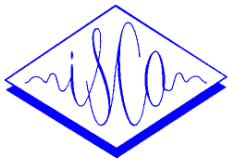

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Predicting the Perceptual Confusion of Synthetic Plosive Consonants in Noise

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

Abstract

In previous work, a novel, time/frequency detection model was developed based on psychoacoustic masking experiments and used to predict the noise masking of speech-like bursts and formant transitions [5]. In this paper, the same model is used to predict the discrimination of voiced synthetic plosive consonants in a variety of noisy environments. Discrimination experiments were conducted using synthetic /bV/, /dV/, and /gV/ syllables and two different additive noise maskers (speech-shaped and perceptually-flat). Experiments were conducted across three vowel contexts (/a/, /i/, and /u/) using CV syllables both with and without a noise burst.

Results show that discrimination thresholds are largely dependent on the noise masker, vowel context, and plosive consonant. For all experimental conditions, the addition of the burst has little effect on thresholds, suggesting that the perception of plosive consonants in noise is dominated by the formant transition cue.

The previously derived, time/frequency detection model was then used to predict the perceptual data. The model is successful in predicting most of the results, but overpredicts discrimination thresholds for /bi/ and /di/.

1. Introduction

Previous studies of perceptual consonant confusions in noise [10, 14] used a large number of /Ca/ syllables in a white-noise masker. It is difficult to conclude from these studies how the confusion of a particular type of consonant, such as a plosive, is affected by vowel context or type of noise. Previous models of consonant confusion in noise have used an information/theoretic approach which consists of analyzing confusion data statistically in an attempt to find out which acoustic/phonetic dimensions account for the greatest variance [14, 12]. Such a data-driven approach may not be easily extended to different speech stimuli and noise-conditions.

In this study, discrimination thresholds are measured for synthetic syllables (/bV/, /dV/, /gV/) in three vowel contexts (/a/, /i/, /u/) and for two different noise maskers (perceptually-flat and speech-shaped). To model the perceptual data, the speech-like stimuli are analyzed with a psychoacoustic masking model. A similar approach has proven successful in predicting both the discrimination and detection of synthetic plosive bursts in noise [2, 4] and the detection of formant transitions in noise [5]. Model predictions are compared to the experimental data.

2. Experimental Stimuli and Protocol

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 spectograms of /Ca/ stimuli (with no burst) used in one such experiment are shown in Fig. 1.

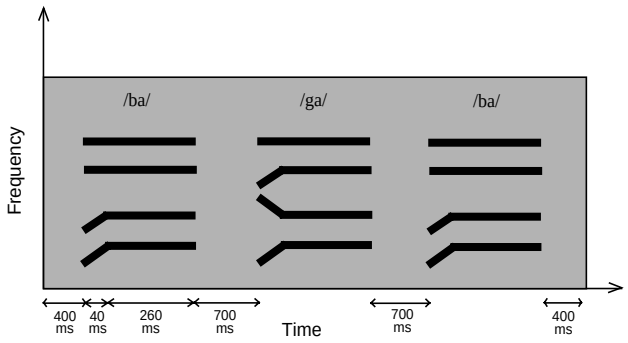

Figure 1 - Schematized spectograms of /Ca/ stimuli (with no burst) used in one discrimination experiment

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 typical bandwidths for F1, F2, F3, and F4, respectively [6]. For the /i/ context, bandwidths of 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 for each CV are shown in Table 1. For vowel contexts /a/ and /u/, 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 [1]. These frequencies were then fine-tuned so that without the burst cue, each of the CVs could be easily identified. The frequencies of the formants remained constant past the 40-ms transition region except for the /u/ vowel context, where F2 decreased from 1600 to 1100 Hz.

	F1	F2	F3	F4
/ba/ onset	500	950	2400	3400
/da/ onset	500	1500	2700	3400
/ga/ onset	500	1650	1850	3400
/a/ (final)	730	1100	2400	3400
/bi/ onset	180	1600	2400	3200
/di/ onset	180	2000	2800	3900
/gi/ onset	180	2500	2900	3400
/i/ (final)	330	2200	3000	3600
/bu/ onset	180	1300	2000	3500
/du/ onset	180	1900	2700	3500
/gu/ onset	180	1700	1800	4000
/u/ (final)	300	1600 (1100)	2250	3500

Table 1 - Initial and final frequencies (in Hz) of the formant transitions for the plosive CVs used in the experiments

CVs with a burst were generated by adding a 10-ms noise burst to the beginning of the 4-formant transitions. The durations and spectral shapes of the bursts were based on measurements of natural stimuli and previous studies using synthetic stimuli [1,4]. The gap between the offset of the burst and onset of the vowel was 5 ms.

For the consonants /b/ and /d/, the burst was at a level of -15 to -20 dB with respect to vowel onset. For /g/, the burst was at a level of -5 dB with respect to vowel onset. These levels were based on both naturally recorded utterances and simulation results from speech production models [13].

Masked thresholds were measured in two types of maskers, perceptually-flat and speech-shaped noise. The perceptually flat-noise had equal energy on an ERB-frequency scale [3] while the speech-shaped noise had a spectrum similar to the long-term average speech spectrum. Each masker had a duration of 3.1 seconds and a level of 66.2 dB SPL, which for the perceptually-flat noise, corresponded 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.

Four subjects with normal hearing participated in the experiments. Each subject was trained for at least 3 hours, before beginning the experiments. An adaptive 3 AFC paradigm with no feedback [7] was used to determine the threshold 79% correct.

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. So for the /ba,da/ distinction, for example, one threshold was measured using /ba/ as a reference (i.e. two /ba/'s and one /da/) and a second threshold was measured using /da/ as a reference (i.e. two /da/'s and one /ba/). Final thresholds were averaged across both trials (using a dB scale).

3. Model Predictions

A schematic of the method for predicting discrimination thresholds (between syllables CV1 and CV2) is shown in Figure 2. Internal templates were first generated (over a range of SNRs) for both CV1 + Masker and CV2 + Masker by processing 100 examples of each stimulus through an auditory front-end which included bandpass filtering, squaring, time windowing, logarithmic compression and additive internal noise. The auditory filter-bank was based on an ERB frequency scale [3], the time window of 5 ms was based on previous temporal resolution experiments [9], and the level of the internal noise was set to fit previous noise-in-noise masking data [5]. The result of the front-end processing was a multivariate distribution of time/frequency "looks" for both CV1 and CV2. Assuming these distributions are Gaussian and statistically independent (as simulation results support), the templates for both stimuli can be represented as matrices of means, $\mu 1_{ij}(SNR)$ and $\mu 2_{ij}(SNR)$, and standard deviations, $\sigma 1_{ij}(SNR)$ and $\sigma 2_{ij}(SNR)$.

Under this framework, the percent correct discrimination was determined over a range of SNRs using a maximum-likelihood criterion[5]. Note that the SNR was defined as the total signal power divided by the total noise power. Thresholds were determined by the SNR at which percent correct equaled 79%, corresponding to the value determined by the adaptive 3 AFC procedure.

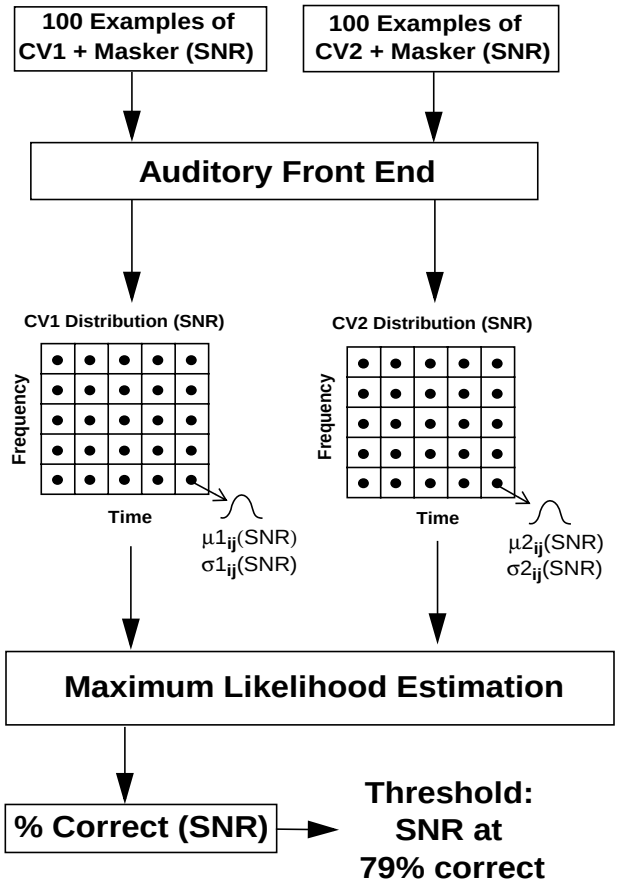

Figure 2 - Schematic of the Method for Predicting Discrimination Thresholds.

4. Results

The results of the discrimination experiment and model predictions are shown in Fig. 3. Masked thresholds for discriminating plosive CVs 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 maskers are denoted by the circles and asterisks, respectively, while model predictions are shown by the solid lines.

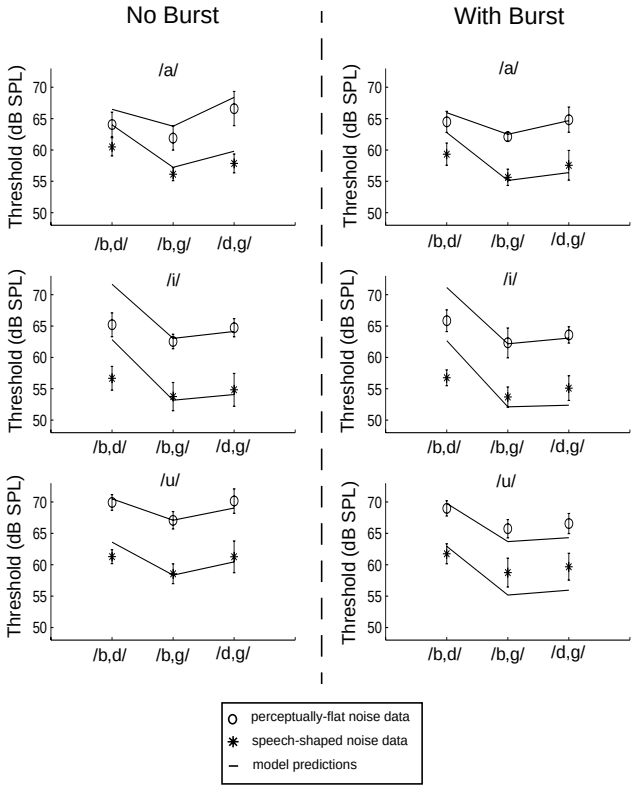

Figure 3 - Thresholds for Discriminating Voiced, Synthetic Plosive Consonants in Perceptually-Flat and Speech-Shaped Noise.

Across all vowel contexts and plosive consonants, there is a 5-10 dB decrease in thresholds between the perceptually-flat and speech-shaped noise conditions. This suggests that subjects may be taking advantage of high frequency cues to discriminate 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 speech sounds, leaving the high-frequency cues relatively uncorrupted.

Other variations in 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 /b,d/ are the largest. In the /u/ context, the smallest thresholds are for the /b,g/ discrimination.

Most of these asymmetries are consistent with the fact that formant transitions which have similar frequency trajectories (and relatively low amplitudes) will be more easily confused in noise. In the /a/ context, the formant trajectories for /d/ and /g/ only significantly differ by their F3 onset frequency. 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. In the /i/ context, /b/ and /d/ both have rising F2 and F3 transitions, which have similar onset frequencies. These similar trajectories are more likely to be masked at higher SNRs, resulting in elevated discrimination thresholds.

The lower thresholds for discriminating /bu/ and /gu/, however, cannot be easily explained by the experimental stimuli. In fact, the time/frequency profiles of /bu/ and /gu/ are more similar than for /bu/ and /du/. Perhaps, subjects were able to take advantage of /gu/'s velar "pinch" (between F2 and F3) to distinguish it from /bu/. This pinch creates a strong energy peak at 1700 Hz which may be highly perceptible in noise.

Remarkably, the detection model is able to predict most of these trends. Fig. 3 only shows considerable errors for the discrimination of /bi/ and /di/. The model's over-estimation of these thresholds may be due to its coarse frequency sampling, since the syllables /bi/ and /di/ both have similar onset frequencies for F2 and F3.

These errors could also be due to a discrimination mechanism not accounted for by the detection model. Recall, the model assumes that the discrimination of speech in noise is simply a maximum-likelihood 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 mechanisms not accounted for in the model.

To reveal the effect of the burst cue, Figure 4 plots the same discrimination data for the burst and no burst conditions together. Thresholds for the perceptually-flat and speech-shaped noise maskers are displayed on the left and right sides of the figure, respectively. Data for the burst and no burst conditions are shown by the circles and asterisks, respectively, while model predictions are shown by the solid and dotted lines, respectively.

Results show that the addition of the burst has very little effect on discrimination. The only considerable drop in thresholds when a burst is added 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 the /bu, gu/ and /du, gu/ discrimination, 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. Nevertheless, it appears that for most of the CVs tested, 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. Previous results show that the masked thresholds of plosive bursts in perceptually-flat noise are between

-7 and 0 dB SNR [4]. Discrimination thresholds for plosive CVs are between -3 and 4 dB. Since the relative levels of the /b/, /d/, and /g/ bursts with respect to the vowel onset are -20, -15, and -5 dB, only the /g/ burst will be heard at the SNR where the CV is being confused.

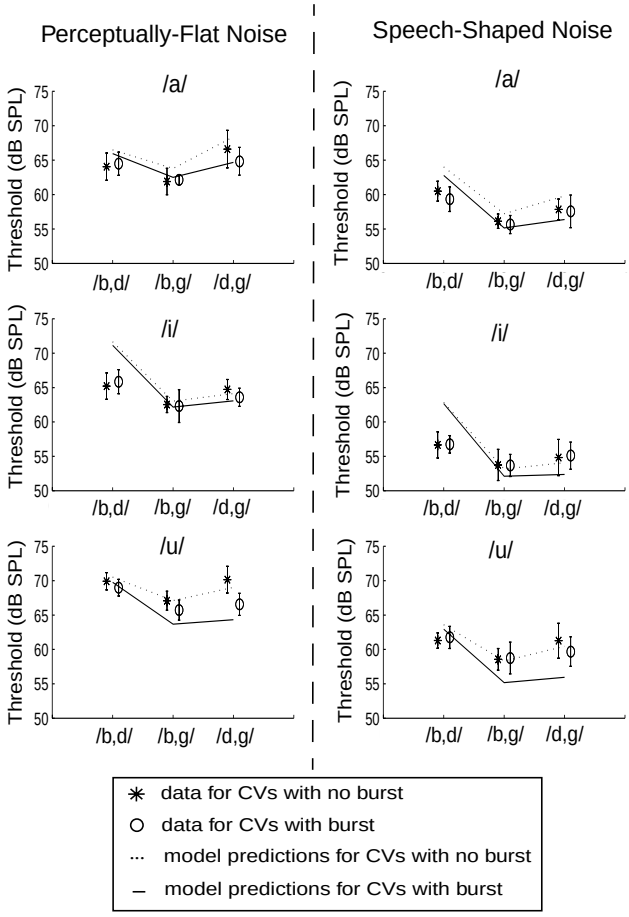

Figure 4 - Comparison of Discrimination Thresholds for Plosive CVs With and Without a Burst.

5. Conclusion

In this paper, a time/frequency detection model, based on psychoacoustic-masking experiments, is used to predict the discrimination of synthetic plosive consonants in two different noise maskers. For all consonants and vowel contexts, there is a 5-10 dB drop in thresholds between the perceptually-flat and speech-shaped noise maskers, suggesting that subjects may be using high frequency cues to discriminate plosives in noise. Other variations in thresholds are largely dependent on the consonant and vowel context.

Despite the complexity of these stimuli, discrimination thresholds can be reasonably predicted by a model that is based on the signals' energy across time/frequency looks. There is increasing evidence in the psychoacoustic literature that short-duration, non-stationary signals such as formant transitions and noise-bursts may be coded by a place-rate mechanism [5, 8, 11]. The success of the time/frequency detection model in predicting the discrimination of synthetic plosive CVs in noise is further support for the place-coding of speech.

Regardless of how the data are modeled, the results of this study suggest that the confusion of speech in noise is largely a function of the consonant, vowel context, and type of noise masker. Any future measurements and models of speech confusion in noise should take these variables into account.

5. Acknowledgements

We would like to thank our subjects for their cooperation. This work was supported in part by NIH-NIDCD Grant No. 1 R29 DC 02033-01A1 and by 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. 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. Am. 81, 1085-1092.
3. Glasberg, B. R., and Moore, B. C. (1990). "Derivation of Auditory Filter Shapes from Notched-Noise Data," Hearing Res. 47, 103-138.
4. Hant, J., Strope, B., and Alwan, A. (1997). "A psychoacoustic model for predicting the noise-masking of plosive bursts," J. Acoust. Soc. Am., 101, 2789-2802.
5. Hant, J. and Alwan, A. (1999) "Modeling the masking of formant transitions in noise", presented at Eurospeech99 - Budapest, Hungary (4) 1895-1898.
6. Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 971-995.
7. Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. 49, 467-477.
8. Madden, J. P. and Fire, K. M. (1996). "Detection and discrimination of gliding tones as a function of frequency transition and center frequency," J. Acoust. Soc. Am. 100, 3754-3760.
9. Madden, J. P. (1994). "The role of frequency resolution and temporal resolution in the detection of frequency modulation," J. Acoust. Soc. Am. 95, 454-462.
10. Miller, G.A. and Nicely, P.E. (1955). "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am. 27, 338-352.
11. Sek, A., and Moore, B. C. (1995). "Frequency discrimination as a function of frequency, measured several ways", J. Acoust. Soc. Am. 2479-2486.
12. Soli, S.D. and Arabie, P. (1979). "Auditory versus phonetic accounts of observed confusions between consonant phonemes," J. Acoust. Soc. Am 66, 46-59.
13. Stevens, K.N. (1998). Acoustic Phonetics. The MIT Press, Cambridge, Massachusetts.
14. Wang, M.D. and Bilger, R.C. (1973) "Consonant confusion in noise: a study of perceptual features," J. Acoust. Soc. Am. 54, 1248-1265.