MODELLING SPEECH PERCEPTION IN NOISE: A CASE STUDY OF THE PLACE OF ARTICULATION FEATURE

Abeer Alwan

Research Laboratory of Electronics and the Department of Electrical Engineering and Computer Science, MIT, Cambridge MA USA

ABSTRACT

In this study, perceptual confusions of the place of articulation feature for syllable-initial /b,d/ stop consonants in noise are examined. Experimental data are compared with a model, based on auditory masking theory, that estimates the level and spectrum of noise needed to mask each formant peak. Results show good agreement between the experiments and the theoretical model, and indicates that F2 transition is essential in signalling the place distinction for these consonants.

1. INTRODUCTION

The goal of the study is to develop procedures for predicting the perceptual confusions of speech sounds in noise. The prediction is based on the following premise: if the acoustic attributes that signal a particular phonetic contrast are known, then, based on auditory masking theory, it should be possible to calculate the level and spectrum of noise that will mask these acoustic attributes, and hence will lead to confusions in listener responses to that phonetic contrast. The methodology here is threefold: 1) quantifying acoustic correlates of phonetic features in naturally-spoken utterances and using the results to synthesize these utterances, 2) using masking theory to predict the level and spectrum of the noise which will mask these acoustic correlates, and 3) performing a series of perceptual experiments, using synthetic stimuli, to evaluate the theoretical predictions.

The feature chosen for the present phase of the study is place of articulation for the stop consonants /b,d/ in CV syllables with the vowels /a/ and /ɛ/.

2. ANALYSIS AND SYNTHESIS

The place distinction for the consonants /b,d/ is signalled mainly by the shape of the trajectory of the second formant frequency (F2) and by the spectral shape of the burst. The F2 transition is thought to carry most of the place information for syllable-initial stop consonants [2]. In the /Ca/ case, the F2 trajectory falls into the vowel for the alveolar /d/ and rises for the labial /b/. With the vowel /ɛ/, the F2 trajectory rises for /b/ and is almost flat for /d/. Figures 1 and 2 show schematized spectrograms of synthetic burstless utterances of /ba, da/ and of /bɛ, dɛ/, respectively, illustrating the differences in the F2 trajectories. These synthetic utterances, which are used later in perceptual experiments, are generated using KLSYN88 [4]. The choice of parameters is based on analyses of natural utterances spoken by a male speaker. The synthetic utterances were 100% identifiable and discriminable in quiet when played back to a group of three subjects.

3. THEORETICAL PREDICTIONS

If it is assumed that masking of formant frequencies is similar to masking of tones at the same frequency (an assumption verified through pilot experiments), the level of noise needed to just mask out each formant frequency can be calculated. For example, the *i*th formant frequency is masked if the level of a white-noise masker in a critical-band around that formant frequency (Nc_i) is 4 dB greater than the amplitude (in dB) of the formant frequency (A_i) [5]. Nc_i is the rms level of the noise (in dB) estimated from the DFT spectrum and corrected by 10 log

(ratio of the analysis-filter bandwidth to the critical bandwidth the *i*th formant frequency). That is, the condition is that $Nc_i \geq A_i + 4$ .Calculations can then be made to determine the time interval over which each formant frequency is masked. Figure 3 illustrates these computations for a white-noise masker at a particular level for which F2 transition in the synthetic /da/ stimulus is partially masked. In this case, F1 is never masked ($A_1 + 4 > Nc_1$), F3 is always masked ($A_3 + 4 < Nc_3$) and only the first 10 ms of F2 is masked. Note that the spectral peak of F2 changes by about 10 dB during the transition period. This is in accordance with observations of amplitude changes in natural speech. The computations are done every pitch period.

4. EXPERIMENTS

The goal of these experiments is to examine the perceptual importance of the F2 trajectory in signalling the place-of-articulation feature distinction for the plosives /b,d/ in syllable-initial position with the vowels /a/ and /ɛ/. Nonsense syllables were used to make sure that lexical effects, such as word frequency, do not bias subjects' responses.

4.1 Stimuli and Experimental Design

Synthetic utterances of /ba, da, bɛ, dɛ/ were attenuated, mixed with white noise, randomized, repeated 10 times and presented to subjects in identification tests. There were 13 stimuli with different signal-noise ratios (SNR) for each utterance. The SNR was varied by changing the signal level in 1 dB steps while keeping the noise level constant. The presentation level, as determined by the peak in the vowel, was 66 to 79 dB SPL.

4.2 Subjects

Four subjects participated in the /Ca/ experiments and three subjects participated in the /Cɛ/ experiments. Two of the subjects were students at MIT. None had any known speech or hearing problems. Training periods, lasting between 1/2 h to 1 h depending on the subject, preceded each listening session.

4.3 Results

4.3.1 /Ca/ case

The results of these experiments show that the /ba/ stimuli were perceived correctly at all noise levels used in the experiment. Figure 4 shows the results of the experiments for each subject individually for the /da/ stimuli. The responses are plotted as a function of the SNR in a critical band of F2 in the steady-state portion of the vowel. These identification functions show an abrupt shift from /da/ to /ba/. The average threshold for the subjects occurs at the stimulus where 23 ms of the F2 transition is masked. The total duration of the F2 transition is 40 ms.

4.3.2 /Cɛ/ case

The results of these experiments show that the /dɛ/ stimuli were identified correctly at all noise levels used in the experiment even when F2 is completely masked. Figure 5 shows the results of the experiments for each subject individually for the /bɛ/ stimuli. The responses are plotted as a function of the SNR in a critical band of F2 in the steady-state portion of the vowel. These identification functions show a shift in perception from /bɛ/ to /dɛ/. However, the identification functions show individual differences in listeners' responses. It is interesting to note that these differences are similar to those found in the listeners' masked thresholds of pure tones in independent tests.

5. Discussion

The results of this study show that the shape of the F2 trajectory is essential in identifying the place of articulation for the consonants /b/ and /d/ preceding the vowels /a/ and /ɛ/. The labial feature is signalled by a flat trajectory when preceding /a/ and a rising trajectory preceding /ɛ/. If noise masks most of the F2 transition such that only the steady-state part of the transition is free of masking, then /dɛ/ is perceived. The feature alveolar, on the other hand, is signalled by a flat trajectory preceding /ɛ/ and a falling trajectory preceding /a/. If noise masks out most of the F2 transition for /da/ such that the movement of F2 is minimal, then the stimulus is perceived as /ba/. This result is in agreement with results of other researchers [1][3] who

observed that the first 20 ms or so of the F2 transition carries important place information for /d/. Their observations were based on perceptual experiments conducted in quiet.

Other experiments examining the perceptual role of stop bursts are underway. Preliminary results indicate that in the /Ca/ case and in the presence of white noise, the burst is masked at very low SNR and, hence, does not play a significant perceptual role. We plan to pursue this approach further in investigating other phonetic contrasts in noise such as manner of articulation and voicing and to test the model under 'shaped' noise conditions.

6. References

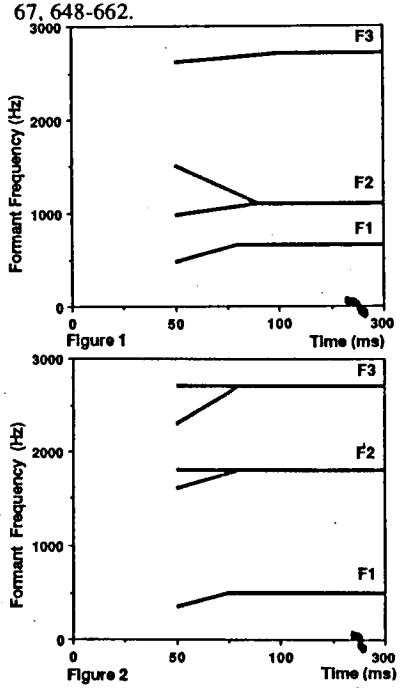
[1] Blumstein, S., and Stevens, K.N. (1980). "Perceptual invariance and onset spectra for stop consonants in different environments," J. Acoust. Soc. Am., 67, 648-662.

[2] Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am., 27, 769-773.

[3] Kewley-Port, D. (1983)."Time-varying features as correlates of place of articulation in stop consonants" J. Acoust. Soc. Am., 73, 322-335.

[4] Klatt, D. H. and Klatt L.C. (1990)."Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., 87, 820-857.

[5] Moore, B. (1982). *An Introduction to the Psychology of Hearing*. Academic Press. London.



Schematized trajectories for the first three formant frequencies for synthetic /ba/ (solid line) and /da/ (dashed line) utterances in Fig.1 and of /dɛ/ (solid line) and /bɛ/ (dashed line) utterances in Fig.2.

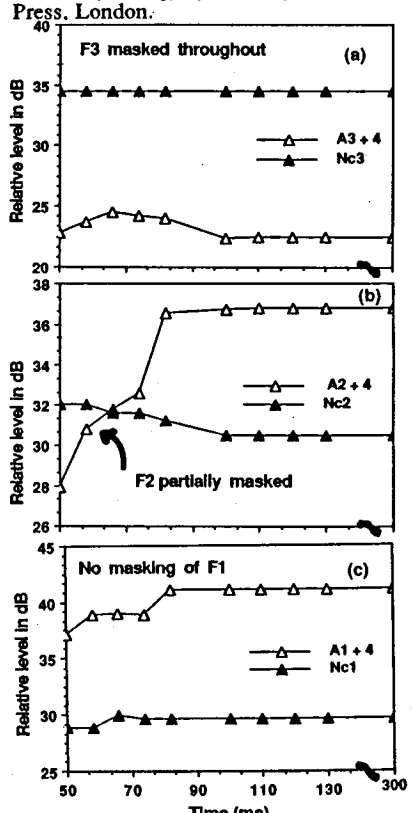


Figure 3. Plots of the relative levels of the amplitudes of the first three formant frequencies plus 4 dB (re 0.0002 bar) along with the noise levels (Nc_i) in the corresponding critical bands. Masking occurs when Nc_i is at least as high as A_i + 4. The data are for a /da/ stimulus.

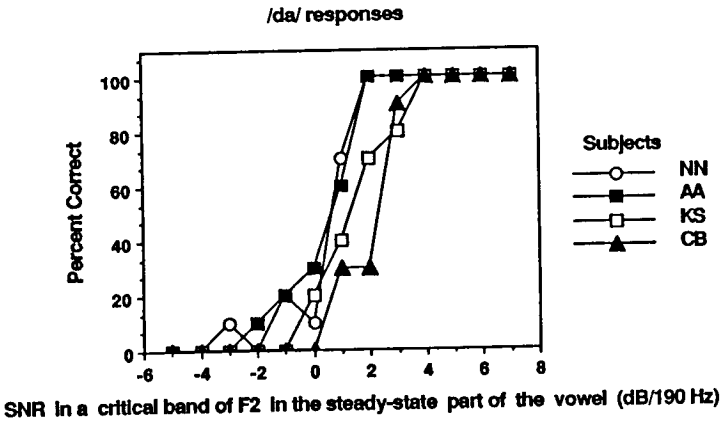


Figure 4. Plots of percent correct versus the signal-to-noise ratio in a criticial band of F2 in the steady-state part of the vowel. The critical band in this case is 190 Hz.The plots are for the /da/ stimuli.

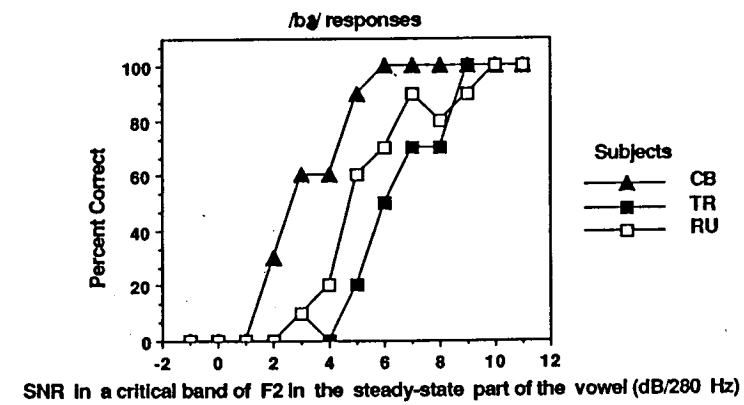


Figure 5. Plots of percent correct versus the signal-to-noise ratio in a criticial band of F2 in the steady-state part of the vowel.The critical band in this case is 280 Hz. The plots are for the /bɛ/ stimuli.