used in its original optimization [18]. The "standard" RASTA filter

$$H(z) = .11\frac{(2+z^{-1}-z^{-3}-2z^{-4})}{1-0.94z^{-1}}$$

was used and performance was not compared with other RASTA variations that optimize the compressive and expansive nonlinearities for the specific acoustic environment.

Spectral estimations on a perceptual frequency scale (MFCC) are more robust than those on a linear scale (LPCC). Adaptation provides sensitivity to onsets, enhancing spectral contrast in time. Unlike the RASTA technique, which can be described as a (smoothed) first-order differentiation, the dynamic model proposed here does not provide zero output for constant input. Instead, the adaptation stages converge to static targets on the I/O curves. Also unlike the RASTA technique, recovery is roughly three times

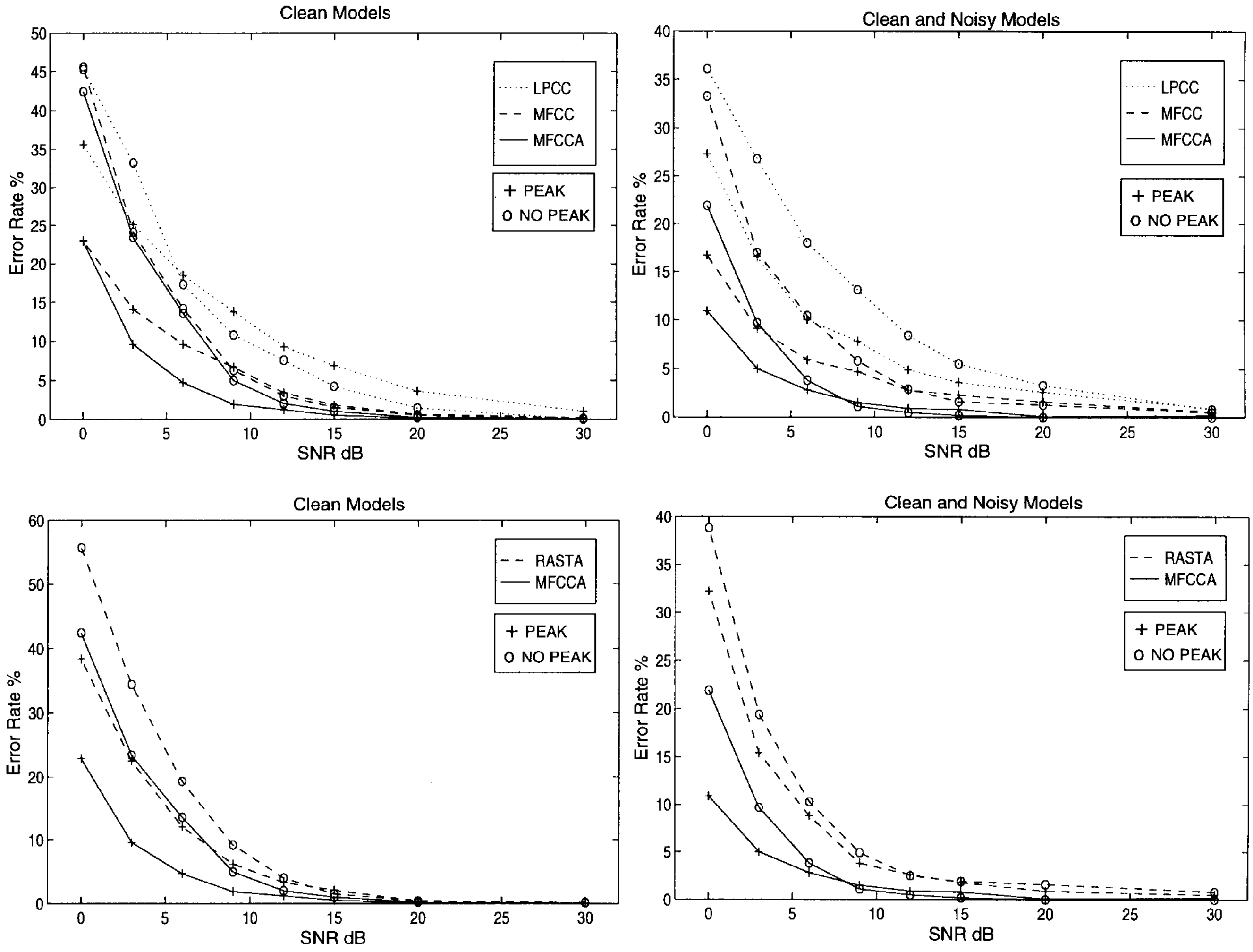


Fig. 13. Talker-independent HMM comparisons: LPCC, MFCC, and MFCCA with (a) clean models, (b) clean and noisy models. MFCCA and RASTA with (c) clean models, (d) clean and noisy models. (+) indicates with peak isolation, (o) indicates without peak isolation.

slower than attack. Finally, peak isolation enhances spectral contrast in frequency. The combination of adaptation and peak isolation provides a spectral estimation sensitive to changing local spectral peaks, enhancing the representation of speech in a noise background. The dynamic mechanism with peak isolation (MFCCAP) reduced the word recognition error in background noise by a factor of two to three over common (MFCC) front ends in each of these evaluations, and provided an improvement over the RASTA technique.

VI. Conclusions

Current speech recognition systems use a simplified auditory model to transform a temporal pressure wave into a sequence of spectral estimations. Specifically, ASR front ends approximate auditory frequency selectivity and magnitude compression. This paper provides two simple nonlinear mechanisms that extend the front end to include adaptation and sensitivity to the frequency location of local spectral peaks. These mechanisms impose additional computational requirements roughly equal to that of the common ASR front end. Forward-masking data parameterize the adaptation mechanisms. Using additive exponential adaptation after logarithmic conversion, the dynamic mechanism predicts a nearly linear decay of the amount of forward masking (in decibels) as a function of logarithmic probe delay, and faster rates of decay of forward masking from more intense forward maskers. The output is below threshold when forward masking is predicted to occur (a decision device is not used), allowing for direct connection to current recognition systems. The peak isolation mechanism is an extension of raised-sine cepstral liftering. Together with the common MFCC front end, these mechanisms imply an auditory system with frequency selectivity and magnitude compression that is highly sensitive to onsets, transitions, and changing local spectral peaks. Each of these mechanisms improves the noise-robustness of a simple word recognition system. Together they reduce the error rate by a factor of two to three over an MFCC front end.

Acknowledgment

The authors thank Dr. D. Morgan and four anonymous reviewers for their helpful suggestions on a previous version of this manuscript.

References

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[3] C. R. Jankowski Jr., Hoang-Doan H. Vo, and R. P. Lippman, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol 3, pp. 286–293, July 1995.

[4] G. Kidd, Jr. and L. L. Feth, "Effects of masker duration in pure-tone forward masking," *J. Acoust. Soc. Amer.*, vol. 72, pp. 1364–1386, Nov. 1982.

[5] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, pp. 947–954, July 1987.

[6] R. L. Smith and J. J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biol. Cybern.*, vol. 17, pp. 169–182, 1975.

[7] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *J. Neural Physiol.*, vol. 42, pp. 1083–1107, July 1979.

[8] J. Ashmore, "A fast motile response in guinea-pig outer hair cells: the cellular basis of the cochlear amplifier," *J. Physiol.*, vol. 388, pp. 323–347, July 1987.

[9] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," *J. Acoust. Soc. Amer.*, vol. 75, pp. 897–907, Mar. 1984.

[10] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1016–1025, Oct. 1986.

[11] H. Fletcher, "Auditory patterns," *Rev. Modern Phys.*, vol. 12, pp. 47–65, Jan. 1940.

[12] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2623–2629, June 1989.

[13] R. S. Goldhor, "Representation of consonants in the peripheral auditory system: A modeling study of the correspondence between response properties and phonetic features," RLE Tech. Rep. 505, Mass Inst. Technol., Cambridge, 1985.

[14] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE ICASSP*, Paris, France, May 1982, pp. 1282–1285.

[15] ——, "Computational models of neural auditory processing," in *Proc. IEEE ICASSP*, San Diego, CA, Mar. 1984, pp. 36.1.1–36.1.4.

[16] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonet.*, vol. 16, pp. 55–76, Jan. 1988.

[17] K. Aikaiwa and T. Saito, "Noise robust speech recognition using a dynamic-cepstrum," in *Proc. ICSLP*, Yokohama, Japan, Sept. 1994, pp. 1579–1582.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[19] W. Jesteadt, S. Bacon, and J. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Amer.*, vol. 71, pp. 950–962, Apr. 1982.

[20] R. Plomp, "Rate of decay of auditory sensation," *J. Acoust. Soc. Amer.*, vol. 36, pp. 277–282, Feb. 1964.

[21] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1523–1525, Nov. 1980.

[22] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Amer.*, vol. 49, pp. 467–477, Feb. 1971.

[23] B. Johnstone, R. Patuzzi, and G. K. Yates, "Basilar membrane measurements and the travelling wave," *Hearing Res.*, vol. 22, pp. 147–153, 1986.

[24] B. C. J. Moore and B. R. Glasberg, "Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise," *J. Acoust. Soc. Amer.*, vol. 73, pp. 1249–1259, Apr. 1983.

[25] A. J. Oxenham and B. C. J. Moore, "Modeling the additivity of nonsimultaneous masking," *Hearing Res.*, vol. 80, pp. 105–118, Oct. 1994.

[26] A. J. Oxenham and C. J. Plack, "Peripheral origins of the upward spread of masking," *J. Acoust. Soc. Amer.*, vol. 99, p. 2542, Apr. 1996.

[27] R. V. Shannon, "A model of temporal integration and forward masking for electrical stimulation of the auditory nerve," in *Cochlear Implants: Models of the Electrically Stimulated Ear*, J. M. Miller and F. A. Spelman, Eds. New York: Springer-Verlag, 1990, pp. 187–205.

[28] T. Dau and D. Pueschel, "A quantitative model of the 'effective' signal processing in the auditory system, I: Model structure," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3615–3622, June 1996.

[29] ——, "A quantitative model of the 'effective' signal processing in the auditory system, II: Simulations and measurements," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3623–3631, June 1996.

[30] J. J. Zwislocki, "Temporal summation of loudness: An analysis," *J. Acoust. Soc. Amer.*, vol. 46, pp. 431–441, Feb. 1969.

[31] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 412–435, July 1994.

[32] O. Ghitza, "Auditory nerve representations as a basis for speech processing," in *Advances in Speech Processing*, S. Furui and M. Sondhi, Eds. New York: Marcel Dekker, 1991, pp. 453–485.

[33] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: I, Vowel-like sounds," *J. Acoust. Soc. Amer.*, vol. 75, pp. 866–878, Mar. 1984.

[34] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: a first step," in *Proc. IEEE ICASSP*, Paris, France, May 1982, pp. 1278–1281.

[35] D. Byrne and H. Dillon, "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear Hearing*, vol. 7, pp. 257–265, Aug. 1986.

[36] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, Aug. 1990.

Brian Strope received the Sc.B. degree (with honors) in electrical engineering from Brown University, Providence, RI, in 1989. He designed workstation hardware for Hewlett-Packard, Fort Collins, CO, for four years. In 1995 he received the M.S. degree in electrical engineering from the University of California, Los Angeles, where he is currently pursuing the Ph.D. degree. His research focus is auditory and perceptual models and their application to speech recognition.

Abeer Alwan (S'82–M'92) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1992.

In 1992, she joined the Electrical Engineering Department, University of California, Los Angeles (UCLA) as an Assistant Professor, where she established the Speech Processing and Auditory Perception Laboratory. In 1996, she became an Associate Professor. Her research interests include modeling speech production and perception mechanisms and applying these models in speech coding, synthesis, and recognition systems.

Dr. Alwan is the recipient of the NSF Research Initiation Award (1993), the NSF Career Development Award (1995), the NIH FIRST Career Development Award (1994), and the UCLA-TRW Excellence in Teaching Award (1994). She is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She is a member of the Acoustical Society of America Technical Committee on Speech Communication, and the IEEE Signal Processing Technical Committees on Audio and Electroacoustics and on Speech Processing.