Modeling the Masking of Formant Transitions in Noise

James J. Hant and Abeer Alwan
Dept. of Electrical Engineering, UCLA
Los Angeles, CA, USA 90095

Abstract

Formant transitions are critical for identifying the place of articulation for consonants. If these transitions are masked by background noise, perceptual confusions can occur. To better understand the masking of formant transitions, masking thresholds were measured for tone glides and single-formant trajectories of varying frequency extent (0 - 3 ERBs), duration (10, 30 and 100 ms), and center frequency (.5, 1.5, 3.5 kHz). Results show that thresholds are independent of frequency extent and only depend on the duration and center-frequency of the transition. A novel, time-frequency detection model, fit to previous noise-in-noise masking experiments (JASA 101, 2789-802 (1997)), is proposed which can predict these data.

1. Background and Motivation

Although there have been several studies on the noise-masking of stationary signals such as tones (e.g. Garner and Miller, 1947; Plomp and Bouman, 1959), few studies have measured the masking of non-stationary stimuli. Collins and Cullen (1978) measured the masked thresholds of both rising and falling tone glides with frequency extents of 200 to 700 Hz and 1200 to 1700 Hz, and durations between 10 and 120 ms. Tone thresholds were about 4 dB lower than thresholds for glides and between durations of 10 and 35 ms, rising glides were more detectable than falling glides. Nabelek (1978) measured glide thresholds over a wider range of frequency extents and durations and only found substantial differences between glide and tone thresholds at the largest frequency extents and shortest durations.

None of these studies, however, has specifically measured the masking of formant transitions. It is not clear whether the thresholds for these multiple-harmonic stimuli are similar to those of glides. Further, no model has been developed to predict the noise-masking of any type of non-stationary stimuli. With this in mind, masking experiments were conducted using glides and formant-transitions of varying center-frequency, duration, and frequency extent. A novel, time-frequency detection model was developed that can predict the masked thresholds of these non-stationary stimuli.

2. A Multi-Look Time/Frequency Detection Model

Traditional models of simultaneous masking (e.g. Fletcher, 1940, Patterson, 1976) have focused on the masking of long-duration, narrowband stimuli. In these models, the signal and noise are filtered through the "optimal" auditory filter centered around the signal's center frequency and if the filtered SNR is greater than a certain threshold, then the sound is heard. However for glides and formant transitions, the "optimal" filter is constantly changing, and thus, there may need to be a mechanism that combines information across multiple filter outputs. One such mechanism is described by the independent noise model (Florentine and Buus, 1981; Durlach *et al*., 1986). In this model, the input signal is filtered through an auditory filter bank and statistically-independent Gaussian noise is added to each frequency channel. The detection is made by optimally combining information across each of these different frequency "looks".

Traditionally, durational effects on masking have been modeled by placing a temporal integrator at the output of the auditory filter (e.g. Plomp and Bouman, 1959). However, to explain the drop in tone thresholds with duration, time constants for the temporal integrator must be of order 80 - 300 ms, which is much larger than the known temporal resolution of the auditory system. In an attempt to account for this discrepancy, Viemeister and Wakefield (1991) suggested that durational effects could be described by a multi-look mechanism across time. They propose that, instead of integrating over a long time window, the listener takes multiple "looks" at a long duration signal and combines the information optimally to detect the signal.

Since the goal of the current study is to predict the masking of formant transitions, which can be both wide-band and non-stationary, the effects of signal bandwidth and duration must be taken into account simultaneously. Toward this end, a combination of the independent noise and the multi-look models is proposed. It is assumed that the listener takes multiple "looks" in both time *and* frequency to detect the signal.

2.1 Preprocessing Stage

Time/frequency looks are generated by processing stimuli through an auditory model. The stages of the model are shown in Figure 1. A sound stimulus is first processed through a filter bank of roex-shaped filters whose bandwidths are determined from previous masking experiments (Glasberg and Moore, 1990). Adjacent filters in the filter bank are separated by one Equivalent Rectangular Bandwidth (ERB). The output of each filter is then squared and processed through a sliding temporal integrator. Each time window has a 4 ms flat section with a raised-cosine of 1 ms on each side, yielding a (1/2 power) equivalent duration of 5 ms. This duration is consistent with psychophysical measures of temporal resolution using glide stimuli (Madden, 1994). The output of each temporal window is logarithmically compressed and internal noise is added to each time/frequency look. Logarithmic compression, consistent with Weber's law of incremental loudness, is necessary to predict the spread in thresholds across duration. The internal noise, which is assumed to be Gaussian and independent, is a rough approximation of the stochastic nature of neural coding in the auditory system.

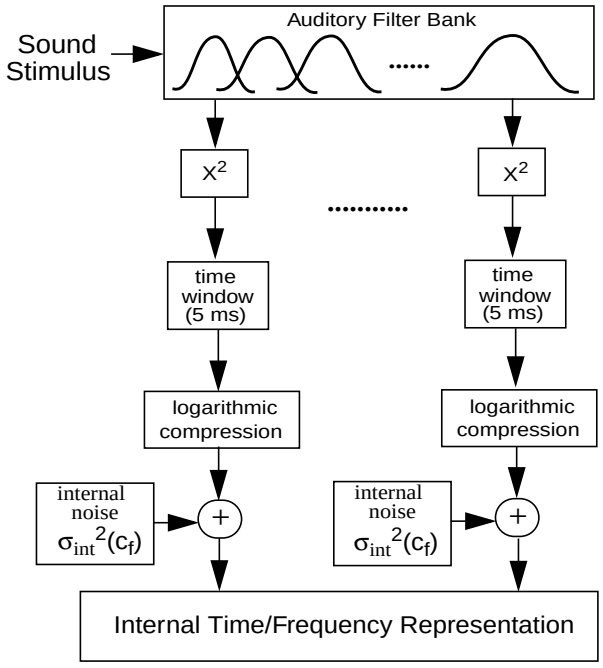

Figure 1 - Auditory Preprocessing Stage

2.2 Decision Device

To detect the signal in the presence of additive noise, the listener is assumed to perform a maximum-likelihood estimation based on the statistics of the time/frequency looks for both the noise and signal+noise stimuli. The basic structure of the decision device is shown in Figure 2.

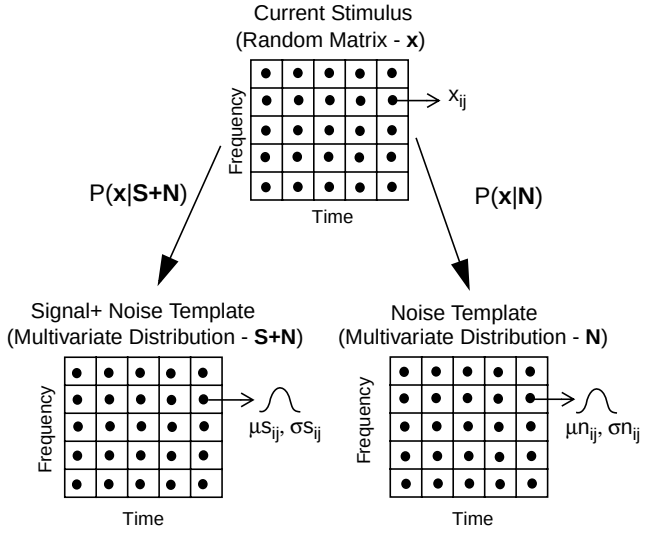

Figure 2 - Structure of the Decision Device

Due to both the stimulus and internal noise, a sound processed through the model generates a 2-dimensional random matrix (**x)** of time/frequency looks. Each element in the matrix corresponds to a time window of 5 ms and a filter bandwidth of 1 ERB, with the size of the matrix dependent on the duration and bandwidth of the stimulus. An ensemble of these matrices will generate a multivariate distribution. Through the course of the experiments, it is assumed that the listener generates and stores multivariate distributions for both the noise and signal+noise stimuli (**N** and **S+N**). The process of detection is simply comparing the random matrix generated by the current stimulus to the two stored distributions of the noise and signal+noise, and choosing the one which is more likely to have generated the stimulus.

To more easily calculate these likelihoods, it is assumed that the distributions for each time/frequency look are Gaussian and uncorrelated. In addition, a sigmoidal weighting function is applied to the likelihood calculation, giving less weight to time/frequency looks in which the means for the noise and signal+noise templates are similar. The weighting function ensures that thresholds do not decrease indefinitely as the signal bandwidth or duration increases. The likelihood calculation can be expressed as:

$$\ln(p(x|S)) - \ln(p(x|N)) = \sum_{i=1}^{Nf}\sum_{j=1}^{Nt} w(|\mu s_{ij} - \mu n_{ij}|) \bullet \left(\frac{1}{2}\ln(\sigma n_{ij}^{\,2}) - \frac{1}{2}\ln(\sigma s_{ij}^{\,2}) + \frac{(x_{ij}-\mu n_{ij})^2}{\sigma n_{ij}^{\,2}} - \frac{(x_{ij}-\mu s_{ij})^2}{\sigma s_{ij}^{\,2}} \right)$$

where: $$w(x) = \frac{1}{2} - \frac{1}{2}\frac{\left(1 - \exp\left(\frac{(x-b)}{a}\right)\right)}{\left(1 + \exp\left(\frac{(x-b)}{a}\right)\right)}$$

μn_{ij}, σn_{ij} - mean and standard deviation for the ith, jth look of the noise

μs_{ij}, σs_{ij} - mean and standard deviation for the ith, jth look of the signal+noise

Nf - total number of frequency looks

Nt - total number of time looks

a, b - sigmoidal weighting parameters

Note: σn_{ij} and σs_{ij} are a euclidean sum of the standard deviations due to stimulus and internal noise

Using this equation, the percent correct for detecting a signal (at a particular SNR) is calculated by taking an ensemble of random stimuli from the signal+noise template, and calculating the percentage of correct responses, i.e those which generate a value greater than 0. The predicted threshold is simply the SNR where the percent correct equals a particular value. For this study, we used a threshold of 72% which corresponds to 79% correct in a 2 AFC task.

2.3 Parameter Fit

There are 3 free parameters in the model which are allowed to vary with center frequency: the standard deviation of the internal noise (σ_{int}) and two sigmoidal weighting parameters, a and b. These parameters were fit to the noise-masked thresholds of bandpass noise signals which varied in bandwidth, duration, and center frequency (Hant *et al*., 1997). Parameter estimates, as a function of center-frequency, were then fit to sigmoidal-shaped curves (on an ERB scale). The resulting fits are shown in Figure 3.

Notice a sharp drop in the internal noise between ERBs of 18 and 25. This drop is needed to explain the decrease in bandpass-noise thresholds between center-frequencies of 1 and 4 kHz, a trend observed for the masked thresholds of tones in noise (Plomp and Bouman, 1959).

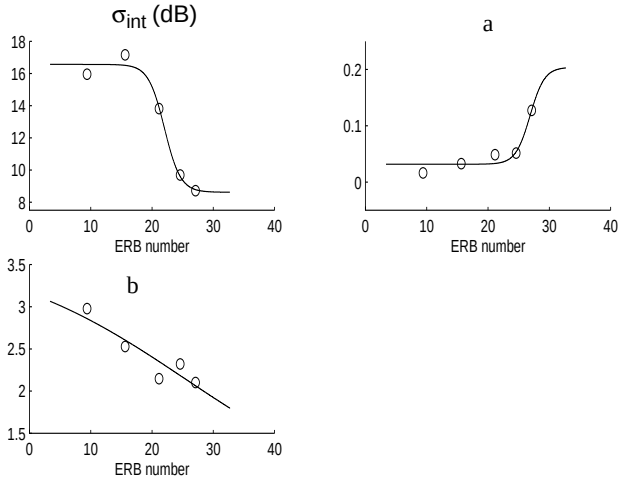

Figure 3 - Best Fit Parameters to Bandpass Noise Data in [7]

3. Data and Model Predictions

With parameters fit to the bandpass noise data, the detection model was then used to predict the masking of non-stationary stimuli similar to formant transitions in consonants. Masking experiments were conducted using glides and single formant transitions which varied in center frequency, frequency extent, and duration. A schematic of these stimuli (in a background noise masker) is shown in Figure 4.

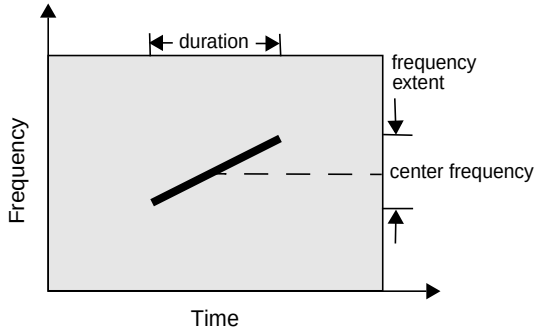

Figure 4 - Schematic of the Glide and Formant Stimuli

Three center frequencies (500, 1500, and 3500 Hz) and three durations (10, 30, and 100 ms) were tested. Frequency extents were based on a frequency scale corresponding the Equivalent Rectangular Bandwidth (ERB) of the auditory filter (Glasberg and Moore, 1990) and defined as the initial frequency minus the final frequency. At center 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.

Single formant transitions were generated in MATLAB by 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 and 200 Hz respectively. The 1500 Hz stimuli had approximate bandwidths of 1 ERB.

The masker used in the experiments was perceptually flat noise (p-flat noise), that is, noise with equal energy per ERB. This masker was at a level of 56 dB/ERB and had a duration of 750 ms. Four subjects (two males, two females) with normal hearing participated in the experiments. Thresholds were determined by an adaptive 2 AFC procedure (Levitt, 1971) which converged to the 79% correct point.

To predict thresholds, 100 samples of the masker and signal + masker stimuli were processed through the model at different SNRs and means and standard deviations were calculated for each time/frequency look. Using the best-fit parameters and 5000 random samples of the signal+masker stimuli, the percent correct was calculated using Equation 1, over a range of SNRs. The threshold was defined as the SNR which generated a response of 72% correct.

Experimental results and model predictions are shown in Figure 5. 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 4 subjects with standard deviations represented by the error bars. Model predictions are shown in solid lines.

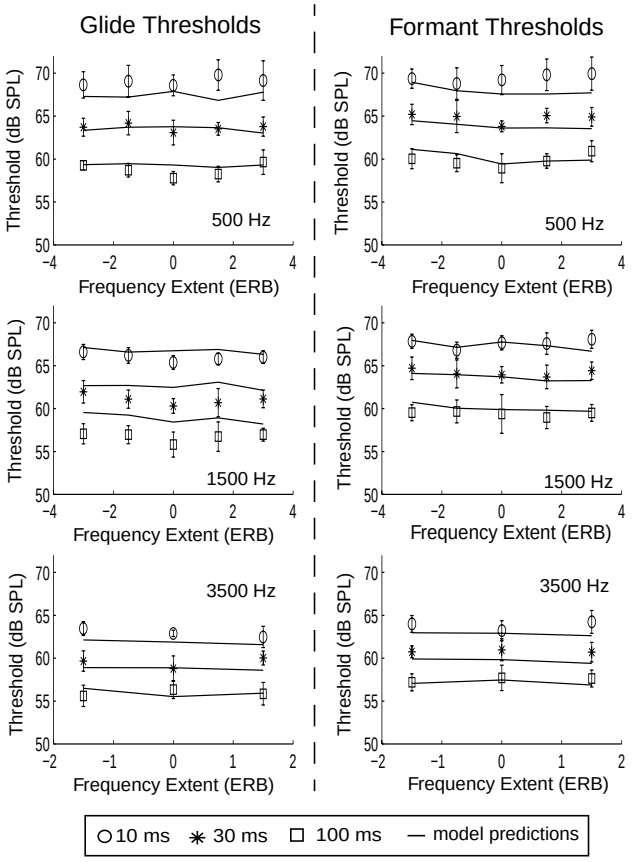

Figure 5 - Masked Thresholds for Glides and Single-Formants: Experimental Results and Model Predictions

Over the range of frequency extents and durations tested, thresholds are only dependent on the duration of the stimulus, and not its frequency extent. At center 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 center frequencies.

The current data, however, are not consistent with those of Collins and Cullen (1978) which showed glide thresholds to be 4 dB greater than the corresponding (steady) tone thresholds. The reason for this discrepancy is not clear. One main difference between the current experiments and those of Collins and Cullen is the method for estimating thresholds (Alternate Forced Choice vs. Adjustment). Perhaps subjects used perceptual cues in the alternate forced-choice experiment, which they were not able to take advantage of in the adjustment experiment.

At center frequencies of 500 and 3500 Hz, formant thresholds are only about 1 dB higher than the corresponding glide thresholds. At 1500 Hz, this difference is greater, approaching 2-3 dB at 100 ms. The small differences between glide and formant thresholds, may be attributed to their differences in bandwidth. The spread of excitation for formant transitions will be larger than for the corresponding glides, which may result in a smaller filter-SNR and slightly larger thresholds. The larger differences between glide and formant thresholds, as seen for the 1500 Hz data, may reflect a difference in how these two different types of signals are processed in the auditory system.

The model is successful in capturing the general trends in the 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 the decrease in thresholds between center-frequencies of 1500 and 3500 Hz.

However, threshold predictions for the 1500-Hz, 100-ms glides are about 1-3 dB higher than the 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 temporal cues to detect glides at 1500 Hz, which they are not using for formants.

Recent discrimination experiments using FM stimuli suggest that short-duration, non-stationary signals, such as formant transitions, may be coded by a place mechanism (Madden and Fire, 1996; Sek and Moore, 1995). The success of the 2D detection model in predicting the masking of glides and formant transitions, is further support for such a 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.

4. Summary and Conclusion

Masked thresholds were measured for glides and single-formant transitions of varying duration, frequency extent, and center-frequency. The results show that masked thresholds are relatively independent of frequency extent and only depend on the signal duration and center-frequency. A novel, multi-look, time/frequency detection model, fit to previous bandpass noise experiments, can predict a majority of these data.

5. Acknowledgements

We would like to thank our subjects for their cooperation. We would also like to thank Brian Strope for his insightful comments. This work was supported in part by NIH-NIDCD Grant No. 1 R29 DC 02033-01A1 and by the Whitaker Foundation.

6. Bibliography

1. Collins, M. J. and Cullen, J.K. (1978). "Temporal Integration of Tone Glides." J. Acoust. Soc. Am. 62, 469-473
2. Durlach, N. I., Braida, B. D., and Ito, Y. (1986). "Towards a model for discrimination of broadband signals." J. Acoust. Soc. Am. 80, 63-71.
3. Fletcher, H. (1940). Auditory Patterns. Rev. Mod. Phys. 12, 47-65.
4. Florentine, M., and Buus, S. (1981). "An excitation-pattern model for intensity discrimination. J. Acoust. Soc. Am. 70, 1646-1654.
5. Garner, W. R., and Miller, G.A. (1947). "The Masked Thresholds of Pure Tones as a Function of Duration," J. Exp. Psychol. 37, 293-303.
6. Glasberg, B. R., and Moore, B. C. (1990). "Derivation of Auditory Filter Shapes from Notched Noise Data," Hearing Research 47, 103-138.
7. Hant, J., Strope, B., and Alwan, A. (1997). "A Psychoacoustic model for the noise masking of voiceless plosive bursts," J. Acoust. Soc. Am. 101, 2789-2802.
8. Levitt, H. (1971). "Transformed Up-Down Methods in Psychoacoustic," J. Acoust. Soc. Am 49, 467-477.
9. Madden, J. P. and Fire, K.M (1996). "Detection and discrimination of frequency glides as a function of direction, duration, frequency span, and center frequency," J. Acoust. Soc. Am. 102, 2920-2924.
10. Madden, J. P. (1994) "The role of frequency and temporal resolution in the detection of frequency modulation." J. Acoust Soc. Am. 95, 454-462
11. Nabelek, I. V. (1978). "Temporal Summation of Constant and Gliding Tones at Masked Auditory Threshold." J. Acoust. Soc. Am. 64, 751-763.
12. Patterson, R. D. (1976). "Auditory Filter Shapes Derived by Noise Stimuli," J. Acoust. Soc. Am. 59, 640-654.
13. Plomp, R. and Bouman, M. A. (1959). "Relation Between Hearing Threshold and Duration for Tone Pulses," J. Acoust. Soc. Am. 31, 749-758.
14. Sek, A., and Moore, B. C. (1995). "Frequency discrimination as a function of frequency, measured several ways", J. Acoust. Soc. Am. 2479-2486.
15. Viemeister, N. F. and Wakefield, G. H. (1991). "Temporal Integration and Multiple-Looks" J. Acoust. Soc. Am. 90, 858-865.