A Psychoacoustic Model for the Noise Masking of Voiceless Plosive Bursts

James J. Hant, Brian P. Strope, Abeer Alwan
Dept. of Electrical Engineering, UCLA
Los Angeles, CA. 90095

Abstract

A model for predicting the masked thresholds of the voiceless plosive bursts /k,t,p/ in background noise is proposed. Because plosive bursts are brief, are generated by a noise source, and have different spectral characteristics, the modeling approach must account for duration, center frequency, signal bandwidth and type. To achieve this goal, noise-in-noise masking experiments are conducted using a broad band masker and bandpass noise signals of varying bandwidth (1-8 CB), duration (10-300 ms), and center frequency (0.4-4 kHz). The results of these experiments are used to parameterize an auditory filter model in which the effective bandwidths of the filters and the signal-to-noise ratio at threshold are frequency and duration-dependent. The duration-dependent filter model is then used to predict the thresholds of both synthetic and naturally-spoken plosive bursts in background noise.

1. Introduction

Plosive bursts are often confused in noisy environments but no model exists to predict these confusions. In this paper, a physchoacoustic model to predict the masking of plosive bursts in background noise is developed. This model will hopefully be a stepping stone towards a more general quantitative model of human speech perception in noisy environments.

Plosive bursts are noisy, brief, and time-varying signals. When trying to model their masking, the major characteristics of these bursts need to be characterized. Because the plosive bursts are brief and noise-like, an understanding of how duration affects the masking of noise-in-noise is important. In addition, because the bursts have different spectral shapes, the effects of signal frequency and bandwidth should also be assessed. Hence, the modeling approach of this study is to conduct perceptual experiments with bandpass noises of varying duration, bandwidth, and center frequency. These experimental results are then used to derive a physchoacoustic model for noise-in-noise masking. Finally, the bandpass noise model is used to predict the masked thresholds of both natural and synthetic plosive bursts.

2. Masking of Bandpass Noises

Four subjects with normal hearing were played bandpass noises of varying bandwidth, center frequency and duration in the middle of a background noise masker. The bandpass noises had bandwidths ranging from 1 to 8 CB, center frequencies from 400 to 4000 Hz, and durations from 10 to 300 ms. The masker used was broadband noise at a spectrum level of 36 dB/Hz and a duration of 750 ms. All signals were centered in time with respect to the masker and turned on and off using a raised cosine window with a rise/fall time of 1 ms. The masked thresholds for the noise signals were determined using an adaptive 2AFC paradigm with no feedback [Levitt 1971,1992].

The thresholds for the bandpass noises centered at 1 kHz are shown in Figure 1. In this figure, the masked thresholds (in dB SPL), are plotted as a function of signal duration with different contours in the plots representing different signal bandwidths. The experimental data are expressed in this manner to emphasize the effect that bandwidth has on the threshold at the different signal durations. Data are averaged across subjects with standard deviations represented by error bars.

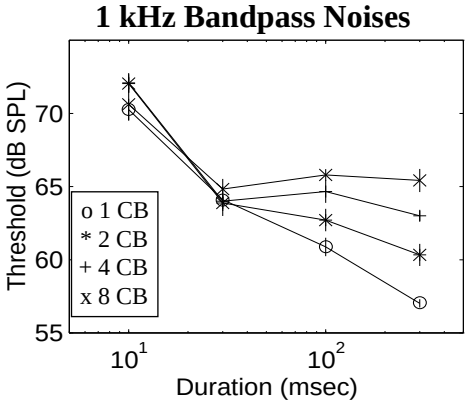


Figure 1 - Average Bandpass Noise Thresholds in a Flat Noise Masker (36 dB/Hz) for the 1 kHz data. Contours represent the different signal bandwidths.

The most striking trend seen in Figure 1 is that at short durations the total energy thresholds are similar across bandwidth, while at long durations, the thresholds increase with increasing bandwidth. This may imply that at short durations, frequency selectivity is greatly reduced. At short durations, if the subject is listening to signals over a wider frequency region, then the auditory system will sum signal energies over a large bandwidth. The result of this energy summation across frequency will be total energy thresholds that are similar across bandwidth. On the other hand, at long durations, if the subject is listening over a narrower frequency region then the auditory system will inefficiently sum energy across bandwidth. As a result, wider bandwidth signals will need to be at a higher level in order to be heard. Other investigators such as Dai and Green[1993] and van den Brink and Houtgast[1990] have described similar bandwidth effects using the results of profile analysis and tone-complex experiments, respectively.

Similar trends are observed at the other center frequencies tested. These data also show thresholds similar across signal bandwidth at the short durations, and increasing across bandwidth at the longer durations.

3. A Model of Duration-Dependent Auditory Filters

In this section, an auditory filter model in which the bandwidths of the filters increase with decreasing duration is proposed to fit the bandpass noise data.

Auditory filters are an expansion of the idea of "critical bands" first introduced by Fletcher[1940]. Fletcher measured the thresholds of static tones in a broadband noise masker and found that the tones were masked only by noise within a certain frequency band of the tone, known as the critical band. Patterson[1976] and several other investigators later used notched noise experiments to quantify the shapes of these "filters". They proposed the following roex function to describe the shape of the auditory filter:

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

p - paramter determining the slope of the filter skirts

where $g = \dfrac{|f - c_f|}{c_f}$

g - normalized magnitude of the frequency deviation from the filter's center frequency, c_f

The basic premise of the filter model is that thresholds are predicted by assuming the signal to noise ratio at the output of the filter, defined as the filtered signal energy divided by the filtered noise energy, is equal to some constant *K*, (shown in the equation below). The bandwidth of the filter is determined by the parameter *p* in the roex function above. Both parameters *K* and *p* are functions of the filter center frequency.

$$K = \frac{\int S(g)\, W(g)\, dg}{\int N(g)\, W(g)\, dg}$$

S(g) - power spectrum of the signal

N(g) - power spectrum of the noise

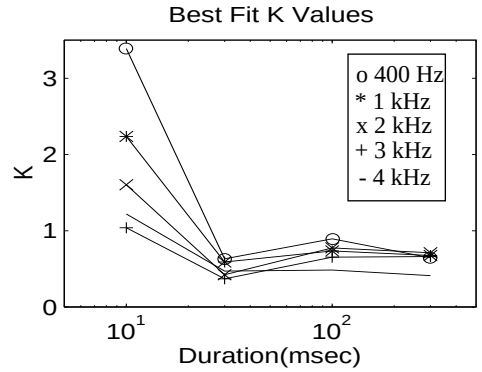
In the duration-dependent model, we assume that the parameters *p* and *K* are also functions of signal duration. The parameter *p* determines how the bandwidth of the auditory filter changes with duration, while the parameter *K* determines how the threshold SNR in each filter changes with duration.

The duration-dependent filter model was used to fit the bandpass noise data, resulting in values of *K* and *p* as functions of duration and center frequency. Some of results of this fit are shown in Figure 2. Figure 2a plots *K* values as a function of duration, where contours represent the different center frequencies. *K* values quantify all duration-dependent effects not associated with changing filter bandwidth, such as temporal integration. The top graph shows fairly flat *K* values from 300 to 30 ms, with large jumps in *K* values from 30 to 10 ms. This may imply that temporal integration occurs over a time period less than 30 ms. If the noise signals were integrated over time constants of order 100 ms, the *K* values would mostly decrease past 30 ms; not a general trend observed in our data.

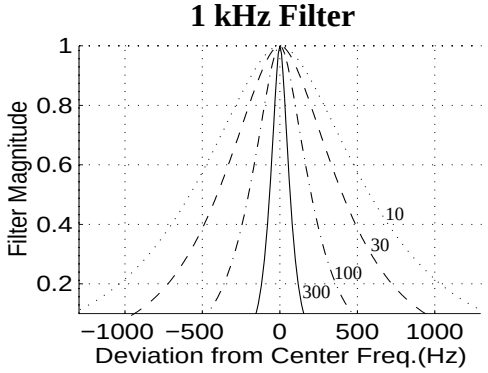
Figure 2b plots the best fit filter shapes for the 1 kHz filter at the different durations. The graph shows that filter bandwidths increase by more than a factor of 8 when the signal duration changes between 300 and 10 ms. This is consistent with the results of profile analysis experiments by Dai and Green [1993] which approximated a filter increase of 7 over the same duration range. Figure 2c plots normalized bandwidths, with respect to the steady state filters at 300 ms, as functions of duration with contours for the different center frequencies. This graph shows that between 300 and 10 ms and for center frequencies of 400 - 4000 Hz, filters widen by a factor of approximately 5 to 9 times their steady state values. In addition, there appears to be a maximum change in bandwidth for the 1kHz filter, with decreasing durational effects at the higher frequencies. The data of Figure 2 suggest, that over the center frequencies tested, filter bandwidths change significantly with duration.

.

2a



2b



2c

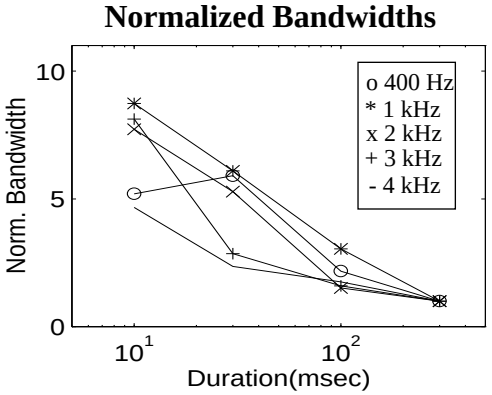


Figure 2 - Model Fit Results. 2a - Best Fit K values. 2b - Estimated filter shapes for the 1 kHz filter at the different durations. 2c - Normalized Bandwidths vs. Duration. for different center frequencies.

4. EXPERIMENTS WITH SPEECH SAMPLES

In this section, we will attempt to use the duration-dependent model to predict the masking of plosive bursts in noise. To do this, both natural and synthetic voiceless plosives are used as stimuli in masking experiments and their thresholds in noise are compared to model predictions.

4.1 Recording and Synthesis of Speech Samples

Consonant-vowel syllables were recorded from three speakers. To isolate the plosive from the syllable, the beginning of each syllable was segmented using a raised cosine window with a rise-fall time of 1 ms and a duration of either 17 or 37 ms. Because exactly 7 ms of silence was left before the start of each syllable, the windowing resulted in natural plosive samples of durations 10 and 30 ms whose bursts were not distorted by the windowing process.

Synthetic bursts were generated resembling the spectral shapes of the natural /k/, /t/, and /p/ tokens. For the /k/ and /t/, burst synthesis was done by exciting the parallel branch of the Klatt formant synthesizer (KLSYN88 - Klatt, 1980). For the /p/ burst, synthesis was done by filtering flat noise through a filter with a constant falling slope in frequency of 30 dB/8000 Hz. All bursts were synthesized at 300 ms. Using a raised cosine window with a rise/fall time of 1 ms, the bursts were then segmented into durations of either 10, 30, 100, or 300 ms. A raised cosine window was used to be consistent with the windowing used in the bandpass noise experiments.

Figure 3 shows examples of spectral shapes for both the natural and corresponding synthetic back /k/ bursts. Notice that the natural back /k/, taken from the carrier word “cot”, has a prominent spectral peak at around 1.8 kHz with a secondary peak near 4 kHz. This shape, along with the shapes of the other natural bursts, are similar to those described by Blumstein and Stevens(1979) and Massey(1994).

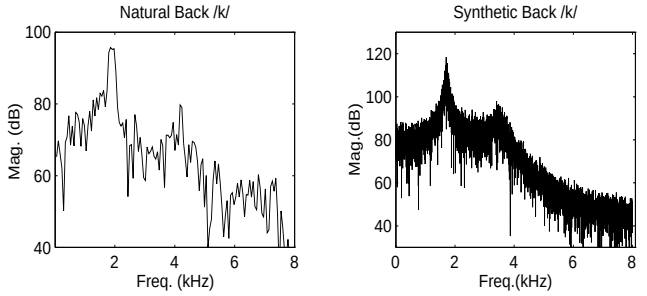


Figure 3 - Spectral Shapes for Natural and Synthetic Back /k/ Bursts. The DFT for the natural burst is taken over a 10 ms window. The DFT for synthetic /k/ is taken over a 300 ms window.

4.2 Model Predictions of Speech Thresholds

With the spectral shapes of the different bursts calculated, the duration-dependent filter model can be used to predict the thresholds of the plosives, as shown in Figure 4. At each duration, the plosive burst signal and background noise is first processed through a filter bank corresponding to the interpolated *p*-values. The filtered signal energy, normalized to the total energy of the signal, is determined by multiplying, in frequency, the filter with the spectral shape of the burst and dividing by the burst’s total energy. Normalized filter SNR values are calculated by dividing this normalized energy by the filtered noise energy. Thresholds are predicted for each filter output by dividing the corresponding *K* value with the normalized SNR. The minimum of the different filter thresholds is considered the final signal threshold.

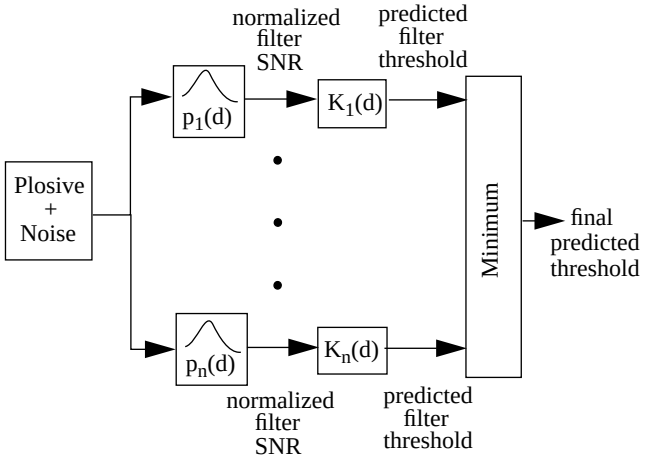


Figure 4 - Schematic for Predicting Plosive Burst Thresholds

The result of this processing is shown in Figure 5. The figure plots the output energies of the filter bank at each duration for the synthetic back /k/ burst. The output energy for each filter, denoted by circles on the graphs, is normalized to the total energy of the signal. At long durations, auditory filters are narrow and hence, the filter bank preserves the original spectral shape of the burst. At the shorter durations, however, auditory filters become wider and there is spectral smoothing. At 10 ms, this smoothing is so severe, that the /k/ burst no longer appears to have a second resonance peak and its once sharp spectral peak at 1.7 kHz is now much broader.

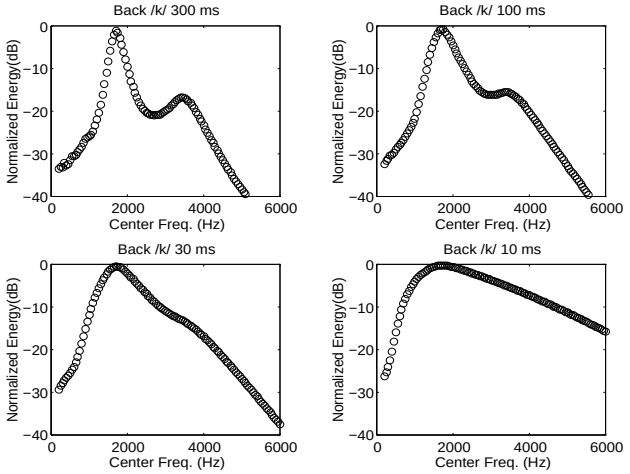


Figure 5 - Predicted spectral smoothing for synthetic back /k/ at different durations. The output energies for each filter, denoted by the circles, are normalized to the total energy of the /k/ burst.

4.3 Perceptual Masking Experiments

The natural and synthetic bursts generated were used as signals in two sets of masking experiments. These speech signals were played for listeners in the middle of a noise masker and listening thresholds were measured. For the synthetic speech experiments, the signals were synthetic /k,t,p/ bursts with durations of 10, 30, 100, and 300 ms. For the natural speech experiments, the signals were naturally recorded /k,t,p/ bursts with durations of 10 and 30 ms.

The masker used in both the synthetic and natural speech experiments was perceptually flat (p-flat) noise, that is, noise with equal energy per critical band. The masker was at a level of 56 dB/CB and had a duration of 750 ms. P-flat noise, as opposed to flat noise, was used primarily because this type of noise is a better approximation of background speech "babble", thus making our experiments more relevant to the masking of speech in naturally noisy environments. In addition, p-flat noise can be used to test the robustness of the filter model to different types of noise maskers. Total energy thresholds were determined using the same 2AFC procedure as in the bandpass noise experiments.

The results of these experiments for both the natural and synthetic back /k/ are shown in Figure 6. Thresholds for the plosives, denoted by the asterisks, are plotted as a function of duration. The thresholds, in terms of total energy (dB SPL), are averaged over both the corresponding speech tokens and the 3 listeners. The standard deviations, expressed by the error bars, take into account both speaker and listener variations. The model predictions for these bursts are expressed by the circles and lines.

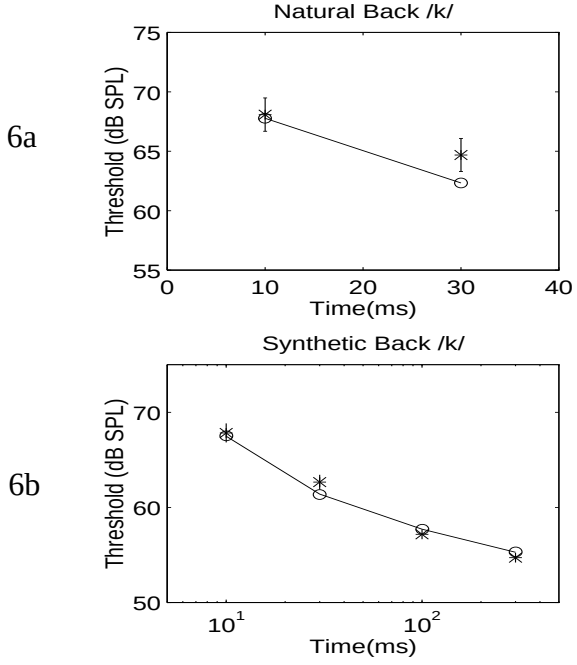


Figure 6 - Thresholds vs. Model Predictions. 6a - shows experimental thresholds for the natural /k/ burst as a function of duration. 6b - shows experimental thresholds for the synthetic /k/ bursts. Model predictions are shown by the circles.

The plots of Figure 6 show that the duration-dependent filter model is successful in predicting thresholds at 10 ms. At 10 ms, the model predictions for both the natural and synthetic /k/ burst fall within the standard deviations of the experimental data. Considering the amount of subject variation implicit in any type of perceptual experiment compounded by the variation in the natural speech samples, these errors are small enough to consider the model a success at 10 ms.

However, the model errors for the 30 ms data are larger, especially for the natural /k/ burst. Possible sources of error include the model's failure to consider cross-filter comparisons and to take into account the time-varying nature of the natural burst. For the /p/ and /t/ bursts, the model performs similarly with small errors primarily at 30 ms.

5. Conclusions

The net effect of a duration-dependent filter interpretation of the auditory system, is that at short durations the speech spectrum is greatly smoothed. Hence, it is possible that detailed spectral shapes are not necessary for identifying the plosive's place of articulation. Figure 5 shows that broad spectral cues, like the concentration of energy at high or low frequencies, are still retained at 10 ms and could play a role in identifying place. However, more subtle cues, like the bandwidth of spectral peaks, are most likely smoothed at these short durations. This spectral smoothing could also have ramifications on speech coding schemes. If our ear cannot resolve frequency at short durations, then coding short duration speech signals, like plosive bursts, with a fine frequency resolution may not be necessary.

Work supported in part by NIH - NIDCD Grant No. 1 R29 DC 02033-01A1 and the Whitaker Foundation.

6. Bibliography

1. Blumstein, S.E. and Stevens, K.N. (1980). "Perceptual Invariance and Onset Spectra for Stop Consonants in Different Vowel Environments," J Acoust. Soc. Am. 67, 648-662.
2. Dai, H. and Green, D.M. (1993). "Discrimination of Spectral Shape as a Function of Stimulus Duration," J. Acoust. Soc. Am. 93(2), 957-965.
3. Fletcher, H. (1940). Auditory Patterns. Rev. Mod. Phys. 12, 47-65.
4. Klatt, D.H. (1980). "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am. 67, 971-995.
5. Levitt, H. (1971). "Transformed Up-Down Methods in Psychoacoustics," J. Acoust. Soc. Am. 49, 467-477.
6. Massey, N. S. (1994) "Transients at Stop-Consonant Releases," M.S. Thesis, MIT (unpublished).
7. Patterson, R.D. (1976). "Auditory Filter Shapes Derived from Noise Stimuli" J. Acoust. Soc. Am. 59, 640-654.
8. Van den Brink, W.A., Houtgast, T. (1990). "Efficient Across-Frequency Integration in Short-Signal Detection," J. Acoust. Soc. Am. 87, 284-291.