ELSEVIER

Speech Communication 40 (2003) 291–313

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise ☆

James J. Hant *, Abeer Alwan

Speech Processing and Auditory Perception Laboratory, Department of Electrical Engineering, School of Engineering and Applied Sciences, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90095, USA

Received 8 August 2001; received in revised form 8 February 2002; accepted 23 April 2002

Abstract

In this paper, a time/frequency, multi-look masking model is proposed to predict the detection and discrimination of speech-like stimuli in a variety of noise environments. In the first stage of the model, sound is processed through an auditory front end which includes bandpass filtering, squaring, time windowing, logarithmic compression and additive internal noise. The result is an internal representation of time/frequency "looks" for each sound stimulus. To detect or discriminate a signal in noise, the listener combines information across looks using a weighted d' detection device. Parameters of the model are fit to previously measured masked thresholds of bandpass noises which vary in bandwidth, duration, and center frequency (JASA 101 (1997) 2789). The resulting model is successful in predicting masked thresholds of spectrally shaped noise bursts, glides, and formant transitions of varying durations. The model is also successful in predicting the discrimination of synthetic plosive CV syllables in a variety of noise environments and vowel contexts.

1. Introduction

Background noise presents a challenging problem for a variety of speech and hearing devices including automatic speech recognition (ASR) systems, speech coders, and hearing aids. Since human listeners are extremely adept at perceiving speech in noise, a better understanding of human perception may help improve the robustness of current designs.

Traditional masking models (e.g. Fletcher, 1940; Patterson, 1976) focus on the masking of long duration, narrowband stimuli by noise. In these models, the signal and masker are filtered through an auditory filter that is centered around the signal's center frequency. If the filtered signal-to-noise ratio (SNR) is greater than a certain threshold, then the signal is heard. To predict the noise masking of a wide-band, non-stationary signal such as speech, however, the effects of both signal duration and bandwidth must be characterized (over a large frequency and duration range). In addition, for a model to predict perceptual confusions of speech sounds in noise, it should be able to predict the results of discrimination as well as detection experiments. To our knowledge, there is no published work which presents a masking model

☆ Portions of this paper were presented at Eurospeech '99 and ARO 2000. This paper is based on parts of Dr. James Hant's Ph.D. Disseration, UCLA, 2000.

* Corresponding author. Tel.: +1-310-336-1388.
E-mail address: james.j.hant@aero.org (J.J. Hant).

0167-6393/02/$ - see front matter
doi:10.1016/S0167-6393(02)00068-7

that can predict how such general stimuli are detected or discriminated in noise.

Traditionally, durational effects on masking have been modeled by placing a temporal integrator at the output of the auditory filter with the highest SNR (e.g. Hughes, 1946; Plomp and Bouman, 1959). To explain the drop in tone thresholds with duration, the time constant of the temporal integrator was set between 80 and 300 ms, which is significantly larger than the temporal resolution of the auditory system (e.g. Plack and Moore, 1991). In an attempt to account for this discrepancy, Viemeister and Wakefield (1991) suggested that durational effects could be described by a multi-look mechanism. They propose that, instead of integrating over a long time window, listeners consider multiple "looks" at a long-duration signal and combine information optimally to detect the signal. The multi-look hypothesis assumes listeners use a multi-dimensional detection mechanism in which they store information from each look in an internal buffer and consider all looks simultaneously to detect the signal. For an optimal combination of looks, the total detectability, d', is the Euclidean sum of the detectabilities for each look in time, d'_i (Green and Swets, 1966),

$$d' = \sqrt{\sum_{i=1}^{\mathrm{Nt}} (d'_i)^2}, \tag{1}$$

where d' is the overall detectability, d'_i is the detectability of the ith look in time, and Nt is the number of time looks.

Eq. (1), however, implies that an optimal combination of information results in a threshold decrease of sqrt(2) or 1.5 dB with the doubling of duration, while tone thresholds decrease by about 3 dB (Plomp and Bouman, 1959). To predict this 3 dB decrease in thresholds, Viemeister and Wakefield applied a weighting function to their detection device so that looks at the beginning of a signal would be weighted less than those at the end. However, auditory models which include adaptation (e.g. Zwislocki, 1969; Strope and Alwan, 1997) and emphasize signal onsets imply that the detectability of looks at the beginning of a signal should be *greater* than those at the end.

Since the multi-look (in time) model only uses information from one frequency channel, it cannot predict thresholds for non-stationary signals, such as glides. Experimental data show that glide thresholds drop nearly 3 dB with the doubling of duration (e.g. Nabelek, 1978; Collins and Cullen, 1978), while a single channel, multi-look model will predict very little change in glide thresholds across duration. More generally, since the multi-look (in time) model does not sum information across filter outputs, it is unable to predict threshold changes with signal bandwidth.

One model which can describe threshold changes with bandwidth is the multi-band excitation pattern model (e.g. Plomp, 1970; Florentine and Buus, 1981). In this model, the input signal is filtered by an auditory filter bank and statistically independent Gaussian (internal) noise is assumed to be present in each frequency channel. Assuming that information from each filter is combined optimally, then (analogous to Eq. (1)) the total detectability of a wide-bandwidth signal, d', is the Euclidean sum of the detectabilities of each frequency channel, d'_j,

$$d' = \sqrt{\sum_{j=1}^{\mathrm{Nf}} (d'_j)^2}, \tag{2}$$

where d' is the overall detectability, d'_j is the detectability of the jth channel, and Nf is the number of frequency channel.

The multi-band excitation model is successful in predicting intensity jnds for tones and wide-band noise signals (Florentine and Buus, 1981). The model, however, predicts threshold drops of 1.5 dB with the doubling of bandwidth, which is less than the 3 dB observed for the masking of (short-duration) bandpass noises and tone complexes (Hant et al., 1997; van den Brink and Houtgast, 1990). Data from both studies also show that the drop in masked thresholds with increasing bandwidth is dependent on signal duration. Specifically, at short durations (10 ms), *intensity* thresholds for bandpass noises and tone complexes are similar across bandwidth, while at long durations (300 ms), *spectrum-level* thresholds are similar across bandwidth. This trend, described by van den Brink

and Houtgast (1990) as more efficient spectral integration at short durations, cannot be predicted by the multi-band excitation model. Durlach et al. (1986) suggested adding correlated or central noise to the multi-band model. These modifications, however, predict a *reduced* drop in thresholds across bandwidth and cannot account for decreases in spectral integration at long durations.

To perceive speech-like signals which are wide-band, non-stationary and of varying durations, listeners may combine information across several filter outputs at different moments in time. To describe such a mechanism, a model that combines aspects of the multi-band excitation and multi-look (in time) models is proposed. In order to perceive a signal in background noise, it is assumed that the listener combines information from multiple looks in both time and frequency using a d' decision device. Recently, another model using time/frequency looks has been proposed to predict the detection and discrimination of Gaussian-windowed tones with varying amounts of spectral splatter (van Schijndel et al., 1999). In that model, it is assumed that listeners sum *energy* over a limited number of time/frequency looks when detecting a signal. At 1 kHz, for example, it is assumed that listeners use 50 looks which are 4 ms long, corresponding to an integration time constant of 200 ms. Although the model works well for Gaussian-windowed tones which cover a limited time/frequency region, it cannot be used to predict the masked threshold of any general wide-band, non-stationary, variable-duration stimulus. Such signals may contain more than 50 time/frequency looks, which cover a large frequency (and duration) range, and it is not clear which looks the model should consider. In addition, van Schijndel et al. use separate decision devices to predict detection and discrimination experiments, making model parameterization difficult and their model less general.

The model proposed in this paper uses a single decision device that takes into account all the time/frequency looks generated by any two stimuli and can thus be used to predict both masked detection and discrimination thresholds. For a detection experiment, for example, the decision device would compare time/frequency looks for the masker and signal plus masker stimuli, while for a discrimination experiment, the decision device would compare looks corresponding to the two signals being discriminated.

Parameters of the proposed multi-look model are fit to previously measured, noise-masked thresholds of bandpass noise signals which vary in duration, bandwidth, and center-frequency (Hant et al., 1997). The resulting model is then used to predict the masked thresholds of spectrally shaped noise bursts (such as those found at the release of plosive consonants), glides, and speech-like formant transitions. Finally, the model is used to predict the discrimination of synthetic voiced plosive consonants in both speech-shaped and perceptually flat noise.

2. Model description

The purpose of the multi-look, time/frequency masking model is to predict the detection and discrimination of signals in the presence of a noise masker; the signals could be wide-band or narrowband, stationary or not, and of any duration. Toward this end, the basic approach of signal detection theory is adopted. It is assumed that listeners develop internal representations for both stimuli presented. These representations are in the form of time and frequency looks generated by processing stimuli through an auditory front end. To detect or discriminate a signal in noise, subjects combine information across time/frequency looks.

2.1. Theoretical considerations

To better characterize how information is combined across time and frequency, a decision device is developed based on the noise-masked thresholds of bandpass noises which vary in duration, bandwidth and center frequency. Fig. 1 plots the masked thresholds of bandpass noises (centered at 1 kHz) as a function of duration with bandwidth as a parameter. These data were originally reported in (Hant et al., 1997). At short durations (10 ms), spectrum-level thresholds drop nearly 8 dB as the signal bandwidth increases from 1 to 8 critical bands (CBs). This drop is consistent

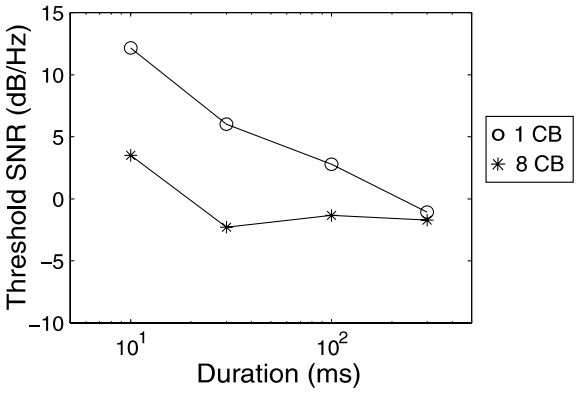

Fig. 1. Masked thresholds of 1 kHz bandpass noises in a flat noise masker. Spectrum-level thresholds (dB/Hz) are plotted as a function of signal duration with signal bandwidth (in CB) as a parameter. These data were originally reported in (Hant et al., 1997).

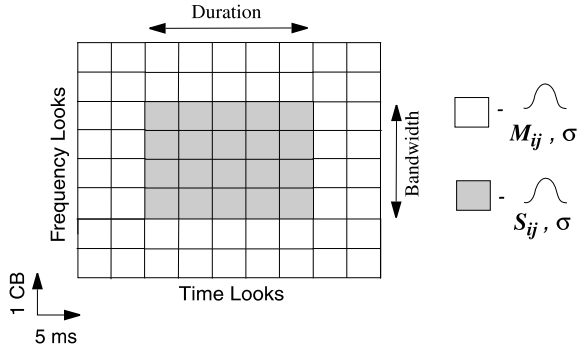

Fig. 2. Schematic of the time/frequency looks for the bandpass noise stimuli.

with an efficient sum of energy (or information) across frequency. At long durations (300 ms), however, spectrum-level thresholds are similar across bandwidth, consistent with a less efficient summation. Similarly, thresholds for 1 CB noises drop nearly 14 dB as the duration increases from 10 to 300 ms, while for the 8 CB noises, the drop is about 6.5 dB. These results are consistent with a more efficient sum of energy (or information) across time for narrow-bandwidth signals. Note, to reduce the effect of spectral splatter, all bandpass noise stimuli were turned on and off using a raised-cosine window with a rise/fall time of 1 ms. Similar trends have been observed for the noise masking of tone complexes (van den Brink and Houtgast, 1990).

A simple simulation is conducted to develop a decision device which can reproduce these trends. Assume that the bandpass noise stimuli are represented by a grid of time/frequency looks shown in Fig. 2. It is assumed that the level of each look is Gaussian-distributed with means M_{ij} and S_{ij} (for the masker and signal plus masker, respectively) and a common variance σ^2. Assuming that the variance is dominated by internal noise, the standard deviation, σ, is a free parameter and can be set to fit experimental data.

Under this framework, thresholds can be predicted using a d' decision device. Below, model predictions of the d' decision device are calculated and compared to the same model with logarithmic compression and a weighting function added. Specifically, the total detectability, d', is a Euclidean sum of the detectabilities for each time/frequency look, d'_{ij} (as shown in Eq. (3)).

$$d' = \sqrt{\sum_{i=1}^{\mathrm{Nt}} \sum_{j=1}^{\mathrm{Nf}} (d'_{ij})^2}, \tag{3}$$

where

$$d'_{ij} = \frac{|S_{ij} - M_{ij}|}{\sigma},$$

M_{ij} is the masker level for each ith, jth look, S_{ij} is the signal plus masker level for each ith, jth look, and σ is the standard deviation for each look.

Predictions for the bandpass noise data are shown in Fig. 3(a). Although an optimal sum of information predicts decreases in thresholds with increasing bandwidth and duration, the magnitude of the changes are much smaller than that for the experimental data. In addition, predictions are not consistent with a decrease in detector efficiency at the long durations and wide-bandwidths, showing similar threshold drops across bandwidth at all durations.

If the decision device is applied after logarithmic compression, however, the magnitude of the threshold drops will increase. Thresholds are again determined by Eq. (3) except that d'_{ij} is calculated after the masker and signal plus masker energies have been logarithmically compressed as shown in Eq. (4). Assuming the variance of each look is

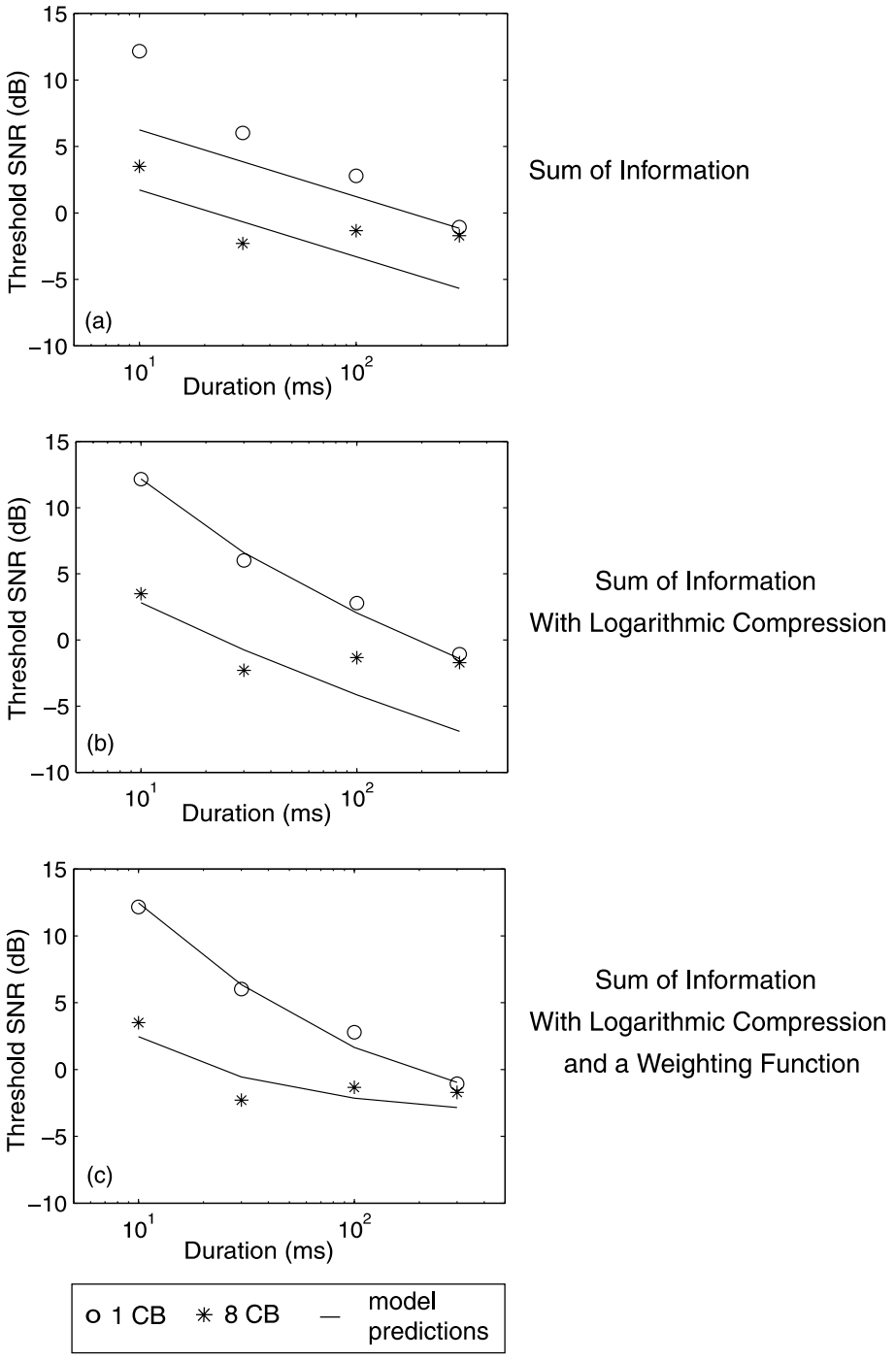

Fig. 3. Theoretical model predictions for the 1 CB noise data using a d' decision device. (a) Sum of information across time and frequency, (b) sum of information after logarithmic compression, (c) sum of information with logarithmic compression and a weighting function.

dominated by internal noise which is added after logarithmic compression, σ is a free parameter (different from that in Eq. (3)) and can be set to fit the experimental data.

$$d'_{ij} = \frac{|\log_{10}(S_{ij}) - \log_{10}(M_{ij})|}{\sigma}, \tag{4}$$

where σ is the standard deviation for the logarithmic model.

Results are shown in Fig. 3(b). The logarithmic model is able to predict an increased drop in thresholds across both duration and bandwidth. The reason for this increased drop is illustrated in Fig. 4 which plots d' versus SNR for the 1 CB noises, with signal duration as a parameter. Thresholds are determined by the intersections of the d' curves with the horizontal line at 0.66 dB (corresponding to 79% in a two AFC task). For the linear model, d' (in dB) increases linearly with SNR, while for the logarithmic model, d' values are compressed at the higher SNRs. This compression results in larger threshold drops across duration for the log model (13.6 dB) compared to the linear model (7.4 dB). With nearly a 3 dB drop in thresholds with the doubling of duration, the logarithmic model may alleviate the need to apply different weights to looks in time as proposed by Viemeister and Wakefield (1991).

Although the log model can predict the magnitude of the threshold drops, it cannot predict a decrease in detector efficiency at the wide-bandwidths and long durations, underestimating the 8 CB thresholds at 100 and 300 ms. To describe a decrease in detector efficiency, it is assumed that listeners only "pay attention" to looks whose difference in means (between the masker and signal plus masker) is above a certain value, θ. This mechanism is implemented by adding a weighting function w to the d' detection device. This function applies no weight to looks where the absolute difference in level ($\log_{10}(S_{ij}) - \log_{10}(M_{ij})$) is below a certain threshold, ensuring that regardless of signal duration or bandwidth, thresholds do not drop below a particular spectrum-level SNR.

$$d' = \sqrt{\sum_{i=1}^{\mathrm{Nt}} \sum_{j=1}^{\mathrm{Nf}} w(|\log_{10}(S_{ij}) - \log_{10}(M_{ij})|)(d'_{ij})^2}, \tag{5}$$

where

$$w(|\log_{10}(S_{ij}) - \log_{10}(M_{ij})| = \begin{cases} 1 & \text{if } |\log_{10}(S_{ij}) - \log_{10}(M_{ij})| > \theta, \\ 0 & \text{if } |\log_{10}(S_{ij}) - \log_{10}(M_{ij})| < \theta. \end{cases}$$

With the weighting function added, the model is able to predict a decrease in spectral integration at long durations (see Fig. 3(c)). Note that for the three decision devices shown in Fig. 3, the standard deviation, σ, determines the relative level of all thresholds while the drop of thresholds across duration and bandwidth is determined by how the

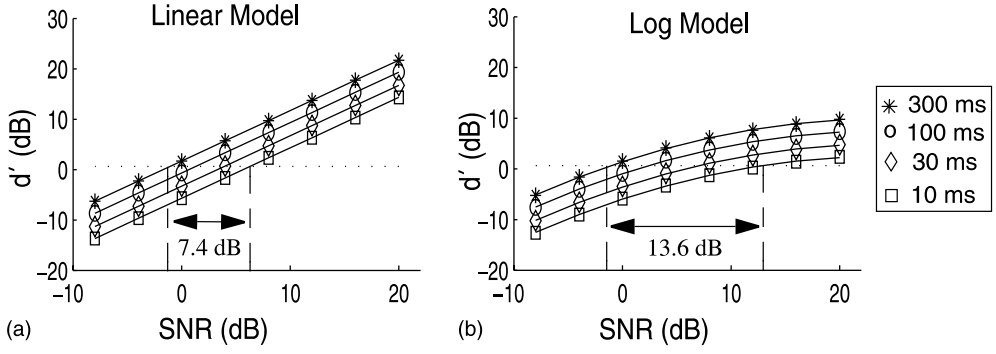

Fig. 4. d' versus SNR for the time/frequency decision device. Model predictions for d' are plotted as a function of SNR with signal duration as a parameter. The horizontal line is at a d' of 1.16 (0.66 dB) and corresponds to a threshold of 79% correct in a two AFC procedure. Results are shown for the (a) linear and (b) logarithmic model. The numbers between the dashed lines correspond to predicted threshold drops between the 10 and 300 ms data.

difference in level between the signal plus masker and masker is calculated (and weighted). In the next section, the weighted d' detection device will be parameterized and used to predict the bandpass noise data at several center frequencies, bandwidths and durations.

2.2. Model overview

To predict masked thresholds, time/frequency looks are first generated for the masker and signal plus masker, by processing both stimuli through an auditory front end. The mean and standard deviation for each time/frequency look is calculated for the masker and signal plus masker, over a range of SNRs. Using the weighted d' decision device (described in Eq. (5)), d' is calculated as a function of SNR. Finally, thresholds are determined by the SNR at which d' equals a particular value.

2.3. Auditory front end

The stages of processing for the auditory front end are shown in Fig. 5.

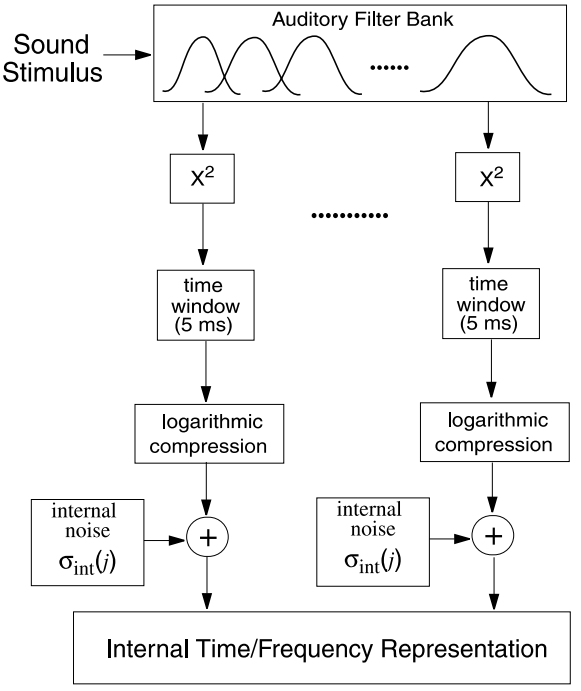

Fig. 5. The auditory front end.

The sound stimulus is first filtered through a bank of auditory filters, whose shapes are determined from previous masking experiments (Glasberg and Moore, 1990). Each filter has a frequency response, $W(g)$, described by the roex function in Eq. (6a), and an equivalent rectangular bandwidth (ERB) which varies with center frequency (c_f) as given by Eq. (6c). The filter bank contains 30 filters, with center frequencies ranging from 105 to 7325 Hz and separated by 1 ERB. To save computation time, narrow-bandwidth signals were processed through a subset of the 30 filters.

$$W(g) = (1 + pg)e^{-pg}, \tag{6a}$$

where

$$p = \frac{3.35c_f}{\mathrm{ERB}} \tag{6b}$$

and

$$\mathrm{ERB} = 24.7(4.37c_f + 1). \tag{6c}$$

For simplicity, filters are assumed to be symmetric and level-independent. The filter bank is implemented by convolving the input signal with a set of (fourth-order) gammatone filters that have frequency responses described by Eqs. (6a)–(6c) and phase responses similar to those measured for the basilar membrane (Patterson et al., 1992). The gammatone impulse responses are sampled at a rate of 16 kHz and truncated to a duration of 100 ms.

The output of each filter is then squared and processed through a temporal integrator every 5 ms. The shape of the temporal window is shown in Fig. 6. The window has a flat section of 4 ms with a raised cosine of 1 ms on each side, yielding an equivalent rectangular duration (ERD) of 5 ms. Previous studies have used roex-shaped windows with ERDs between 3.8 and 6.6 ms to predict the detection of brief, intensity decrements in a wideband noise (Plack and Moore, 1991) and the discrimination of sinusoidal signals which change in frequency continuously from those that change in a series of discrete steps (Madden, 1994).

In this study, smooth, overlapping windows are used to reduce the spectral splatter for each time/

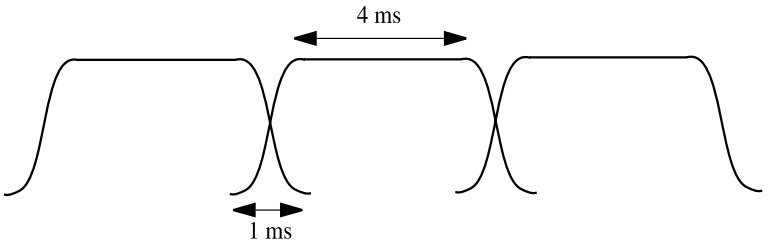

Fig. 6. Shape of the temporal window.

frequency look. This overlap, however, results in correlations between looks in time. A flat-window with raised-cosine skirts is used instead of a roex shape to reduce these correlations. In addition, a sum of the overlapping windows shown in Fig. 6 applies an equal weighting across the duration of the stimulus, while a sum of overlapping roex windows does not.

The output of each temporal window is logarithmically compressed, and independent Gaussian noise is added to each time/frequency look. Logarithmic compression, consistent with Weber's law of incremental loudness, has been used to predict the intensity discrimination of noise signals (e.g. Green, 1960; Raab and Goldberg, 1975). The internal noise is a first-order approximation of the stochastic nature of neural encoding in the auditory system. The variance of this noise (σ^2_{int}) is allowed to vary with center frequency, but not with duration or bandwidth.

2.4. Model statistics

In order to use the d' detection device described in Eq. (5), each time/frequency look must be Gaussian-distributed and statistically independent from the other looks. To assess the validity of these assumptions, 500 flat-noise samples were processed through the model and their statistics were measured. Fig. 7(a) plots the distribution for a single time/frequency look (corresponding to the output of the filter centered at 1 kHz) with no internal noise added. Despite several levels of processing by the model, some of which are non-linear, the distribution for a single look is reasonably approximated by a Gaussian. The standard deviation for each time/frequency look (due to external noise) ranges from 4.5 dB at 100 Hz to 1.5 dB at 7500 Hz.

To quantify the correlation between looks, each 2D, time/frequency matrix (X) generated by the noise samples is column ordered into a vector ($\boldsymbol{x}$). An example of this column ordering is shown in Eq. (7).

$$X = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \dots & \dots & \dots & \dots \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{bmatrix} \quad \begin{matrix} \text{time looks} \rightarrow \\ \text{frequency looks} \downarrow \end{matrix} \tag{7}$$

$$x^{\mathrm{T}} = [\, a_{11} \dots a_{M1} \quad a_{12} \dots a_{M2} \quad \dots \quad a_{1N} \dots a_{MN} \,].$$

Correlation functions for a single time/frequency look centered at 1 kHz are shown in Fig. 7(b). The width of the central peak represents correlations between looks in frequency, while the two side peaks show correlations between looks in time. The only significant correlations are for looks which are directly adjacent in frequency and time, resulting from time and frequency windows which are slightly overlapping. However, if independent internal noise is added to each look (as shown in Fig. 7(c)), correlations between looks become nearly negligible.

2.5. Decision device

Assuming that time/frequency looks are Gaussian and statistically independent, the internal representation for any stimulus can be represented

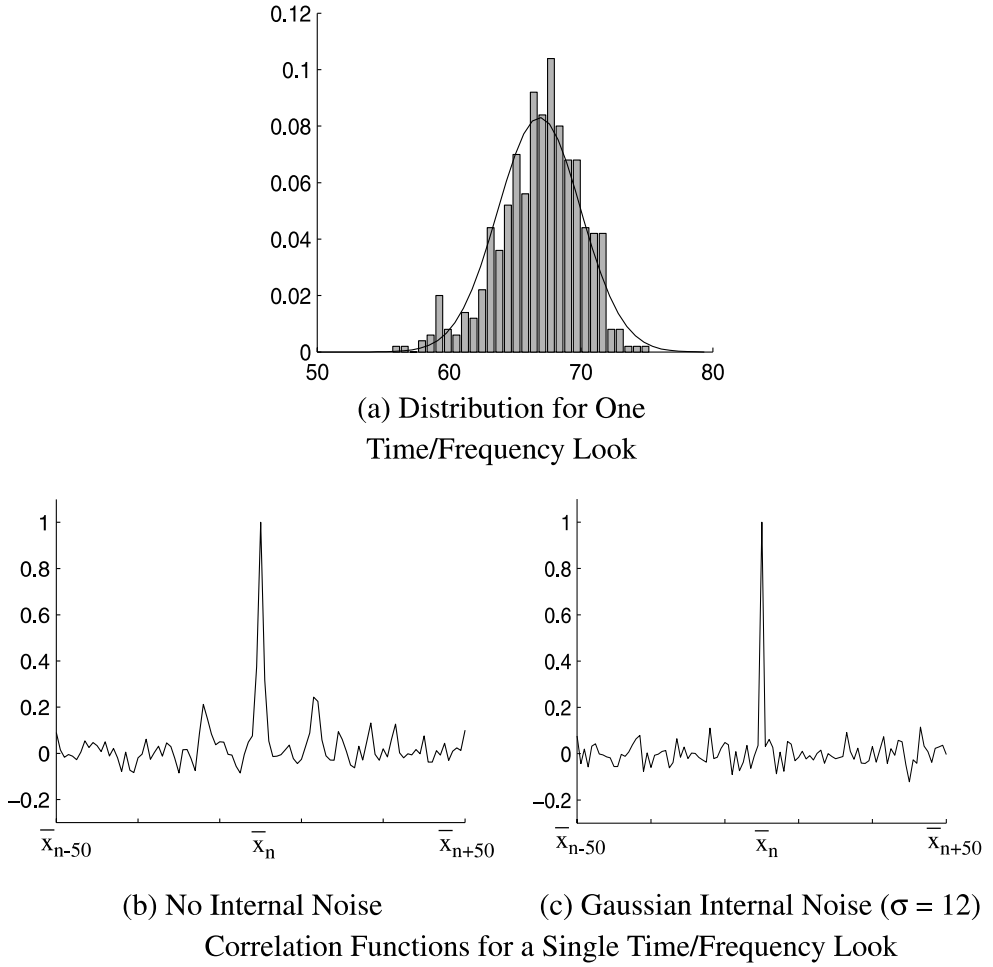

Fig. 7. Model statistics using flat noise as input. (a) Distribution for a single time/frequency look at the output of the filter centered at 1 kHz. (b,c) Correlation functions for a single time/frequency look both without and with added Gaussian internal noise ($\sigma = 12$ dB), respectively.

by a matrix of means and variances; i.e. (μm_{ij}, σm_{ij}^2) and (μs_{ij}, σs_{ij}^2) for the masker and signal plus masker, respectively. Note that means and variances are calculated after logarithmic compression and the variance for each look is a sum of the variances due to external and internal noise. By using a common variance for both the masker and signal plus masker, the detection device described by Eq. (5) can be implemented as follows:

$$d' = \sqrt{\sum_{i=1}^{\mathrm{Nt}} \sum_{j=1}^{\mathrm{Nf}} w_j(|\mu s_{ij} - \mu m_{ij}|)(d'_{ij})^2}, \tag{8}$$

where

$$d'_{ij} = \frac{|\mu s_{ij} - \mu m_{ij}|}{\sqrt{\left(\frac{\sigma s_{ij}^2 + \sigma m_{ij}^2}{2}\right)}}$$

and

$$w_j(|\mu s_{ij} - \mu m_{ij}|) = \begin{cases} 1 & \text{if } |\mu s_{ij} - \mu m_{ij}| > \theta(j), \\ 0 & \text{if } |\mu s_{ij} - \mu m_{ij}| < \theta(j). \end{cases}$$

Here, the common variance used for each time/frequency look is approximated by the average of the variances for the masker and signal plus masker. Assuming that the variances for the masker

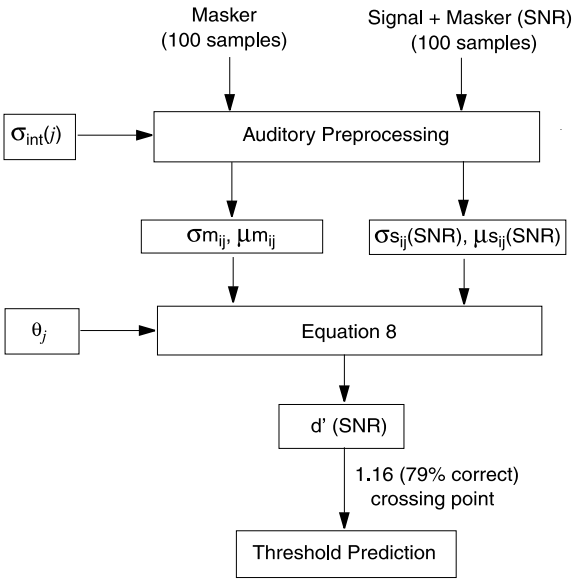

Fig. 8. Schematic of the method for predicting thresholds.

and signal plus masker are dominated by internal noise (which is the same for both), this approximation is fairly accurate. Recall that the weighting function, w_j, and parameter $\theta(j)$, are allowed to vary with the frequency of the look, j.

Using Eq. (8), masked thresholds can be predicted for any wide-band, non-stationary stimulus. A schematic of the prediction method is shown in Fig. 8. To predict thresholds, 100 examples of the masker and signal plus masker stimuli are first processed through the auditory front end at different SNRs. Note that the SNR is simply defined as the total signal power divided by the total noise power. From the 100 examples, means (μm_{ij} and μs_{ij}) and standard deviations (σm_{ij} and σs_{ij}) are calculated for each time/frequency look. Using these values, the d' at each SNR is calculated using Eq. (8). Finally, the SNR at which d' equals 1.16 (corresponding to 79% correct for a two AFC procedure), is defined to be the threshold.

3. Model fit to bandpass noise data

3.1. Parameter fit

There are two free parameters in the model: the standard deviation of the internal noise, $\sigma_{\text{int}}(j)$, and a weighting parameter, $\theta(j)$. Both parameters are allowed to vary with center frequency (but not with duration or bandwidth). Specifically, $\sigma_{\text{int}}(j)$ determines the absolute level of all thresholds at a particular center frequency, while $\theta(j)$ determines the SNR at which thresholds no longer decrease. The drop in thresholds with increasing bandwidth and duration is described by an increase in the number of time/frequency looks that subjects can use for detection.

Using the method outlined in Fig. 8, masked thresholds were predicted for bandpass noises of varying center frequency (0.4, 1.0, 2.0, 3.0 and 4.0 kHz), duration (10, 30, 100 and 300 ms) and bandwidth (1, 2, 4 and 8 CBs). These thresholds were originally reported in (Hant et al., 1997). At each center frequency, j, parameters $\theta(j)$ and $\sigma_{\text{int}}(j)$ were adjusted in an iterative procedure to minimize the mean-squared error between the model predictions and 16 data points (4 bandwidths $\times$ 4 durations). Parameter estimates, as a function of center frequency, were then fit to sigmoidal-shaped curves.

Fig. 9 plots the best fit parameters $\sigma_{\text{int}}(f)$ and $\theta(f)$, as a function of ERB number, f, along with the sigmoidal fits.

The equations for the sigmoidal fits are

$$\sigma_{\text{int}}(f) = 16.62 + 7.88\left(-\frac{1}{2} + \frac{1}{2}\frac{(1-\exp(f-21.80))}{(1+\exp(f-21.80))}\right), \tag{9}$$

$$\sigma_{\text{int}}(f) = 3.81 + 2.39\left(-\frac{1}{2} + \frac{1}{2}\frac{\left(1-\exp\left(\frac{f-16.15}{14.0}\right)\right)}{\left(1-\exp\left(\frac{f-16.15}{14.0}\right)\right)}\right), \tag{10}$$

where f is the ERB number.

Note that the sigmoidal curves were fit after $\sigma_{\text{int}}(f)$ and $\theta(f)$ had been determined at each center frequency, and were meant to interpolate parameters so that the model could be used to predict thresholds for signals over a continuous frequency range.

Notice the sharp drop in internal noise, $\sigma_{\text{int}}(f)$, for ERBs numbered 18–25. This decrease is needed

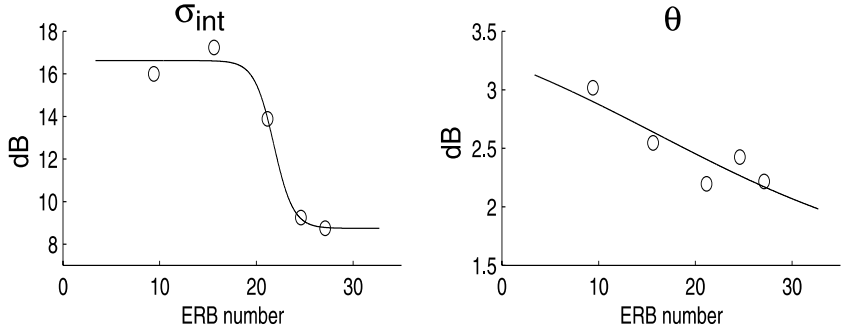

Fig. 9. Best fit parameters σ_{int}, θ corresponding to bandpass noises with center frequencies of 0.4, 1, 2, 3, 4 kHz (9.4, 15.6, 21.2, 24.5, 27.1 ERB) are plotted as a function of ERB number and denoted by circles. Sigmoidal fits to these parameters are shown by the solid lines.

to predict the drop in spectrum-level thresholds for frequencies higher than 1 kHz. Interestingly, this decrease in σ_{int} occurs around the frequencies where phase locking (and thus the coding of fine-time information) becomes less prominent (Kiang et al., 1965). Such a decrease is physiologically plausible, if one assumes the amount of internal noise for coding signal energy is inversely related to the proportion of fibers that are coding rate, as opposed to fine-time information. At frequencies below 2 kHz, where phase locking in thought to be strongest, a fraction of the fibers may be delegated to coding temporal information and thus, one would expect a larger internal noise for coding rate (or energy) information. At frequencies above 3 kHz, where phase locking is thought to be weaker, a majority of fibers will be coding rate (or energy) information and one would expect less internal noise. The values for $\sigma_{\text{int}}(f)$ fall within the range used by Farar et al. (1987) to predict the discrimination of synthetic plosive bursts in background noise (21.7 dB at 10 ms to 3.6 dB at 300 ms).

Fig. 9 also shows a slight decrease in the threshold parameter, $\theta(f)$, with increasing center frequency. This drop is needed to describe a slight decrease in the wide-bandwidth, long-duration thresholds at the higher center frequencies.

3.2. Results and discussion

Fig. 10 shows the experimental data and model predictions at each center frequency. Spectrum-level thresholds are plotted versus duration, with signal bandwidth as a parameter. Model predictions are shown by the solid lines.

The model successfully predicts a decrease in thresholds with increasing signal duration and bandwidth, even though model parameters do not vary across either dimension. In addition, the model successfully predicts smaller threshold drops across duration at the wide-bandwidths, and smaller threshold drops across bandwidth at the long durations. Mean-squared errors for model predictions are between 0.322 and 0.493 dB.

The model, however, underpredicts the decrease in thresholds between 10 and 30 ms, especially at the higher center frequencies. These threshold drops, which range from 6–9 dB, are larger than those predicted by either a multi-look (Viemeister and Wakefield, 1991) or an integration model (Plomp and Bouman, 1959). Instead, it appears that the bandpass noise thresholds at 30 ms may be the result of a mechanism which is not solely based on the signal's energy. Perhaps, as was suggested in (Hant et al., 1997), signal transients play a role. More experiments are necessary to quantify such a mechanism.

Previous models of temporal and spectral integration are unable to predict all the trends in the data. A multi-band excitation pattern model, for example, which uses frequency channels that have been processed through a temporal integrator (e.g. Durlach et al., 1986), can successfully predict drops in thresholds with increasing duration and bandwidth. However, at all durations, the model predicts threshold drops of 1.5 dB per doubling of

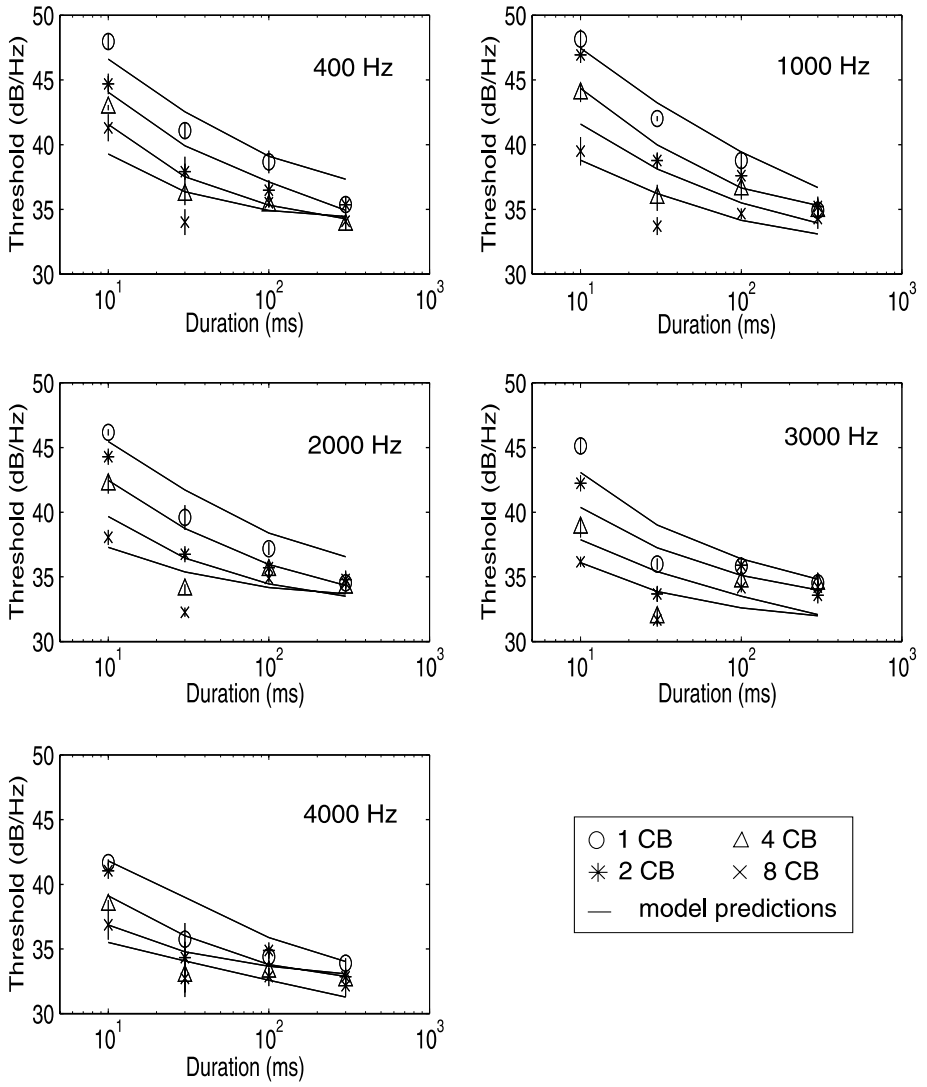

Fig. 10. Model predictions of the bandpass noise thresholds reported in (Hant et al., 1997). Spectrum-level thresholds (in dB/Hz) for bandpass noises centered at 0.4, 1, 2, 3 and 4 kHz are shown. A flat noise masker with a spectrum level of 36 dB/Hz was used. Thresholds, averaged across four subjects, are plotted as a function of signal duration with signal bandwidth (in CB) as a parameter. The standard deviations across subjects are expressed by the error bars. Model predictions are shown by the solid lines.

bandwidth, inconsistent with the data. Similar errors will occur if duration-dependent internal noise is added to each frequency channel (Farar et al., 1987).

Hant et al. (1997) described the bandpass noise data in terms of a traditional filter-SNR model in which the "effective" bandwidth of each filter was duration dependent. If filters are broad at short durations, then subjects will sum signal energies over a wide frequency region and intensity thresholds will be similar across bandwidth. If filters are narrow at long durations and the filter with the highest SNR is used to detect the signal, then spectrum-level thresholds will be similar across bandwidth. van den Brink and Houtgast (1990) found similar bandwidth and durational

effects for the masking of tone complexes and parameterized the data in terms of an increase in spectral integration at short durations and in temporal integration at narrow bandwidths. In the current approach, these bandwidth and durational trends are the result of an evolving statistical estimate of the signal, using time/frequency looks.

The advantage of the current time/frequency model is that the duration (or bandwidth) of the signal does not have to be known a priori in order to predict masking thresholds. Durational and bandwidth effects are accounted for implicitly in the model. In addition, since the model uses information from multiple filter outputs, at varying moments in time, it can be used to predict the masked threshold of wide-band and non-stationary stimuli of varying durations.

4. Model predictions of masked thresholds for tone glides and formant transitions

4.1. Stimuli and experimental procedure

With parameters fit to the bandpass noise data, the model was then used to predict the masking of non-stationary stimuli. Masking experiments were first conducted using rising and falling tone glides and formant transitions which varied in final frequency, frequency extent, and duration. A schematic of these stimuli is shown in Fig. 11.

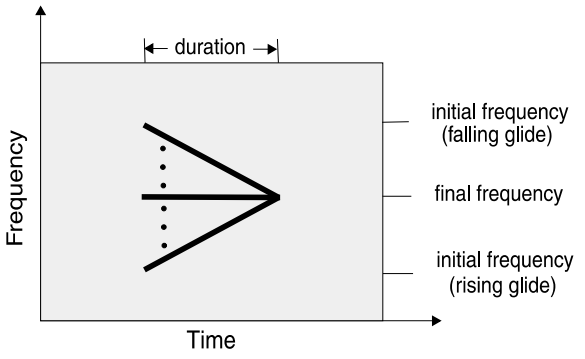

Fig. 11. Schematic of the glide and formant-transition stimuli. Three final frequencies (500, 1500 and 3500 Hz) and three durations (10, 30 and 100 ms) were tested. Frequency extents were based on the ERB frequency scale (Glasberg and Moore, 1990) and defined as the initial frequency minus the final frequency.

Table 1
Initial and final frequencies for the glide and formant-transition stimuli

Final frequency (Hz)	−3 ERB (Hz)	−1.5 ERB (Hz)	0 ERB (Hz)	1.5 ERB (Hz)	3 ERB (Hz)
500	299	399	500	628	778
1500	1023	1242	1500	1803	2159
3500	–	2944	3500	4153	–

Three final frequencies (500, 1500 and 3500 Hz) and three durations (10, 30 and 100 ms) were tested. Frequency extents were based on the ERB frequency scale (Glasberg and Moore, 1990) and defined as the initial frequency minus the final frequency. At final frequencies of 500 and 1500 Hz, frequency extents of (−3, −1.5, 0, 1.5, 3) ERBs were tested, while at 3500 Hz, frequency extents of (−1.5, 0, 1.5) ERBs were tested. Table 1 shows the initial and final frequencies of all stimuli. Rates of frequency change range from 0 to 65.9 Hz/ms. To reduce the effect of spectral splatter, signals were turned on and off using a raised-cosine window with a rise/fall time of 1 ms.

Single formant transitions were generated in MATLAB using the overlap-and-add method. An impulse train, with an F0 of 100 Hz, was filtered with second-order resonators that had center frequencies (and bandwidths) corresponding to a specific portion of the formant trajectory. These time-slices were added together using overlapping raised-cosine windows with rise/fall times of 2 ms. The 500 and 3500 Hz formant trajectories had approximate bandwidths of 60 Hz (0.76 ERB) and 200 Hz (0.5 ERB), respectively. The 1500 Hz stimuli had an approximate bandwidth of 186.6 Hz (1 ERB).

The masker used in the experiments was perceptually flat noise, with a level of 56 dB per ERB (total level of 71.2 dB SPL). Fig. 12 shows the spectrum for this masker. The masker duration was 750 ms and all signals were centered in time with respect to the masker.

Four subjects (two males, two females) with normal hearing participated in the experiments. Subjects ranged in age from 19 to 27 years. Stimuli were presented diotically to listeners in a sound attenuating room via Telephonics TDH49P

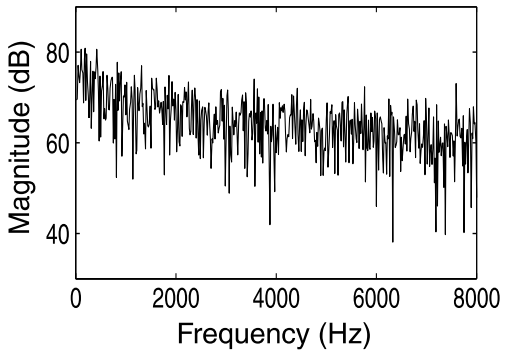

Fig. 12. Spectrum for the perceptually flat noise masker.

headphones. Computer software generated the test tokens as 16 bit/16 kHz digital numbers. An Ariel ProPort 656 board performed digital-to-analog conversion. The resulting analog waveforms were amplified using the pre-amp of a Sony 59ES DAT recorder, which was connected to the headphones. The entire system was calibrated within $+/-0.5$ dB before each experiment using a Larson Davis 800B Sound Level Meter.

Masked thresholds were determined using an adaptive 2I, two AFC paradigm with no feedback (Levitt, 1971). Three correct responses determined a successful sub-trial while one incorrect response determined an incorrect sub-trial. Thresholds, therefore, are defined to be the 79% correct points. Step sizes were initially set to 4 dB, then reduced to 2 dB after the first reversal, and finally to 1 dB after the third reversal. From a total of 9 reversals, the average of the last 6 determined the threshold for each trial. The mean of two trials determined the final threshold. Subjects were trained for 2 h before beginning the experiments. No training effects were apparent in the final data. Threshold predictions were generated using the method outlined in Fig. 8.

4.2. Results and model predictions

Experimental results and model predictions are shown in Fig. 13. On the left side of the figure, glide thresholds are plotted as a function of frequency extent with signal duration as a parameter. The corresponding formant thresholds are plotted on the right side of the figure. Thresholds are averaged across four subjects with standard deviations represented by the error bars. Model predictions are shown by the solid lines.

4.2.1. Experimental results

Experimental results show an interesting trend: over the range of frequency extents and durations tested, thresholds are only dependent on the duration of the stimulus, and not the frequency extent. At frequencies of 500 and 1500 Hz, the threshold drop between 10 and 100 ms is close to the 10 dB predicted by an (efficient) integration of signal energy across duration. At 3500 Hz, this threshold drop is slightly smaller, a trend which is consistent with the masking of tones in noise (Plomp and Bouman, 1959) and can be predicted by a decrease in the integration time constant at the higher frequencies.

The current data, however, are not consistent with those of Collins and Cullen (1978) which showed thresholds for 200–700 Hz and 1200–1700 Hz glides to be 4 dB greater than for corresponding (steady) tones. They also found that between durations of 10 and 35 ms, rising glides were more easily detectable than falling glides, but later showed that this asymmetry is only significant for rates of frequency change >96 Hz/ms (Cullen and Collins, 1982). Nabelek (1978) reported similar trends, but only at large frequency extents (>750 Hz) and short durations (<50 ms). For frequency extents of 200 Hz, Nabelek measured similar thresholds for both glides and tones, which is consistent with the current data. The reason for the discrepancy between the Collins and Cullen (1978) study and the current one is not clear, but may be partially due to the method for estimating thresholds (Alternate Forced Choice in the current study versus Method of Adjustment in the previous studies).

At final frequencies of 500 and 3500 Hz and a duration of 100 ms, formant thresholds are about 1 dB larger than for the corresponding glide thresholds. At 1500 Hz, this difference is greater, approaching 2–3 dB. The small differences between glide and formant thresholds may be attributed to differences in bandwidth. The spread of excitation for formant transitions will be larger than for the corresponding glides, which may re-

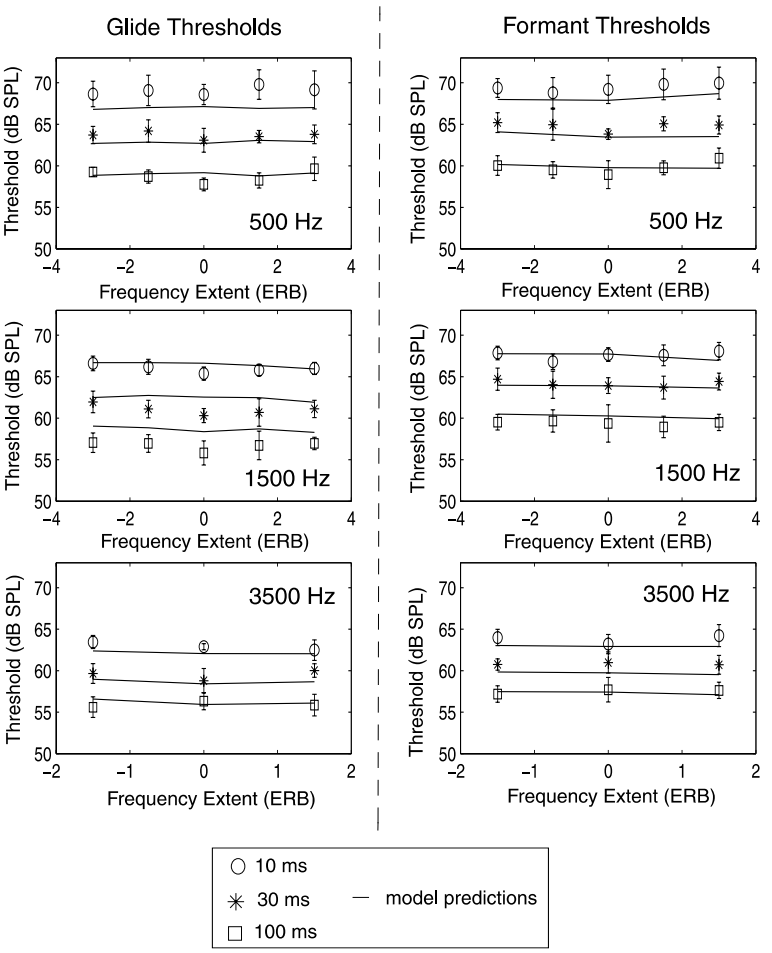

Fig. 13. Masked thresholds and model predictions for glides and single formant transitions. On the left side of the figure, thresholds for glides with varying final frequency (0.5, 1.5, 3.5 kHz) are plotted as a function of frequency extent (in ERB) with signal duration as a parameter. The corresponding formant thresholds are plotted on the right side of the figure. Thresholds are averaged across four subjects with standard deviations represented by the error bars. Model predictions are shown by the solid lines.

sult in a smaller filter-SNR. If energy is not summed efficiently across filter outputs (which is the case for the 100-ms stimuli) this will result in slightly larger thresholds. The larger differences between glide and formant thresholds seen for the 1500 Hz data may be the result of differences in the fine-time temporal structure between these two signals, resulting from the fact that formants are modulated by the fundamental frequency while the tone glides are not. Fine-time temporal cues, which may be more prominent for the glides than for the corresponding formant transitions, could result in a higher detectability for the glides at lower SNRs.

4.2.2. Model predictions

As shown in Fig. 13, the multi-look, time/frequency detection model, with parameters fit to previous bandpass noise data, is successful in capturing the general trends in the current data, predicting thresholds which are independent of frequency extent and decrease by about 9 dB between durations of 10 and 100 ms. The model is also successful in predicting smaller thresholds with increasing frequency.

However, threshold predictions for the 1500 Hz, 100-ms glides are about 1–3 dB higher than the experimental data. The reason for this error is

not clear. The model is successful in predicting formant thresholds in the same frequency region. Perhaps, subjects are using fine-time temporal cues to detect glides at 1500 Hz, which are not present in the formant transitions. There is also a slight underestimation of thresholds for the 10-ms glides and formants at 500 Hz. Fig. 10 shows a similar error for the 1 CB, 10 ms bandpass noise data at 400 Hz.

Recent discrimination experiments using FM stimuli suggest that short duration, non-stationary signals, such as formant transitions, may be coded by a place-rate mechanism (Moore and Sek, 1998; Madden and Fire, 1996; Sek and Moore, 1995). The place-rate mechanism assumes auditory signals are coded by the total rate (or energy) of neural firing at the output of a particular auditory filter or "place" along the basilar membrane. The fine temporal details of each filter output are not considered. The success of the time/frequency detection model in predicting the masking of glides and formant transitions, is further support for the place-rate mechanism. With the exception of the 100-ms, 1500 Hz glides, masking thresholds can be predicted by a model which is purely based on the signal's distribution of energy across frequency and time.

5. Model predictions of masking thresholds for synthetic plosive bursts

The multi-look model was also used to predict previously measured, masked thresholds of synthetic plosive bursts at four durations (10, 30, 100 and 300 ms) (Hant et al., 1997). The spectra of these bursts are shown in Fig. 14.

Note, to reduce the effect of spectral splatter, the burst stimuli were turned on and off using a raised-cosine window with a rise/fall time of 1 ms. Both the front and back /k/ burst have compact spectral peaks, while the /p/ and /t/ bursts have broader spectral peaks concentrated at the low and high frequencies, respectively. Fig. 15 plots the masked thresholds and model fits for these stimuli as a function of signal duration. Masked thresholds, averaged across three subjects, are denoted by the circles, with error bars representing the standard deviations across subjects. Model predictions are shown by the solid lines.

The model predicts thresholds well, with a maximum error of around 2 dB. The model successfully predicts the large durational dependence of burst thresholds for back /k/ and the smaller changes in threshold for /p/ and /t/. These durational dependencies are expected. Since the burst for a back /k/ has a relatively narrow spectral peak, its threshold response is similar to the narrow-bandwidth noises, showing a large drop across duration. The /t/ and /p/ bursts, on the other hand, have wider spectral peaks, and thus have similar thresholds to the wide-bandwidth noises, showing smaller drops across duration (see Fig. 10).

At 10 ms, which is about the duration of a naturally spoken burst, threshold predictions are 65–67 dB for /k/, about 70 dB for /p/, and 62 dB for /t/. These predictions can be explained by the parameter fits shown in Fig. 9. Since the /t/ burst has most of its energy concentrated at the high frequencies, it will be subject to a lower internal noise and will thus have a lower threshold than the /p/ burst, which has most of its energy concentrated at the lower frequencies.

The model's success at 30 ms is somewhat surprising. Model fits to the bandpass noise data (see Fig. 10) show that the model has difficulty predicting the "kink" in bandpass noise thresholds at 30 ms. These kinks, however, do not appear in the burst thresholds. Instead, it appears that the mechanisms responsible for the drop in bandpass noise thresholds between 10 and 30 ms, do not play an important role in the masking of plosive bursts.

The model, however, slightly overestimates the back /k/ thresholds at 100 and 300 ms and underestimates the front /k/ thresholds at 10 and 30 ms. These errors are similar to those seen for the 1 CB bandpass noise thresholds between center frequencies of 1 and 3 kHz (see Fig. 10). The errors in predicting the /k/ burst thresholds may also be due to the fact that the spectral peaks of these stimuli occur in a frequency region where the internal noise, σ_{int}, changes most drastically (see Fig. 9). Any slight errors in the estimation of parameters in this region could have a large effect on threshold predictions.

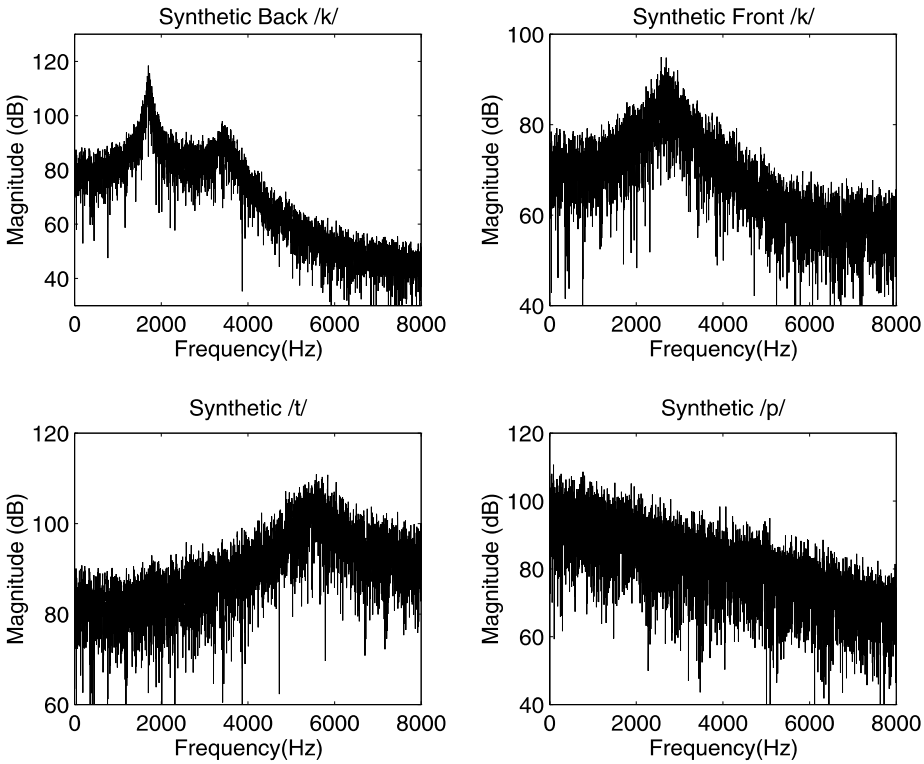

Fig. 14. DFT spectra for synthetic plosive bursts (Hant et al., 1997).

6. Model predictions of the discrimination of synthetic CVs in noise

In Sections 4 and 5 the multi-look model was successful in predicting the noise-masked thresholds of the two main acoustic cues for identifying plosive consonants, namely bursts and formant transitions. In this section the model is used to predict the discrimination of synthetic plosive CV syllables in three vowel contexts (/a/, /i/ and /u/) and in two types of noise maskers (perceptually flat and speech shaped).

In background noise, subjects were presented with two reference CV stimuli and one test CV stimulus in random order. Subjects were then forced to decide whether the test stimulus occurred first, second, or third. Experiments were conducted for CVs both with and without the burst cue. Schematized spectrograms of the /Ca/ stimuli (with no burst) used in one such experiment are shown in Fig. 16. By calculating d' using time/frequency looks for the two CV syllables being discriminated, the model described in Section 2 could be expanded to predict the results of the discrimination experiments.

6.1. Synthesis of the experimental stimuli

The 4-formant transitions were synthesized in MATLAB by the overlap and add method. An impulse train (with an F0 of 100 Hz) was first filtered with four second-order resonators in cascade that had center frequencies (and bandwidths) corresponding to a specific portion of the (F1 through F4) formant trajectory. These time-slices were then added together by using overlapping raised-cosine windows with rise–fall times of 2 ms. Each window overlapped by 1 ms. For the /a/ and /u/ contexts, the bandwidths for the four resonators were 60, 90, 150 and 200 Hz, corresponding to

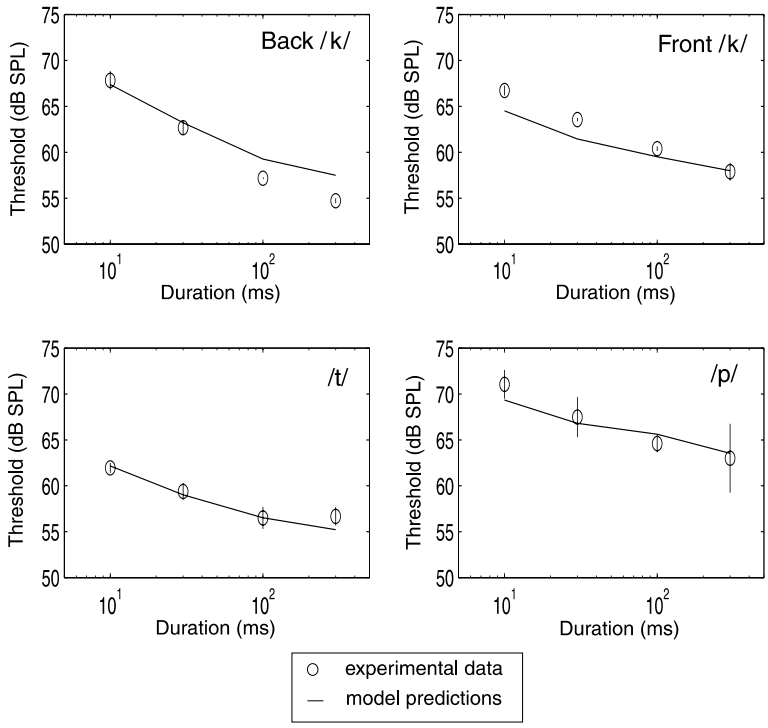

Fig. 15. Masked thresholds and model predictions for synthetic plosive bursts. Masked thresholds for synthetic plosive bursts are plotted as a function of signal duration. Thresholds, denoted by the circles, are averaged across three subjects with standard deviations represented by the error bars. Model predictions are shown by the solid lines.

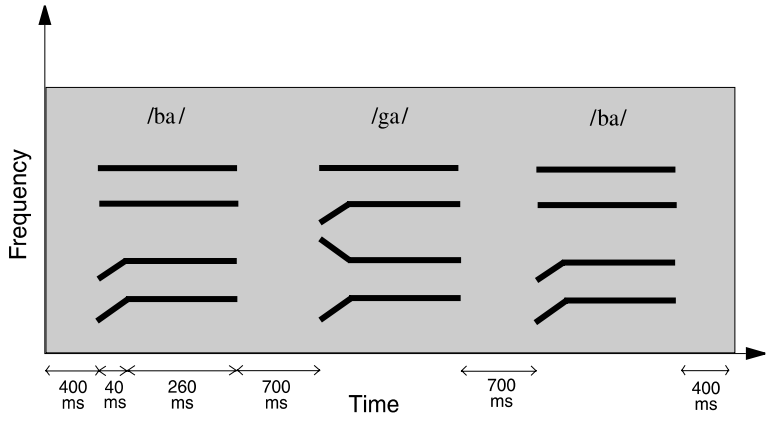

Fig. 16. Schematized spectrogram for the /ba,ga/ discrimination experiment (time axis is not drawn to scale).

typical bandwidths for F1, F2, F3 and F4, respectively (Klatt, 1980). For the /i/ context, bandwidths for F1 through F4 were 60, 150, 200 and 300 Hz, respectively. Note that the cascade synthesis resulted in time-varying amplitudes for each formant.

The initial and final frequencies of the formant transitions are shown in Tables 2–4. For the /a/

Table 2
Initial and final formant frequencies for synthetic plosive CVs in the /a/ context (in Hz)

	Fl	F2	F3	F4
/b/ onset	500	950	2400	3400
/d/ onset	500	1500	2700	3400
/g/ onset	500	1650	1850	3400
Vowel (final)	730	1100	2400	3400

Table 3
Initial and final formant frequencies for synthetic plosive CVs in the /i/ context (in Hz)

	F1	F2	F3	F4
/b/ onset	180	1600	2400	3200
/d/ onset	180	2000	2800	3900
/g/ onset	180	2500	2900	3400
Vowel (final)	330	2200	3000	3600

Table 4
Initial and final formant frequencies for synthetic plosive CVs in the /u/ context (in Hz)

	F1	F2	F3	F4
/b/ onset	180	1300	2000	3500
/d/ onset	180	1900	2700	3500
/g/ onset	180	1700	1800	4000
Vowel (final)	300	1600–1000	2250	3500

and /u/ vowel contexts, the onset and final frequencies for each formant were based on naturally spoken utterances while for the /i/ context, the values were based on those from (Blumstein and Stevens, 1980). These frequencies were then fine-tuned so that without the burst cue, each of the CVs could be easily identified.

CVs with a burst were generated by adding a 10-ms noise burst to the beginning of the 4-formant transitions. The duration of the burst was based on measurements of natural stimuli and previous studies using synthetic stimuli (Blumstein and Stevens, 1980). The gap between the offset of the burst and onset of the vowel was 5 ms. For /d/ and front and back /g/, the spectral shapes of the bursts were identical to those used in the burst-masking experiments (see Fig. 14). For /b/, the burst was generated by filtering the /b/ burst, shown in Fig. 14, with a low-pass, second-order Butterworth filter with a cutoff frequency of 1600 Hz. This was done to improve the naturalness of the /bV/ stimuli.

Table 5
Relative overall level of the synthetic plosive burst with respect to vowel onset

	/a/ (dB)	/i/ (dB)	/u/ (dB)
/b/	−20	−15	−20
/d/	−20	−15	−20
/g/	−5	−5	−5

Table 5 shows the relative levels of the bursts with respect to the vowel onset, for each plosive and vowel context. These levels were based on both naturally recorded utterances and simulation results from speech production models (Stevens, 1998).

6.2. Maskers

Masked discrimination thresholds were measured in two types of maskers, perceptually flat and speech-shaped noise. The spectrum of the perceptually flat noise was shown in Fig. 12 and the spectrum of the speech-shaped noise is shown in Fig. 17.

Each masker had a duration of 3.1 s and a level of 66.2 dB SPL, which for the perceptually flat noise, corresponds to 51 dB/ERB. The CVs were separated by 700 ms, with 400 ms of noise before the onset of the first CV and after the offset of the third CV.

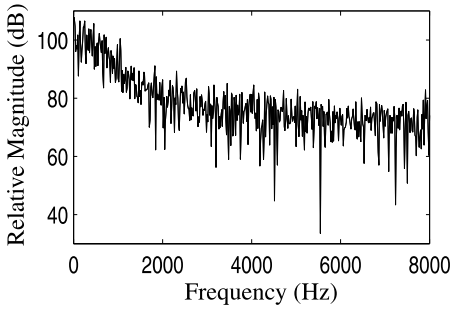

Fig. 17. Spectrum of the speech-shaped noise.

6.3. Subjects and experimental protocol

Four subjects with normal hearing participated in the experiments. Each was trained for at least 3 h, before beginning the experiments. An adaptive three AFC paradigm with no feedback was used to determine the threshold 79% correct, which corresponds to a d' of 1.62 (Green and Swets, 1966).

Masked discrimination thresholds were measured for the discrimination of /b/ and /d/, /b/ and /g/, and /d/ and /g/, for the three vowel contexts and two noise conditions. Two trials were conducted for each CV pair, in which the reference CV was switched. Final thresholds were averaged across both trials (using a dB scale).

6.4. Results

Results are shown in Fig. 18. Masked discrimination thresholds are plotted for each vowel context as a function of the plosive-consonant pair being discriminated. Results for the CVs with and without a burst are shown on the right and left sides of the dashed line, respectively. Thresholds for the perceptually flat and speech-shaped noise masker are denoted by the circles and asterisks,

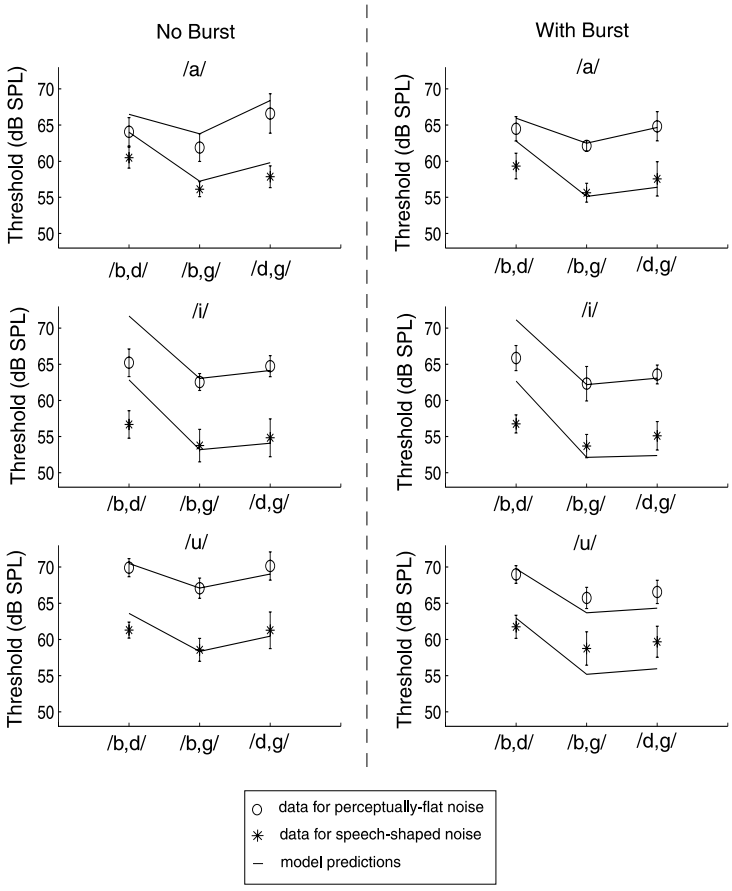

Fig. 18. Thresholds for discriminating synthetic plosive-consonant CVs in both perceptually flat and speech-shaped noise. Thresholds are plotted for each vowel context as a function of the plosive-consonant pair being discriminated. Results for CVs with and without a burst are shown on the right and left sides of the dashed line, respectively. Thresholds for the perceptually flat and speech-shaped noise masker are denoted by the circles and asterisks, respectively, while model predictions are shown by the solid lines.

respectively, while model predictions are shown by the solid lines.

Across all vowel contexts and plosive consonants, there is a 5–10 dB decrease in masked discrimination thresholds between the perceptually flat and speech-shaped noise conditions. This suggests that subjects may be taking advantage of high-frequency cues to detect plosives in noise. Perceptually flat noise masks all frequency regions equally (on an ERB scale) and will thus greatly affect the high-frequency cues. Speech-shaped noise only significantly masks the low-frequency regions of the consonants, leaving the high-frequency cues relatively uncorrupted.

Other variations in masked discrimination thresholds are largely dependent on the noise masker, vowel context, and plosive consonant. For the /a/ context and a perceptually flat noise masker, for example, thresholds for discriminating /d,g/ are largest, while for the /i/ context, thresholds for discriminating /b,d/ are the largest. In all contexts, the smallest thresholds are for the /b,g/ discrimination.

Most of these asymmetries suggest that CVs with similar formant transitions will be more easily confused in noise. In the /a/ context, for example, the formant trajectories for /d/ and /g/ only significantly differ by their F3 onset frequency (23.7 ERB for /da/ versus 20.5 ERB for /ga/). Since the amplitude of F3 (relative to F1) is small and the F2 trajectories for /d/ and /g/ are similar, the discrimination of these two consonants is more likely to be confused at the higher SNRs, especially in perceptually flat noise. In the /i/ context, /b/ and /d/ both have rising F2 and F3 transitions, whose onset frequencies differ by 1.8 and 1.3 ERB, respectively. These similar trajectories are likely to be confused at higher SNRs, resulting in elevated discrimination thresholds.

Results also show similar thresholds for the burst and no-burst conditions. The only considerable drop in thresholds occurs for discriminating /du/ and /gu/. Smaller drops occur for /da/ and /ga/, /di/ and /gi/, and /bu/ and /gu/. These drops are somewhat expected, since the /g/ burst has the highest relative level compared to the vowel onset, and is thus, more likely to be audible at the lower SNRs.

For most of the CVs tested, it appears the discrimination of synthetic plosives in noise is dominated by the formant transition cue. This is expected considering that most of the burst cues will be masked at an SNR which is higher than the CV's discrimination threshold. Recall from Fig. 15 that the masked thresholds of plosive bursts in perceptually flat noise are between -7 and 0 dB SNR. Discrimination thresholds for plosive CVs are between -3 and 4 dB (Fig. 18). Since the relative levels of the /b/, /d/ and /g/ bursts with respect to the vowel onset are -20, -15 and -5 dB, typically only the /g/ burst will be heard at the SNR where the CV is being confused.

6.5. Model predictions

Masked discrimination thresholds were determined using the detection device described in Eq. (8). However, instead of computing the detectability d' between the masker and signal plus masker, d' was computed between the two signals (plus masker) being discriminated.

Remarkably, the detection model is able to predict most of these trends. Fig. 18 only shows considerable errors for the discrimination of /bi/ and /di/. The model's overestimation of these thresholds may be due to its coarse frequency sampling, since the /bi/ and /di/ syllables only differ slightly in their F2 and F3 onset frequencies.

These errors could also be due to a discrimination mechanism not accounted for by the detection model. Recall that the model assumes that the discrimination of speech in noise is simply a comparison between two templates of time/frequency looks. To distinguish speech stimuli such as /bi/ and /di/, which have similar time/frequency profiles, subjects may be utilizing other discrimination cues not accounted for in the model, such as the fine-time structure of both stimuli.

For the discrimination of /bu/ and /gu/ and /du/ and /gu/, the model overpredicts the drop in thresholds with the addition of the burst. Recall that the model assumes an optimal combination of the burst and transition cues for discriminating plosives in noise. The data, however, suggest that for discriminating /bu/ and /gu/ and /du/ and /gu/,

perhaps subjects are unable to combine both cues optimally.

7. Summary and conclusion

In this paper, a multi-look, time/frequency masking model is proposed to predict the detection and discrimination of speech-like stimuli in noise. Time/frequency looks are generated by processing stimuli through an auditory front end which includes bandpass filtering, squaring, time windowing, logarithmic compression, and additive internal noise (with a variance of σ^2_{int}). The model uses a weighted d' detection device which calculates a Euclidean sum of the detectabilities for the time/frequency looks whose difference in means (between the signal and non-signal stimuli) is greater than a threshold, θ.

Parameters σ_{int} and θ are fit to previously measured noise-masked thresholds of bandpass noises (Hant et al., 1997). The resulting model is able to reproduce basic trends in the perceptual data, showing an increase in spectral integration at the short durations and an increase in temporal integration at the narrow bandwidths. The model, however, underpredicts the drop in thresholds between 10 and 30 ms, a result that may be due to perceptual mechanisms sensitive to signal transients.

With parameters fit to the bandpass noise data, the detection model is then used to predict the masking of glides, speech-like formant transitions, and plosive bursts. With the exception of the 100 ms, 1500 Hz glides, the model successfully predicts the perceptual data. Finally, the model is used to predict the discrimination of synthetic plosive CV syllables in both perceptually flat and speech-shaped noise. With the exception of the /bi,di/ discrimination and the burst condition in the /u/ context, the model is able to predict the discrimination data.

The success of the model in predicting such a wide range of stimuli suggests that the perception of wide-band, non-stationary signals in noise (for the bandwidths, durations, and frequency extents tested) is probably dominated by a place-rate mechanism. Whether the signal has a coherent temporal structure (i.e. glides, formant transitions, and CVs with no burst) or a more random temporal pattern (i.e. bandpass noises and plosive bursts), detection and discrimination thresholds can be predicted by a decision device that is based on the signal's energy across time/frequency looks. In future work, this assertion will be tested further by conducting other experiments with speech-like stimuli.

Acknowledgements

We would like to thank our subjects for their cooperation. We would also like to thank Dr. Brian Strope for many interesting discussions at the early stages of this work. This work was supported in part by NIH-NIDCD grant no. 1 R29 DC 02033-01A1, the Whitaker Foundation, and by the NSF.

References

Blumstein, S.E., Stevens, K.N., 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments. J. Acoust. Soc. Amer. 67, 648–662.

Collins, M.J., Cullen, J.K., 1978. Temporal integration of tone glides. J. Acoust. Soc. Amer. 63, 469–473.

Cullen, J.K., Collins, M.J., 1982. Audibility of short-duration tone-glides as a function of rate of frequency change. Hear. Res. 7, 115–125.

Durlach, N.I., Braida, L.D., Ito, Y., 1986. Towards a model for discrimination of broadband signals. J. Acoust. Soc. Amer. 80, 63–72.

Farar, C.L., Reed, C.M., Ito, Y., Durlach, N.I., Delhorne, L.A., Zurek, P.M., Braida, L.D., 1987. Spectral-shape discrimination. I. Results from normal-hearing listeners for stationary broadband noises. J. Acoust. Soc. Amer. 81, 1085–1092.

Fletcher, H., 1940. Auditory patterns. Rev. Mod. Phys. 12, 47–65.

Florentine, M., Buus, S., 1981. An excitation-pattern model for intensity discrimination. J. Acoust. Soc. Amer. 70, 1646–1654.

Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47, 103–138.

Green, D.M., 1960. Auditory detection of a noise signal. J. Acoust. Soc. Amer. 32, 121–131.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Krieger, New York.

Hant, J.J., Strope, B., Alwan, A., 1997. A psychoacoustic model for predicting the noise-masking of plosive bursts. J. Acoust. Soc. Amer. 101, 2789–2802.

Hughes, J.W., 1946. The threshold of audition for short periods of stimulation. Proc. R. Soc. B 133, 486–490.

Kiang, N.Y.-S., Watanabe, T., Thomas, E.C., Clark, L.F., 1965. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve (Res. Monogr. no. 35). MIT Press, Cambridge.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Amer. 67, 971–995.

Levitt, H., 1971. Transformed up–down methods in psychoacoustics. J. Acoust. Soc. Amer. 49, 467–477.

Madden, J.P., 1994. The role of frequency resolution and temporal resolution in the detection of frequency modulation. J. Acoust. Soc. Amer. 95, 454–462.

Madden, J.P., Fire, K.M., 1996. Detection and discrimination of gliding tones as a function of frequency transition and center frequency. J. Acoust. Soc. Amer. 100, 3754–3760.

Moore, B.C., Sek, A., 1998. Discrimination of frequency glides with superimposed random glides in level. J. Acoust. Soc. Amer. 104, 411–421.

Nabelek, I.V., 1978. Temporal summation of constant and gliding tones at masked auditory threshold. J. Acoust. Soc. Amer. 64, 751–763.

Patterson, R.D., 1976. Auditory filter shapes derived by noise stimuli. J. Acoust. Soc. Amer. 59, 640–654.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., 1992. Complex sounds and auditory images. In: Cazals, Y., Horner, K. (Eds.), Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing, June 1991. Pergamon Press, Oxford, pp. 429–446.

Plack, C.J., Moore, B.M., 1991. Decrement detection in normal and impaired ears. J. Acoust. Soc. Amer. 90, 3069–3076.

Plomp, R., 1970. Timbre as a multidimensional attribute of complex tones. In: Plomp, R., Smoorenberg, G.F. (Eds.), Frequency Analysis and Periodicity Detection in Hearing. Sijthoff, Leiden, pp. 376–396.

Plomp, R., Bouman, M.A., 1959. Relation between hearing threshold and duration for tone pulses. J. Acoust. Soc. Amer. 31, 749–758.

Raab, D.H., Goldberg, I.A., 1975. Auditory intensity discrimination with bursts of reproducible noise. J. Acoust. Soc. Amer. 57, 437–447.

Sek, A., Moore, B.C., 1995. Frequency discrimination as function of frequency, measured in several ways. J. Acoust. Soc. Amer. 97, 2479–2486.

Stevens, K.N., 1998. Acoustic Phonetics. The MIT Press, Cambridge, MA.

Strope, B., Alwan, A., 1997. A model of dynamic auditory perception and its application to robust word recognition. IEEE Trans. Speech Audio Process. 5, 451–464.

van den Brink, W., Houtgast, T., 1990. Efficient across-frequency integration in short-signal detection. J. Acoust. Soc. Amer. 87, 284–291.

van Schijndel, N.H., Houtgast, T., Festen, J.M., 1999. Intensity discrimination of Gaussian-windowed tones: Indications for the shape of the auditory frequency–time window. J. Acoust. Soc. Amer. 105, 3425–3435.

Viemeister, Wakefield, G.H., 1991. Temporal integration and multiple-looks. J. Acoust. Soc. Amer. 90, 858–865.

Zwislocki, J.J., 1969. Temporal summation of loudness: an analysis. J. Acoust. Soc. Amer. 46, 431–441.