Modeling Speech Perception in Noise: The Stop Consonants as a Case Study

RLE Technical Report No. 569

Abeer A.H. Alwan

February 1992

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307

Modeling Speech Perception in Noise: The Stop Consonants As a Case Study

RLE Technical Report No. 569

Abeer A.H. Alwan

February 1992

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307

This work was supported in part by a C.J. LeBel Fellowship, in part by the National Institutes of Health under Grant 8 RO1-DC00075, and in part by a grant from the National Science Foundation.

MODELING SPEECH PERCEPTION IN NOISE: THE STOP CONSONANTS AS A CASE STUDY

by Abeer Abdul-Hussain Alwan

Submitted to the Department of Electrical Engineering and Computer Science on January 3, 1992 in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

ABSTRACT

This study develops procedures for predicting perceptual confusions of speech sounds in noise by integrating knowledge of the acoustic properties which signal phonetic contrasts of speech sounds with principles of auditory masking theory. The methodology that was followed had three components: 1) quantifying acoustic correlates of some phonological features in naturally-spoken utterances and using the results to generate synthetic utterances, 2) developing a perceptual metric to predict the level and spectrum of the noise which will mask these acoustic correlates, and 3) performing a series of perceptual experiments to evaluate the theoretical predictions.

The focus of the study was the perceptual role of the formant trajectories in signalling the place of articulation for the stop consonants /b,d/ in consonant-vowel syllables, where the vowel was either /ɑ/ or /ɛ/. Nonsense syllables were chosen for the perceptual study so that lexical effects such as word frequency did not bias subjects' responses. Computer-generated, rather than naturally-spoken, syllables were used to provide better control of the stimuli.

In the analysis/synthesis stage, the acoustic properties of the stop consonants /b,d/ imbedded in naturally-spoken CV syllables were quantified and the results were then used to synthesize these utterances with the formant synthesizer KLSYN88 (Klatt, and Klatt, 1990). In the context of the vowel /ɑ/, the two synthetic syllables differed in the F2 trajectory: the F2 trajectory was falling for /dɑ/ and was relatively flat for /bɑ/. In the C/ɛ/ context, both F2 and F3 trajectories were different for the consonants: F2 and F3 were flat for /dɛ/, whereas they were rising for /bɛ/.

A metric was then developed to predict the level of noise needed to mask a spectral peak corresponding to a formant peak. The metric was based on a combination of theoretical and empirical results. Two types of masking were studied: within-band masking (where the formant was within the bandwidth of the noise masker) and above-band masking (where the formant was above the upper cutoff frequency of the masker). Results of auditory masking theory, which was established primarily for pure tones, were used successfully to predict within-band masking of formant peaks. The predictive measure in this case was the signal-to-noise ratio in a critical band around the formant frequency. The applicability of the results of masking theory to formant peaks was tested by conducting a series of discrimination and detection experiments with synthetic, steady-state vowels.

In the above-band masking case, it was found that predictions based on the two methods known for predicting aspects of this kind of masking (ANSI standards (1969) and Ludvigsen's equation (1985)) did not agree with experimental results. An empirical algorithm was developed to account for the experimental data.

In the final stage of the study, a series of identification tests with synthetic CV utterances in noise was conducted. Two noise maskers were used in the experiments: white noise, and band-pass noise centered around the F2 region. The spectral prominences associated with F2 and F3 have a lower amplitude during the transitions from the consonant than in the steady-state vowel, so that it is possible, using a steady-state noise, to mask portions of a formant transition without masking the formant peak in the vowel. Subjects' responses were analyzed with the perceptual metric developed earlier. Results showed that when the F2 transition for C/ɑ/ or the F2 and F3 transitions for C/ɛ/ were masked by noise, listeners interpreted the stimuli as though the formant transitions were flat. That is, /dɑ/ was heard as /bɑ/, and /bɛ/ was heard as /dɛ/.

It was also found that when only the F2 trajectory is masked, achieved by selectively masking F2 with a band-pass noise masker, then amplitude differences in the F3 and F4 regions could be used as cues for place information in the C/ɑ/ case even though the trajectories of these higher formants did not differ for the two consonants.

Thesis supervisor: Professor Kenneth Noble Stevens

Title: Clarence J. LeBel Professor of Electrical Engineering

Acknowledgements

My deepest gratitude goes to my advisor Kenneth Stevens. His insights, guidance, and genuine interest in research have been motivating and inspiring for me throughout my studies at MIT. His faith and encouragement to pursue new lines of research has led to this thesis and made this work a truly rewarding experience.

Thanks to the committee members: Louis Braida, Bertrand Delgutte, Charlotte Reed, and Patrick Zurek for their careful readings of the manuscript and valuable comments. In particular, I thank Louis for his critical comments, Bertrand for always bringing up interesting questions and for suggestions which helped clarify several sections in the thesis, Charlotte for many insightful comments, and Pat for stimulating discussions on psychophysics.

Thanks to Christine Rankovic for several discussions on 'upward spread of masking', for her help in headphone calibrations, for being a subject, and for her friendship.

Thanks to JK, JM, MM, and NN for being subjects and to Sharon Manuel, Torstein Pedersen, and Rosalie Uchanski for participating in pilot experiments.

Thanks to my colleagues in the EE department at UCLA for their support and patience.

I am deeply indebt to past and present members of the Speech Communication Group for creating a friendly and enriching environment. Special thanks go to Corine Bickley and Stefanie Shattuck-Hufnagel for participating in the experiments and proofreading parts of this thesis. Thanks to Stef also for 'endless' discussions on many technical and not so technical issues. Thanks to Jane Webster for being a subject and for sharing with us her fantastic cakes. Many thanks to Seth Hall and Melanie Matthies for their help in solving computer-related problems, and to Keith North for his assistance in experimental setups whenever needed. Thanks to Arlene Wint, Carol Espy-Wilson, Joe Perkell, Caroline Huang, and Eva Holmberg for their help and support all along.

Thanks to all my friends here and abroad. In particular, thanks to Julien Piot and Isabelle Delbos for friendship, wonderful memories and chocolate, and to Munther Dahleh, Jinane Abounadi, Eman Hashem and Rami Safadi for very enjoyable times together; Munther's realistic and 'deconstructionist' views of the world helped me deal with many difficult issues. I am very grateful to Intisar Hamid for making me feel like a member of her family since I first met her 10 years ago. Thanks to Nadia Alhasani and Raya Ani for their friendship and their help in putting this thesis together.

If I were to believe in angels, then Gabriella DiBenedetto would certainly be one!. Gabriella and her family provided endless support and love. Grazie mille, amica.

My thanks also go to my extended family of aunts, uncles, cousins, and my grandfather whose affection never failed, despite the long years and distance away, and for 'hanging in there'.

Finally, I would like to thank my family: my parents, Nagham, Ali, Maythem and Hassouni for their love, encouragement, and courage during very difficult times.

This work has been supported in part by a LeBel fellowship, an NIH grant, and an NSF grant.

To my father and mother

Contents

List of Figures

Chapter 1

Introduction and Literature Review

In everyday life we often listen to degraded speech. The degradation could be due to other speech-like signals or to non-speech-like signals. The purpose of this study is to contribute to a broad research program whose aim is to understand and model the perception of speech in noise.

Speech sounds in all languages are thought to be realizations of a small number of constituents or *features.* Theories about these discrete, rather than continuous, representations of speech are based on articulatory, acoustic, and perceptual considerations (Jackobson et al., 1963) or are based on production mechanisms with less emphasis on the perceptual and acoustical dimensions (Chomsky and Halle, 1968).

The mapping of these features to the acoustic domain is not necessarily one-to-one, but rather, one-to-many. For example, place of articulation for syllable-initial stop consonants is signalled by one or a combination of acoustic cues such as the spectral shape at consonantal release, second formant frequency transition following the release, rate of formant transitions, etc. The assessment of the perceptual importance of each cue is verified through perceptual experiments where typically one acoustic property is manipulated but all other properties are kept constant. A classic example is the study by Delattre et al. (1955) where it was found that the transition of the second formant frequency carries much of the place information for syllable-initial stop consonants.

In this study, we will examine the perceptual importance of acoustic correlates of certain features under conditions where the speech signals are corrupted by noise. Although noise is very frequently the limiting factor in normal communication, most previous perceptual studies have been based on experiments conducted in quiet. The goal here is to develop procedures for predicting perceptual confusions of speech sounds in noise by integrating knowledge of the acoustic properties of the speech signal with that of the properties of the human auditory system. As a case study, the perception of the stop consonants /b,d/ in syllable-initial position with the vowels /ɑ/ and /ɛ/ in noise is considered. Specifically, the perceptual role of the formant frequencies in signalling the place of articulation distinction for these consonants is examined. The complex and dynamic production mechanism of stops has interested many speech researchers and, as a result, there is a large body of perceptual studies on stop consonants in quiet. Thus we have the opportunity to contrast and compare our results with those from 'quiet' conditions.

It is hoped that this investigation will contribute towards a more complete and quantitative theory of speech perception. In addition, understanding the way listeners with normal hearing perceive speech in noisy environments can potentially provide insights into the perceptual mechanisms of listeners with hearing impairments, since it has been shown that speech reception performance of normal-hearing listeners under noisy conditions is similar to that of listeners with certain hearing impairments under quiet conditions (Zurek and Delhorne, 1987). Modeling auditory processes in noise has several other practical applications. For example, the results of masking in the human auditory system were used successfully to optimize the performance of a digital speech coder (Schroeder et al., 1979) and the same technique was later used in a high-quality audio coding scheme (Brandenburg and Johnston, 1990). Another possible application would be improving the performance of automatic speech recognizers under noisy conditions. Currently, the performance of speech recognizers deteriorates significantly at signal-to-noise ratios high enough for humans to hear and understand perfectly.

In the following sections a brief summary of background material is presented, followed by an outline of the general strategy adopted in the study.

1.1 Previous Studies

1.1.1 Masking

Reference will be made to two kinds of masking: within-band masking, where the frequency of a signal is within the bandwidth of a noise masker, and above-band masking, where the signal frequency is above the upper cutoff frequency of the masker.

Within-band masking and critical bands

The peripheral auditory system, which consists of the outer, middle, and inner ears, performs frequency analysis of sounds. The pioneering work of Fletcher (1940) suggested that the inner ear (cochlea) acts as a bank of filters with frequency-dependent bandwidths; that is, the bandwidth of each filter depends on its center frequency. These bandwidths were found as a result of masking experiments done by Fletcher in 1940 who labeled the bandwidths as 'critical bands'. In his experiments Fletcher determined the masked thresholds of pure tones when presented in white noise and hypothesized that a) only those frequencies surrounding the tones in its 'critical band' contribute to its masking, and b) the signal power is equal to the noise power in a critical band at the masked threshold or, equivalently, the critical band is determined by the ratio S_m/N_0 where S_m is the signal power at the masked threshold and N_0 is the noise power per unit bandwidth or the spectrum level of the masker.

While the first hypothesis is well accepted, the second has been challenged. The bandwidth values obtained from Fletcher's indirect method of determining the masked thresholds of tones have been found to be 2.5 times narrower than those obtained from other more direct measures of bandwidth. These more direct measures resulted from a variety of psychoacoustic tasks such as masking experiments in which the threshold of a tone in the presence of noise of different bandwidths is measured (Hawkins and Stevens, 1950; Greenwood, 1961), two-tone masking experiments (Zwicker, 1954), and others. The value of the critical band as a function of frequency obtained from different sets of measurements is shown in Figure 1.1 (Moore, 1982). The figure shows that the bandwidth estimates from different studies are remarkably similar. The values obtained by Fletcher were later referred to as 'critical ratios' while the values obtained from the direct methods were referred to as 'critical bands'. Nevertheless, the ratio S_m/N_0 is a relevant parameter in 'within-band' masking experiments, where the tone frequency is within the bandwidth of the masker. For example, when a tone is presented in a background of white noise it is possible to predict a priori the power of noise per Hertz (N_0) needed to just mask a tone of a certain frequency and level (S_m) using the known values of S_m/N_0 at the masked threshold. For example, 10 log S_m/N_0 for 570, 1000, and 2500 Hz are 17, 18, and 21 dB, respectively (Reed and Bilger, 1973). The relationship between the masked threshold of a tone and N_0 is linear. That is, an increase in N_0 (in dB) results in the same increase in the masked threshold.

Masking of a tone can also occur if the masker and the tone are not simultaneous. There are three kinds of masking depending on whether the masker is preceded by, followed by, or simultaneously presented with, the probe tone. These three kinds of masking are called backward, forward, and simultaneous masking, respectively. The effects of simultaneous masking are the most well-studied of the three types, and it has been suggested that for complex signals such as speech the effects of simultaneous masking override the effects of non-simultaneous masking (Sorin, 1987). The critical-band theory, although a simplification for peripheral auditory processes, relates to *simultane-*

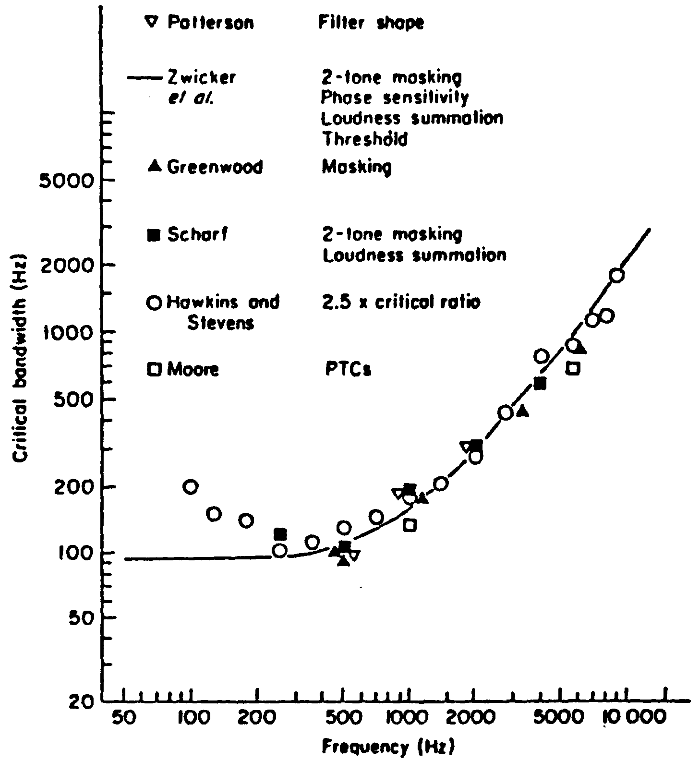


Figure 1.1: The value of the critical bandwidth as a function of frequency. The results of different sets of measurements are shown (Moore, 1982).

ous tone-in-noise masking when the noise is unmodulated band-limited white Gaussian noise. It is difficult to relate this theory directly to backward or forward masking.

Above-band masking and upward-spread of masking

Above-band masking refers to masking of a probe tone at a frequency that is above the upper edge of the noise masker. It is as if there is a virtual or 'effective' spectrum of the masker which extends beyond its physical spectrum. This kind of masking is highly dependent on the masker level and on the spacing between the upper edge of the masker and the tone. The nonlinear growth of above-band masking with masker level is referred to in the literature as 'upward-spread of masking'. An example of above-band masking is a case where the upper cutoff frequency of a low-pass noise masker is at 1200 Hz, and a tone is at 2500 Hz. If the overall level of the noise masker is 80 dB

SPL and the resulting masked threshold of the tone at 2500 Hz is 25 dB SPL, then due to upward-spread of masking, a 10 dB increase in the noise level will not necessarily result in a 10 dB increase in the masked threshold of the tone; the masked threshold will probably increase by more than 10 dB.

The earliest experiments which showed the effects of upward spread of masking of tones by narrow bands of noise were those by Bilger and Hirsh (1956), Carter and Kryter (1962), and Zwicker (1963). Curve-fitting procedures were developed to describe the data of Bilger and Hirsh, and of Carter and Kryter, and these procedures became part of an algorithm for predicting speech intelligibility in noise (ANSI procedures, S3.5-1969). The ANSI procedures predict the slope of the effective spectrum of a noise masker if the spectrum level and cut-off frequency of the masker are known. Zwicker's data, on the other hand, were accounted for by an equation proposed by Ludvigsen (1985). Ludvigsen was interested in estimating upward spread of masking in hearing-impaired subjects. He found that inserting a term, which simulated hearing loss, in his proposed equation resulted in good predictions of the performance of hearing-impaired subjects in the presence of a low-pass masker.

If the effective spectrum of the masker is known, then the masked threshold of a tone at frequencies higher than the upper edge can be computed by adding the critical ratio of the tone to the effective noise spectrum level at that frequency. To the knowledge of the author, the only study which compared the predictions of either procedure to experimental data from normal subjects was that of Rankovic et al. (in press). In that study, the ANSI procedures were shown to be successful in predicting some aspects of the masking patterns obtained from five subjects. However, masked threshold predictions were in some cases off by 10 dB from the experimental results.

To summarize, two methods can be used to predict upward spread of masking: ANSI procedures and Ludvigsen's equation. These methods have not been compared to determine their accuracy in threshold predictions.

1.1.2 Speech Perception in Noise

In this section relevant studies dealing with the perception of speech sounds in noise by normal-hearing listeners are reviewed. The focus is on those studies which examined the way fine phonetic/acoustic attributes are masked in the presence of noise.

The classic paper in this area is that by Miller and Nicely (1955) in which the effects of masking noise and filtering (both high-pass, and low-pass) on the perception of some consonants were examined. In that study, natural consonant-vowel syllables were presented to subjects in identification tests. The vowel was the vowel /ɑ/ (as in "father"), and the consonant was of one of the sixteen allowable syllable-initial consonants of English, with the exception of semivowels: /p,t,k,b,d,g,m,n,v,f,θ,ð,s,ʃ,z,ʒ/. The syllables were read from a randomized list by each of the five female subjects who participated in the experiments. The subjects rotated as speaker and listener within each of the three experimental conditions. In the first listening condition, the signal-to-noise ratio (SNR) was varied from −18 dB to +12 dB in 6 dB steps while keeping the speech bandwidth fixed between 200 and 6500 Hz. The SNR was varied by keeping the noise level fixed while varying the gain in the speech channel. The signal level was taken to be the peak deflection of the syllables on a VU meter monitoring the speech signals. In the second listening condition the speech was low-pass filtered at six different cut-off frequencies ranging between 300 and 5000 Hz, and in the third condition the speech was high-pass filtered with cut-off frequencies ranging between 1000 to 4500 Hz. The filtered speech was presented at a fixed SNR of +12 dB. Only the results of the masking-noise experiments will be discussed here.

Confusion matrices were obtained and the perceptual confusions were summarized by classifying the consonants according to a set of five linguistic features: voicing, nasality, duration, affrication, and place of articulation. Since the feature 'strident' was not used in their analysis, the authors chose to use duration as a feature because they believed that the strident fricatives /s,ʃ,z,ʒ/ are longer than the rest of the consonants and felt that the duration feature would help to set these four consonants apart from the rest. Figure 1.2 summarizes their results. In this figure the relative transmission (the ratio of transmitted to input information per 'channel' or feature) in percent is plotted as a function of the SNR. Results of three smaller experiments are also plotted in the figure. The smaller experiments used: 1) only stop consonants in syllable-initial position, 2) only stop consonants in final position following the syllable /tɑ/, and 3) only the eight fricatives in syllable-initial position. As shown in the figure, voicing and nasality were equally discriminable at an SNR as low as -12 dB, whereas place of articulation was hard to distinguish if the SNR was less than $+$ 6 dB. Affrication and duration were equally discriminable and were superior to place but inferior to voicing and nasality. The data also indicated that the features were perceived independently of each other since there was little 'cross talk' or interaction between the five feature channels.

Two informative pictures revealing the underlying structure of the original confusion matrices in the Miller and Nicely (hereafter referred to as MN) data are shown in Figure 1.3 (Shepard, 1972). Figure 1.3a shows the result of hierarchical-clustering analysis (using an algorithm developed by Johnson, 1967) when applied to the pooled data of the six SNR conditions. Five criterion levels (or minimum intracluster proximity) were chosen. Figure 1.3b, on the other hand, shows the effects of SNR on the consonant confusions at a certain criterion level (.17). We can infer from these two pictures that voicing and nasality are preserved well even under severe signal degradations (SNR=$-$12dB); this is in agreement with MN general conclusions. Place of articulation, on the other hand, is more salient for the nasals (/m/ versus /n/) than it is for the weak fricatives (/f/ versus /θ/). Neither duration nor affrication (features chosen by MN) seem to be effective in increasing or decreasing the similarity among consonants. It is

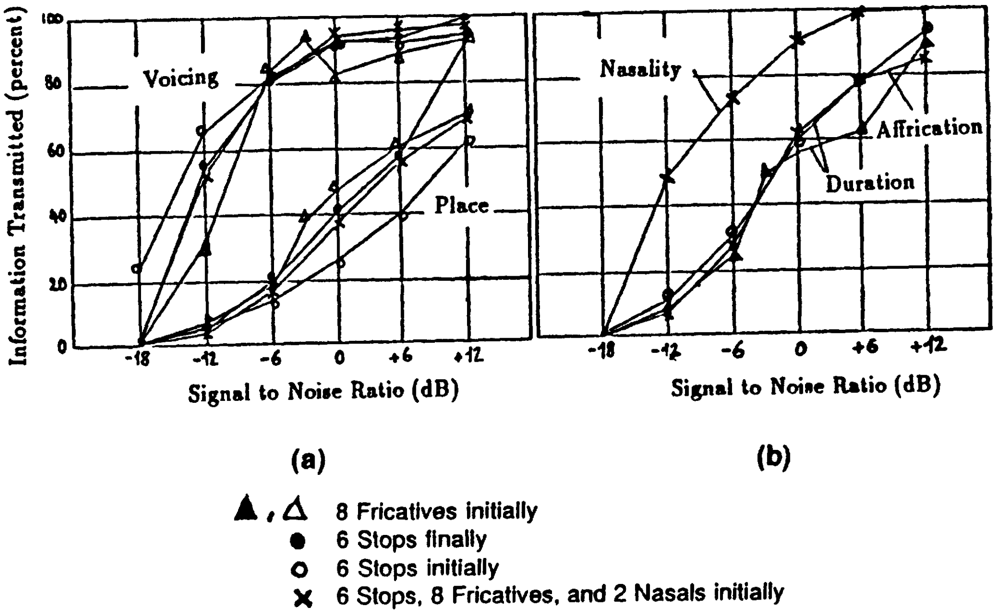


Figure 1.2: The relative information transmitted about a) voicing and place, and b) nasality, duration, and affrication as a function of SNR (Miller and Nicely, 1955).

important to keep in mind that MN chose five features to classify the consonants and then analyzed the data within a feature-based information theory approach, whereas Shepard used the raw data from the confusion matrices to find clustering patterns without postulating features.

Other researchers (Carroll and Wish, 1974; Wish and Carroll, 1974; Soli and Arabie, 1979; Soli, Arabie, and Carroll, 1986) have attempted to describe the perceptual confusions among consonants in MN data based on the acoustic properties of the speech signal rather than underlying phonetic features. The typical approach in these studies is to analyze the confusion data statistically and then interpret the dimensions which account for most of the variance, based on what is known about the acoustic attributes of the phonemes. These attributes were presumably based on the canonical forms of these consonants because the material used in MN data was not recorded and hence, not available for subsequent acoustic analysis. The salient properties under severe noise

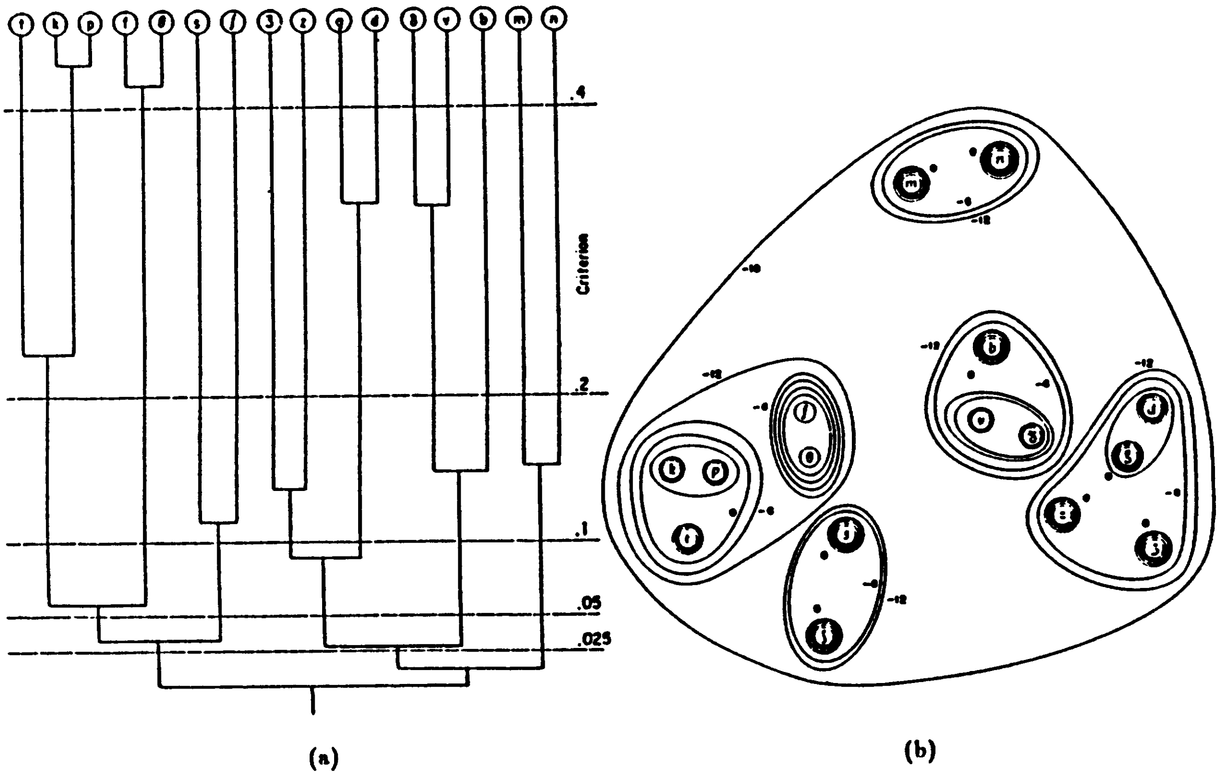


Figure 1.3: a) Hierarchical clustering representation for 16 consonants based on the pooled data of Miller and Nicely, b) representation of the effect of SNR on confusions among the consonants. At each SNR a closed contour is drawn around the consonants which are confused together at a criterion level of .17 (Shepard, 1972).

conditions were found to be voicing, signaled by low-frequency energy, and nasality, signaled by a nasal resonance (as defined by MN). Another property which was found to be salient under mild degradation conditions was the movement of the second formant frequency (F2) from the consonant to the vowel for the voiced consonants. A rise or fall in F2 was taken to be indicative of a change in place of articulation (Wish and Carroll, 1974; Soli et al., 1986). However, a close examination of their data shows that the rise/fall in F2 signals only a labial/non-labial place distinction. Gradual versus abrupt onset, a cue for the stop/fricative distinction, was found to be highly susceptible to noise for voiced consonants (Soli et al., 1986) and less so for the voiceless consonants. The results are interesting in so far as they are attempts to interpret the analysis results based on what is known about the articulatory-acoustic transformation for certain phonemes. However, the weakness of these studies is that the confusion matrices were explained with a set of acoustic attributes the choice of which was not based on acoustical analyses of the speech tokens used in the original experiments (since the speech material was not recorded), or by results of perceptual experiments.

Another study which examined consonant confusions in noise in terms of phonetic features is an extensive study by Wang and Bilger (1973). In their identification tests, subjects were presented with four sets of CV and VC nonsense syllables, recorded by an adult male speaker, both in quiet and in the presence of masking noise. The syllables represented all possible phonologically permissible VC and CV combinations of English consonants with the three vowels /ɑ,i,u/. There were six SNR conditions ranging from -10 to $+15$ dB, with a 5 dB difference in SNR between each condition and the next. As in the MN study, the authors characterized the phonemes in terms of phonetic features except that their feature set was much larger, containing nineteen features. Twelve of the nineteen were binary phonological features taken from Chomsky and Halle distinctive-feature description of English consonants (Chomsky and Halle, 1968). The rest of the features were taken from feature analyses of perceptual data by other researchers. The authors found a significant interaction between syllable position and vowel identity. For example, consonants accompanied by /u/ were easier to identify than those accompanied by /ɑ/. The effects of /i/ on the consonant intelligibility, on the other hand, depended on consonant position; /i/C syllables were easier to perceive than other VC syllables, whereas C/i/ syllables were the most difficult syllables to perceive.

The authors also found that the relative importance of perceptual features changes as a function of listening conditions (quiet versus noise-masked), and that this relative importance is not invariant across syllable sets. The perceptual relevance of a feature was measured by the percentage of transmitted information the feature accounted for. For example, the feature [high] was found to be the most intelligible feature in CV sets whereas the importance of this feature was diminished in VC sets. The only exceptions were the features [nasal] and [voice]. The robustness of these features was invariant across syllable sets and listening conditions. Thus, the authors concluded that articulatory and phonological features (with the exception of the features nasal, voice, and round) do not necessarily represent natural perceptual features. They cite examples where a feature is phonologically distinctive in a certain syllable set, yet data analysis of subjects' responses showed that the same feature has no perceptual relevance.

The common goal of all these studies was to find out if a phonetically- or acoustically-based feature system could account for the perceptual confusions among consonants. One of the few studies which attempted to approach this issue in a more quantitative way was a study by Farrar et al. (1987). In that study, the authors were able to predict results of discrimination experiments in noise based on a model of the auditory system. The signals used were broadband noises with spectral shapes appropriate for the unvoiced consonants /p,t,k,f,s,ʃ/ modeled after natural spectra of these consonants preceding back unrounded vowels. The spectrum of the noise masker resembled an average long-term speech spectrum. The duration of the frication noise was 300 ms and of the noise bursts, 30 ms. In addition, four durations (10, 30 , 100, 300 ms) of the noise bursts were tested for the consonants /p,t/. In the discrimination tests, listeners were presented with pairs of stimuli; each pair differed only in a place of articulation feature (for example, /p/ vs. /k/, /f/ vs. /ʃ/, etc.). A filter-bank model of the auditory system was then used to process the stimuli. The stages implemented in the model involved: filtering into non-overlapping frequency bands, estimation of average power, logarithmic transformation of the power estimates, addition of internal Gaussian noise, and ideal central processing. Results of the discrimination tests were predicted well by this model, with the exception of results illustrating durational effects. It would be interesting to see how well the model predicts the results of identification tests and how the predictions vary if line spectra, rather than continuous spectra, are used as stimuli.

Several measures have been introduced in the literature to predict the speech-recognition ability of normal-hearing and hearing-impaired listeners. The most commonly used measure is the Articulation Index (AI) (French and Steinberg, 1947). The basic idea behind the AI is that the speech-to-noise ratio in different frequency bands of the speech spectrum is indicative of average intelligibility. If the speech-to-noise ratio in a particular band is above a threshold value, then that band contributes to the overall speech recognition performance. In addition, different frequency bands have different weights (importance). The different weights have been determined empirically. Several researchers (e.g., Humes et al., 1986; Rankovic et al., in press) have shown that with proper modifications to the AI to accommodate different experimental conditions (speech materials and levels, kinds of degradation, sets of talkers and listeners, etc.) a correlation can be found between the overall performance of listeners and the AI. However, since the long-term average speech spectrum is used to compute the AI, the index does not provide a precise measure for predicting when perceptual confusions among speech sounds will occur; this limitation is especially true if the acoustic correlates of fine phonetic distinctions occur over short periods of time -as is usually the case- and do not significantly affect the long-term average of the speech waveform.

1.1.3 Perception of Place of Articulation for Stop Consonants in Quiet

Stop consonants are good examples of the one-to-many articulatory-to-acoustic transformations that characterize the production of speech. The perceptual mechanism by which the different acoustic cues are decoded and integrated by listeners is still not fully understood. Several theories have been proposed to explain the perception of stop consonants. In particular, place-of-articulation cues for stops have received wide attention from speech researchers since the fifties. In this section, theories which attempt to describe quantitatively the perception of stop consonants are discussed.

One theory of stop perception claims that F2 transitions are sufficient cues for signalling the place of articulation for stops (Delattre et al., 1955). This theory was based on perceptual experiments with two-formant synthetic CV syllables in which the place of articulation for each stop consonant was identified with a particular F2 transition. The F2 pattern associated with each stop was context-dependent and was assumed to point to a virtual 'locus' at a particular frequency. However, only F2 trajectories for /d/ seemed to originate from a common locus, which was at 1800 Hz. Despite its failure in explaining how stops are perceived, the locus theory raised a lot of interest in examining F2 to determine stop place. Most recently, a study by Sussman et al. (1991) found that F2 values at the onset and at the mid-point of the vowel in natural CV/t/ tokens can be used to classify correctly place of articulation for the stops /b,d,g/ in 10 vowel environments. A discriminant analysis using derived slopes and y-intercept of the $F2_{onset}$ and $F2_{vowel}$ values led to near perfect classification of the stop place. The authors hypothesized that these context-free acoustic cues (slopes and y-intercepts) are perceptually relevant. Although F3 was not used in their classification scheme, the authors note that "preliminary graphic analyses have implicated a greater F3 onset role in segregating stop place in high front vowel environments (p.1322)."

A different kind of theory was proposed by Blumstein and Stevens (1979, 1980). This theory hypothesized that the invariant (with respect to context) acoustic/perceptual cue for each stop consonant is the gross short-time spectral characteristics in the 10-20 ms following consonantal release. The theory was based on an acoustic study and a series of perceptual experiments with five-formant synthetic CV stimuli synthesized with and without bursts.

A study by Searle and his colleagues (1979, 1980) showed that features extracted from spectral displays processed by one-third octave filters (approximating auditory tuning curves) of speech signals provided cues for voicing and place information for word-initial stop consonants. Discrimination accuracy based on these auditory-based, running spectra was 77%. Inspired by Searle's study, Kewley-Port and colleagues (1983) claimed that three time-varying properties, and not a static representation like that proposed by Blumstein and Stevens, in the initial 20-40 ms following consonantal release contains context-free place information. The three time-varying spectral properties, derived from an earlier acoustical study by Kewley-Port (1983), were: spectral tilt of the burst, the existence of a mid-frequency peak sustained for at least 20 ms, and a delayed F1 onset value.

Both theories (Blumstein and Stevens (1979, 1980) and Kewley-Port et al. (1983)) were evaluated by Lahiri et al. (1984) in a cross-language study of voiced and voiceless diffuse-stop consonants (labial, dental, and alveolar). The conclusion was that metrics derived from the two theories mentioned above do not classify adequately the stops in the three languages studied (French, Malayalam, and English), and a different metric for classifying these consonants was proposed. The metric was based on the distribution of spectral energy from the burst release to the onset of voicing. Lahiri et al. claimed that changes in spectral energy constitute an invariant acoustic and perceptual property for the stops studied. The perceptual significance of the metric was verified in perceptual experiments with synthetic /b/V and /d/V stimuli in five vocalic environments.

In summary, the above theories agree that the interval following consonantal release carries important place information. Whether the different acoustic cues are processed independently, integrated at one point in time, or evaluated at different points along the transition from the consonant to the vowel is still not known.

1.2 Thesis Outline

The literature review shows that predicting perceptual confusions of speech sounds in noise has been an important area of research since the fifties, yet it still is an unresolved problem. This study attempts to develop procedures for predicting such confusions 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. This level of noise should lead to confusions in listener responses to that phonetic contrast. For this purpose, a quantitative feature-based approach is adopted in which the methodology is threefold: 1) quantifying acoustic correlates of some features in naturally-spoken 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 to evaluate the theoretical predictions.

In this study, these principles are examined for the place of articulation for the stop consonants /b,d/ in consonant-vowel syllables, where the vowel is either /ɑ/ or /ɛ/. The emphasis here is on examining the perceptual role of the formant trajectories in signalling the place-of-articulation distinction for these consonants. Nonsense syllables are chosen for the perceptual study so that lexical effects such as word frequency do not bias subjects' responses. Computer-generated, rather than naturally-spoken, syllables are used to provide better control of the stimuli.

First, the acoustic properties of the stop consonants /b,d/ imbedded in naturally-spoken

CV syllables are quantified. The results of the analysis are then used to synthesize these utterances using the formant synthesizer KLSYN88 (Klatt, and Klatt, 1990). Since we are interested in the perceptual role of only the formant trajectories, burstless utterances are synthesized. The analysis/synthesis stage is described in Chapter 2.

The analytical tools necessary to predict the masking of formant frequencies of the synthetic CV utterances in noise are developed in Chapter 3. The tools are based on a combination of theoretical and empirical results. Two types of masking are studied: within-band masking (where a tone or formant is within the bandwidth of the noise masker) and above-band masking (where a tone or formant is above the upper cutoff frequency of the masker). Theoretical predictions for the within-band masking case are based on results of auditory masking theory. In this case, a prediction as to whether a tone is masked or not depends on the signal-to-noise ratio in a critical band around that tone. In the above-band case, experiments are conducted and the two methods known for predicting aspects of this kind of masking (ANSI standards (1969) and Ludvigsen's equation (1985)) are compared with the experimental results.

The final stage of the study, which is described in Chapter 4, involves conducting a series of identification tests with synthetic CV utterances in noise. Two noise maskers are used in the experiments: white noise, and band-pass noise centered around the F2 region. In the white-noise case, all formant frequencies are within the bandwidth of the masker, whereas in the band-pass case only F2 is within the masker bandwidth while F3 and higher formants are above the upper edge of the masker. In the latter case, the higher formants will be masked at certain levels of the noise masker due to above-band masking. Analysis of the subjects' responses is done with the aid of the analytical tools developed in Chapter 3.

A summary and assessment of the experimental results along with a discussion of the limitations of the experiments are presented in Chapter 5. The final chapter also includes ideas for future work in this area.

Chapter 2

Analysis and Synthesis of the Stop Consonants /b/ and /d/

In this chapter, results of acoustic analysis of natural consonant-vowel (CV) utterances will be discussed. Based on these results, a set of parameters is selected to generate synthetic utterances. The synthesis is then implemented using a software formant synthesizer KLSYN88 (Klatt and Klatt, 1990). The synthetic utterances are later used in perceptual experiments which are described in Chapter 4.

2.1 Analysis

2.1.1 Corpus and Recording Method

The corpus used for the analysis consisted of a set of nonsense C/ɑ/ and C/ɛ/ utterances. The consonant was either /b/ or /d/. The purpose of the acoustic analysis was to generate a set of parameters which could be used to synthesize the utterances with the aid of a formant synthesizer. Data from only one speaker were analyzed since it has been shown that using the data of one speaker yields better synthetic stimuli than using average data from several speakers (Klatt, in preparation). A male native speaker of American English (KNS) served as a subject.

Each CV utterance was written on three different index cards. The cards were then shuffled and presented to the subject who was asked to read what was written on each card keeping his intonation the same on each syllable. There was a total of 12 utterances. The utterances were recorded in a sound-treated room using an Altec microphone, a Shure microphone mixer and a Nakamichi Lx-5 tape recorder. The recordings were made using TDK D60L tapes without the use of Dolby or DBX. The speech material was then digitized at 10 kHz using a low-pass filter with a cut-off frequency of 4.8 kHz.

2.1.2 Analysis of Corpus

Method

The main purpose of the analysis was to examine the formant trajectories associated with the consonants /b/ and /d/. Analysis of the natural utterances was done using KLSPEC, a software package developed by Klatt (1984). All spectral representations were computed using a Hamming window. Broad-band spectrograms were made by computing a discrete Fourier transform every 1 ms with a 6.4 ms window and smoothed using Gaussian filters of width 300 Hz.

Results

Broad-band spectrograms of one repetition of each of the CV utterances are shown in Figure 2.1. As illustrated in those spectrograms, the stops are characterized by (a) a voice bar, (b) a brief burst of noise at the release of the consonant and (c) particular formant trajectories from the consonant to the vowel.

The voice bar is a low-frequency energy signal radiated from the neck during the occlusion of the voiced stops and is similar acoustically for both consonants. Both the spectrum of the burst and the trajectories of the formant frequencies depend on the vocal-tract shape during the production of the stop consonants and hence are cues for

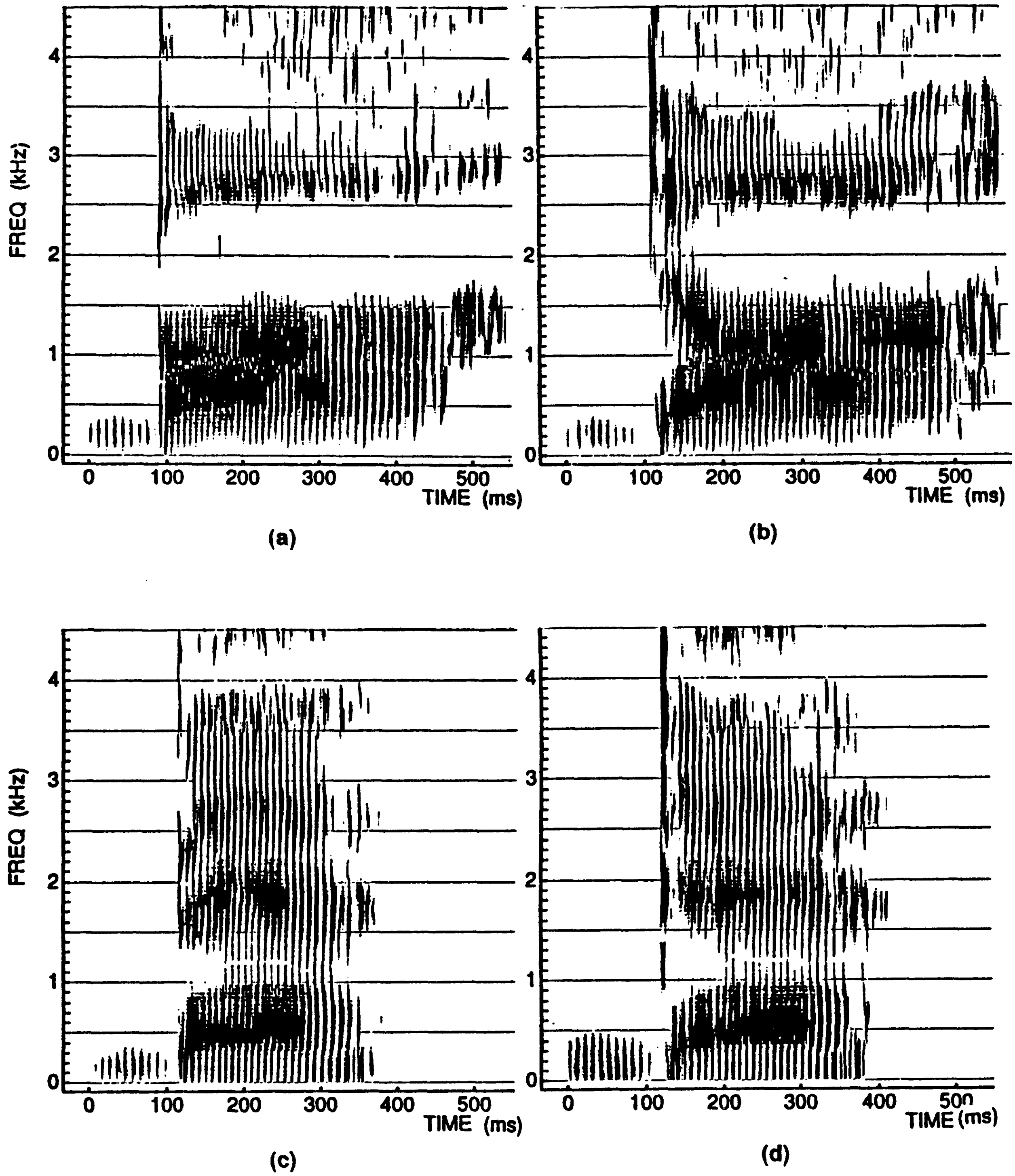


Figure 2.1: Spectrograms of the natural utterances: (a) /bɑ/, (b) /dɑ/, (c) /bɛ/, and (d) /dɛ/ as spoken by a male speaker.

their place of articulation. The spectrograms show how the formant trajectories are different for the two consonants. With the vowel /ɑ/, F2 rises into the vowel for the labial and falls for the alveolar and with the vowel /ɛ/, both F2 and F3 rise into the vowel for /b/ and are relatively flat for /d/. F1 rises from the consonant into the vowel in all cases. Figure 2.2 shows spectra sampled at the burst (solid lines) overlaid with spectra sampled at the mid-point of the following vowel (dashed lines) of natural CV utterances. The spectra were computed with a Hamming window of duration 6.4 ms and the signals were not preemphasized. The short window was chosen to capture the transient nature of the burst. The noise burst excites only resonances associated with the cavity in front of the constriction (referred to as front-cavity resonances), and hence the spectral shape of the labial is relatively flat (no front cavity) whereas the spectrum for /d/ has prominences at high frequencies (short front cavity).

Formant frequency locations and amplitudes were estimated from short-time Discrete Fourier Transform (DFT) spectra. The analysis window was 128 samples (or 12.8 ms for a 10kHz sampling rate) long and the computations were done every pitch period. The estimation was performed on 'burstless' natural utterances in which the bursts were sliced off. The reason for burst removal was to estimate the frequencies at the vowel onset more accurately (otherwise, burst information could be averaged together with the vowel onset). The range of values of the first three formant frequencies at the onset of the vowel (Fi_o) and average values at the midpoint of the vowel (Fi_m) for the natural CV utterances are listed in Tables 2.1 and 2.2. Formant transition time was considered to be the time from the onset of the vowel to the time when the formant frequency first reached its steady-state value. Measurement of transition times was done mainly with the aid of spectrographic displays. The average transition time for the /bɑ/ utterances was 30 ms for all formant frequencies, and for the /dɑ/ utterances the average times were 50, 65, 60 ms for F1, F2, and F3, respectively. For the /bɛ/ utterances, the transition times were 25, 30, 40 ms for F1, F2, and F3 respectively and the F1 transition for /dɛ/ was 30 ms. F2 and F3 trajectories for /dɛ/ were relatively

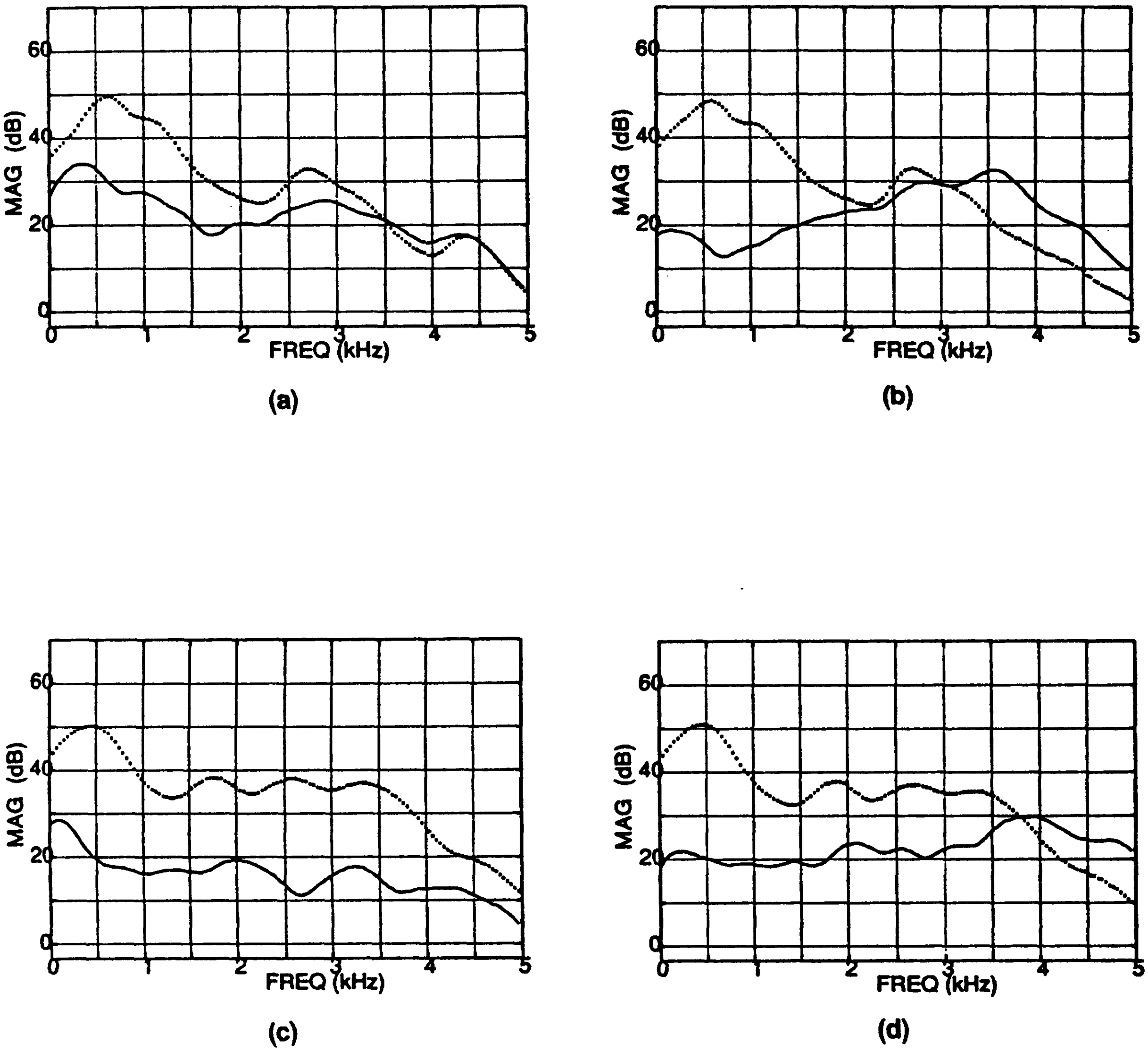


Figure 2.2: Smoothed-DFT spectra sampled at the release of the stop burst (solid lines) superimposed with spectra sampled at the mid-point of the following vowel (dashed lines). The spectra were sampled in the natural utterances: (a) /bɑ/, (b) /dɑ/, (c) /bɛ/ and (d) /dɛ/. The analysis used a 6.4 ms Hamming window. Smoothing was done with a 300 Hz wide filter. The signals were not preemphasized.

Table 2.1: Range of values for the first three formant frequencies at the onset of the vowel for the natural CV utterances. There were three tokens for each utterance.

Range of formant frequencies (Hz) at vowel onset				
	C/ɑ/		C/ε/	
	b	d	b	d
F1	350-480	280-350	320-450	320-380
F2	940-960	1500-1650	1550-1640	1800-1900
F3	2400-2600	2600-2700	2300-2400	2700-2750

Table 2.2: Average values of formant frequencies at the midpoint of the vowels.

Formant Frequencies (Hz) at midpoint of the vowel		
	/ɑ/	/ε/
F1	610	510
F2	1100	1800
F3	2700	2700

flat.

Plots of the time course of the relative amplitudes of each of the spectral prominences corresponding to the first three formant frequencies in comparison with the amplitude of F1 at the steady-state part of the vowel are shown in Figures 2.3 and 2.4. Each point on the plots is an average value across three tokens. Notice that with the vowel /ɑ/ (Figure 2.3) the time function of the relative amplitude of F1 is similar for the two consonants, the amplitude of F2 for /b/ is higher than it is for /d/ by 6 to 10 dB in the initial 15 ms, and the relative amplitude for F3 is higher for /d/. With the vowel /ε/ (Figure 2.4), the change in formant amplitudes between the consonant and the vowel is small.

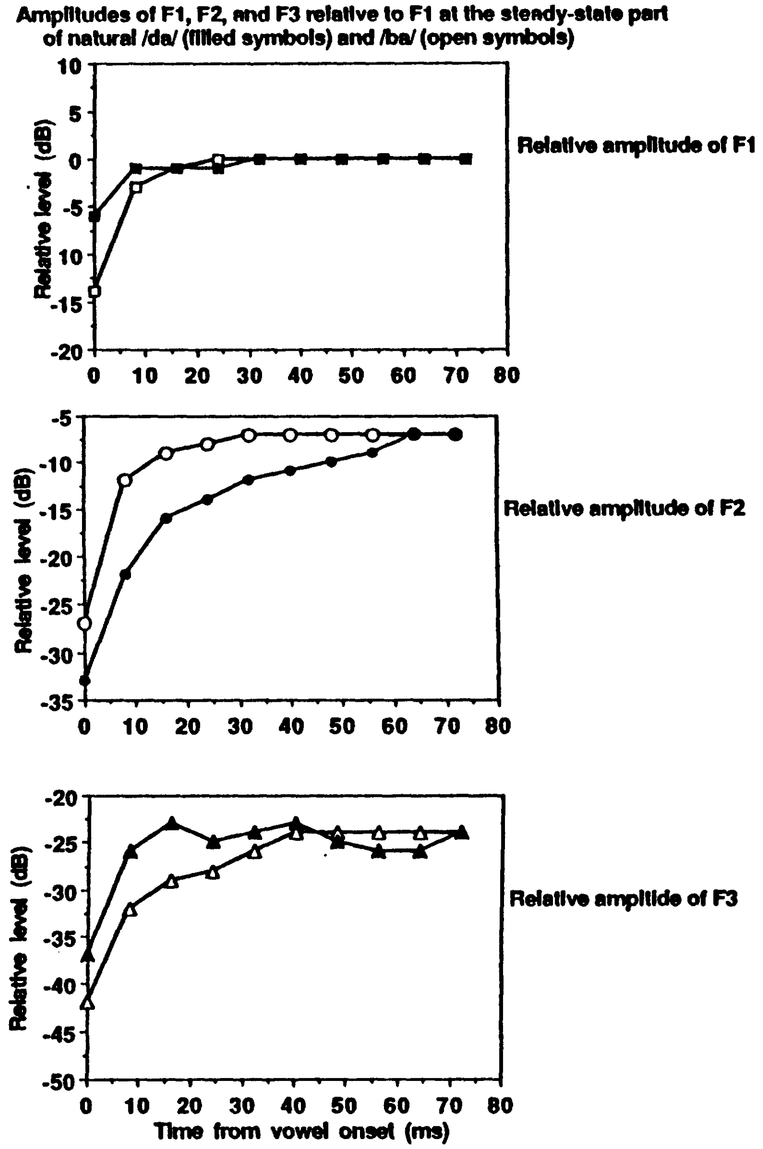


Figure 2.3: Plots of the relative amplitudes in dB of the first three formant frequencies in comparison with that of F1 at the steady-state part of the vowel for the /ba/ (open symbols) and /da/ (filled symbols) natural utterances. Estimates were based on DFT spectra computed every pitch period. Each data point is an average value from three tokens.

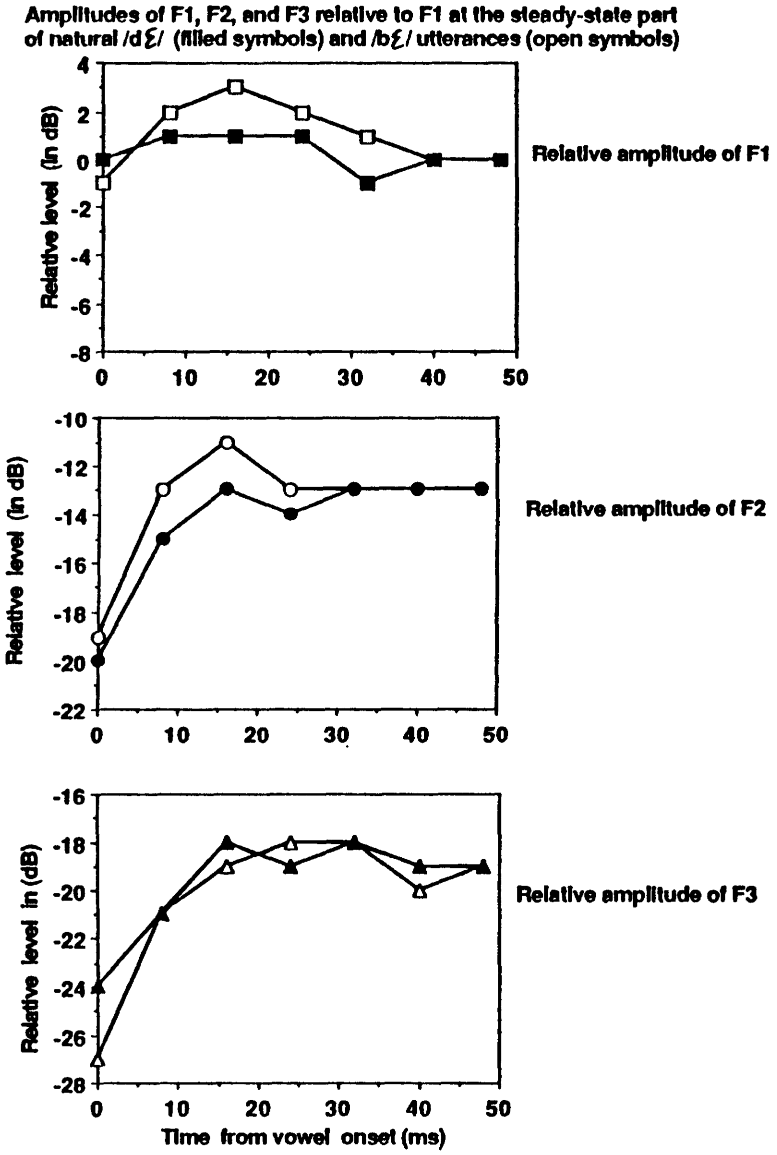


Figure 2.4: Plots of the relative amplitudes in dB of the first three formant frequencies in comparison with that of F1 at the steady-state part of the vowel for the /bɛ/ (open symbols) and /dɛ/ (filled symbols) natural utterances. Estimates were based on DFT spectra computed every pitch period. Each data point is an average value from three tokens.

The difference in formant amplitudes between the two consonants is due mainly to differences in formant trajectories. This point will be elaborated on further in the following sections.

2.2 Synthesis

Synthetic CV utterances were generated using the cascade part of the formant synthesizer KLSYN88 (Klatt and Klatt, 1990). A block diagram of the synthesizer is shown in Figure 2.5. The cascade/parallel parts of the synthesizer refer to the cascade/parallel connections of the digital resonators that simulate the vocal-tract transfer function. If the cascade part of the synthesizer is used, the relative amplitudes of the formant peaks are adjusted automatically according to the acoustic theory of speech production (Fant, 1960). The use of the parallel part of the synthesizer requires adjustments of the amplitudes of individual formant peaks. Non-nasal sonorants are synthesized best with the cascade part, while the parallel parts are used for synthesis of sounds characterized by poles and zeros in the vocal-tract transfer function (such as nasals, and fricatives) or for sounds which do not adhere to the normal amplitude relations such as one or two-formant vowels (Klatt, 1980).

Synthesis parameters did not necessarily mimic acoustic features extracted from natural utterances. Rather, the minimal-pair syllables (differing in place-of-articulation features only) had the same time-varying characteristics except for differences in one or two formant trajectories. Fine-tuning of synthesis parameters was such that the reference synthetic utterances were perfectly identifiable and discriminable in the absence of noise.

All synthetic syllables were 250 ms in duration. For each stimulus, the amplitude of voicing started abruptly, remained fixed for 240 ms, and then decreased by 10 dB in the last 10 ms. The fundamental frequency remained fixed at 125 Hz for 140 ms, and

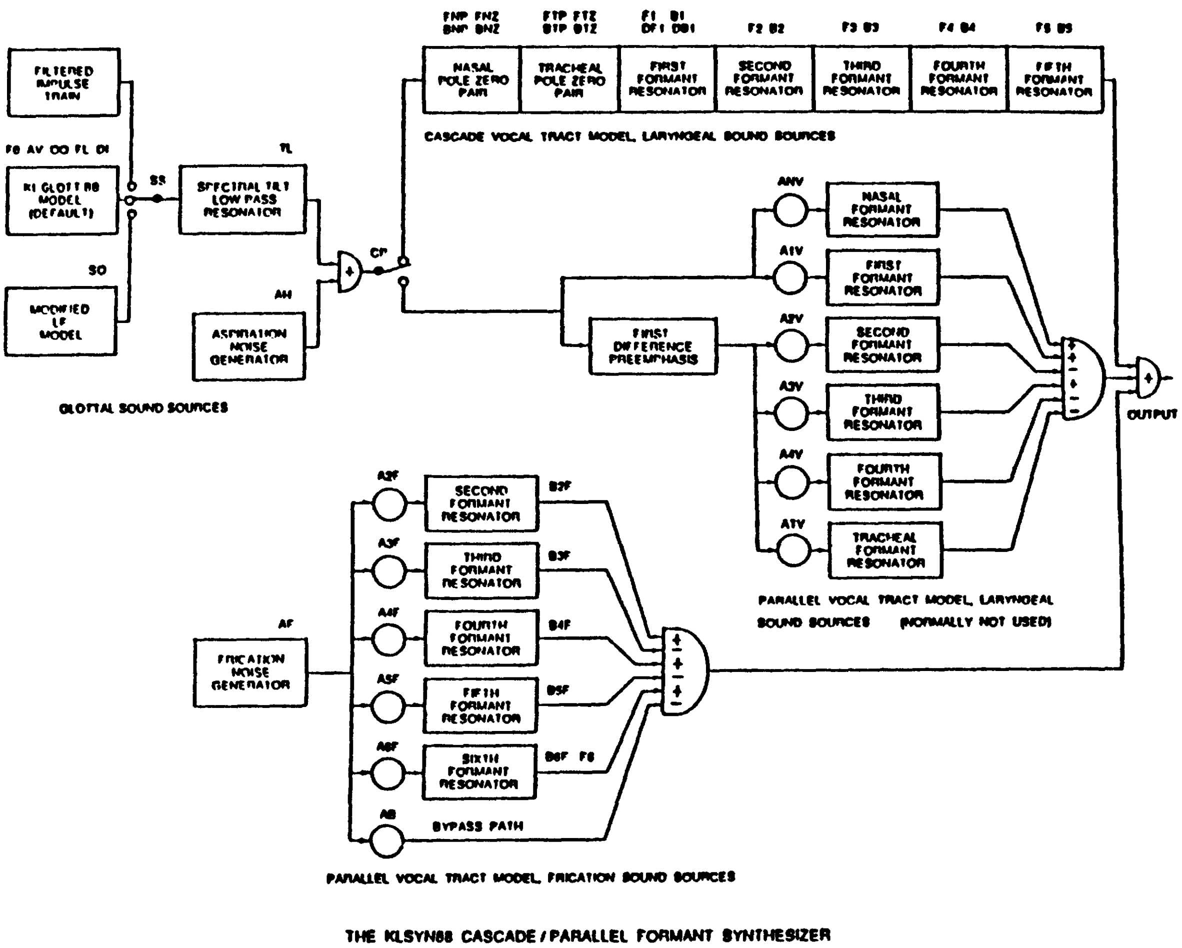


Figure 2.5: Block diagram of the Klatt cascade/parallel formant synthesizer (Klatt and Klatt, 1990). The cascade part was used to generate synthetic CV utterances.

then fell to 100 Hz in the last 110 ms of the stimulus. The choice of abrupt onsets for the amplitude-of-voicing and fundamental-frequency contours was based on informal experiments in which listeners observed that abrupt, rather than gradual, onsets for these contours led to an enhanced stop-like quality.

2.2.1 C/ɑ/ Syllables

Synthetic /bɑ/ and /dɑ/ utterances were generated without stop bursts because we were interested in examining the perceptual importance of only the formant trajectories in distinguishing between /b/ and /d/. Since our analysis of natural data revealed that the main acoustic cue signalling the difference between the two consonants is the F2 trajectory, all synthesis parameters, except for F2, were the same for both /bɑ/ and /dɑ/.

The synthesis involved mainly specifying the formant trajectories and bandwidths. Figure 2.6 shows the formant trajectories for /bɑ/ (solid lines) and /dɑ/ (dotted lines). The formant trajectories varied in a piecewise-linear fashion which was an idealization of formant trajectories observed in natural speech. However, this idealization yielded natural-sounding stimuli. All synthesis parameters for the two utterances were the same except for F2. The formant frequencies at the steady-state part of the vowel were chosen to be the same as average values obtained from the analysis of natural speech (F1=610, F2=1100, F3=2700 Hz). Pilot experiments were conducted to help decide which onset values of the formants (within the range obtained from natural speech) and which transition times would yield the best-sounding stimuli. Two phonetically-trained subjects participated in these experiments. In open-response tests, the subjects were asked to identify the consonant in each syllable heard and rate its naturalness on a 1 to 3 scale. Based on the pilot experiments the following onset values were chosen for the synthesis: F1 and F3 were 470 and 2600 Hz, respectively, for both consonants, and

F2 was 970 Hz for /b/ and 1500 Hz for /d/. The transition times for F1, F2, and F3 were 25, 40, and 50 ms, respectively. F4 was kept fixed at 3300 Hz .

Estimates of formant bandwidths, on the other hand, are difficult to make directly from spectra, and hence an analysis-by synthesis approach was used. The bandwidth values were adjusted so that the relative levels of the formant peaks in the steady-state vocalic part of the synthetic utterances were similar to those of natural speech. The bandwidths were held constant at the following values: B1=100 Hz, B2=90 Hz, B3 and B4=200 Hz.

As a consequence of the different F2 trajectories for the two consonants, the amplitudes of the prominences of F2 and all other formant peaks were different as well. Figure 2.7 shows plots of the time course of the relative amplitudes of the first three formant peaks in comparison with that of F1 at the steady-state part of the vowel for the synthetic C/ɑ/ utterances. The open symbols in Figure 2.7 refer to measurements for the /bɑ/ utterance and the filled symbols, for /dɑ/. Formant peak amplitudes were estimated from the DFT spectra computed every pitch period using a Hamming window of duration 128 samples.

The change in formant amplitudes could be explained from the acoustic theory of speech production (Fant, 1960). The amplitude of the *ith* formant peak is proportional to the magnitude of the vocal-tract transfer function at that frequency (M_i). M_i can be approximated by (Alwan, 1986):

$$M_i \simeq 20 \log \frac{F_i}{B_i} \prod_{k=1,k\neq i}^{n} \frac{F_k^2}{(F_i - F_k)(F_i + F_k)} \tag{2.1}$$

where F_i is the *ith* formant-frequency location and B_i is the bandwidth. Hence, the amplitude of each formant peak depends on its location, its bandwidth, and the location of the other formant frequencies. The amplitude of F3, for example, for /dɑ/ during the transition period was higher than that for /bɑ/ (Figure 2.7c) because of the

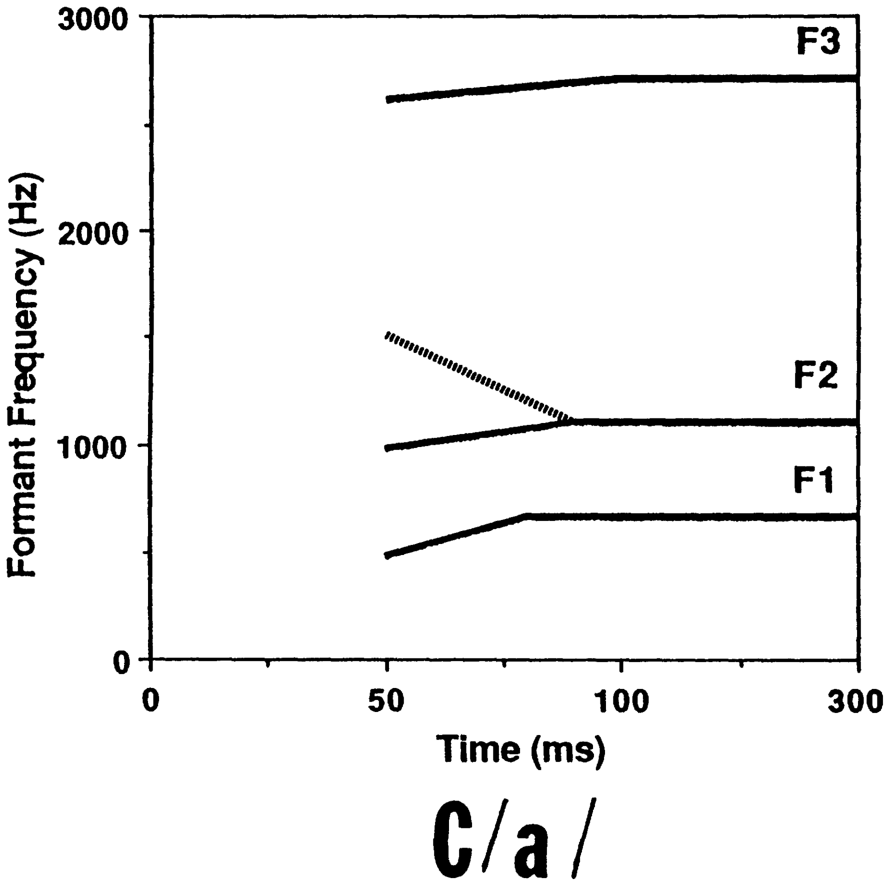


Figure 2.6: Schematized trajectories of the first three formant frequencies for the synthetic /ba/ (solid lines) and /da/ (dashed lines) utterances.

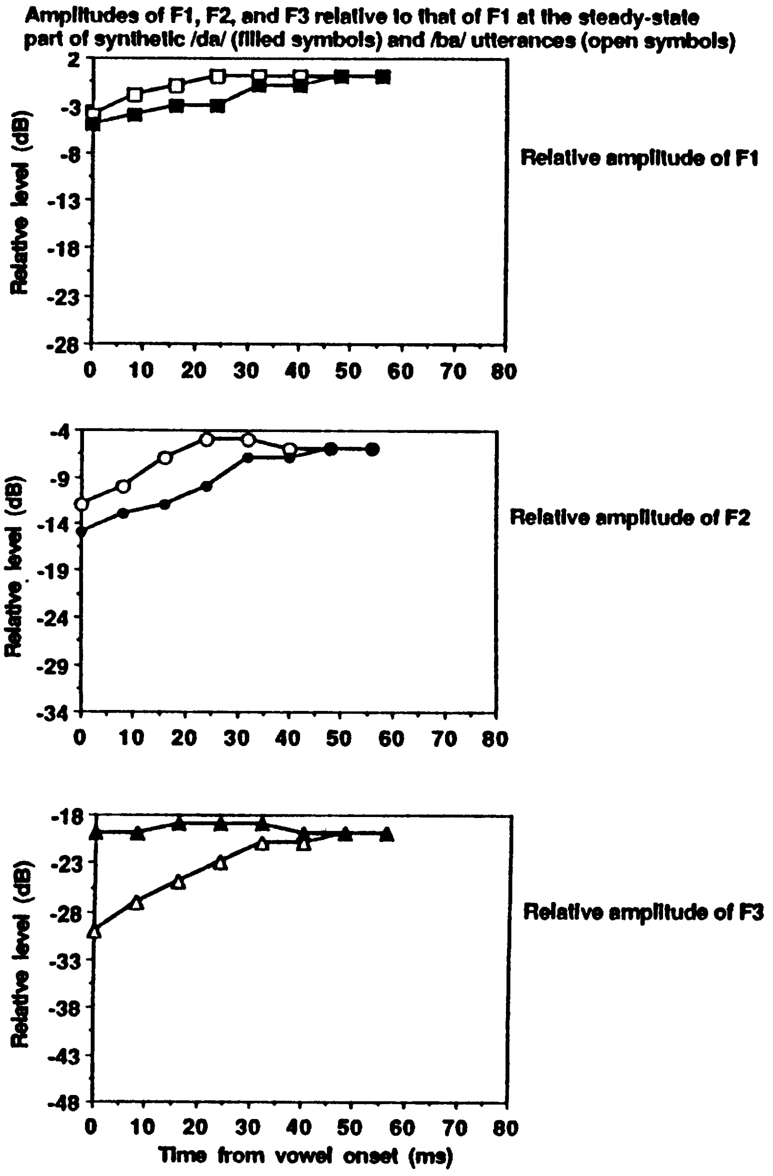


Figure 2.7: Plots of the relative amplitudes in dB of the first three formant frequencies in comparison with that of F1 at the steady-state part of the vowel for the /ba/ (open symbols) and /da/ (filled symbols) synthetic utterances. Estimates were based on DFT spectra computed every pitch period.

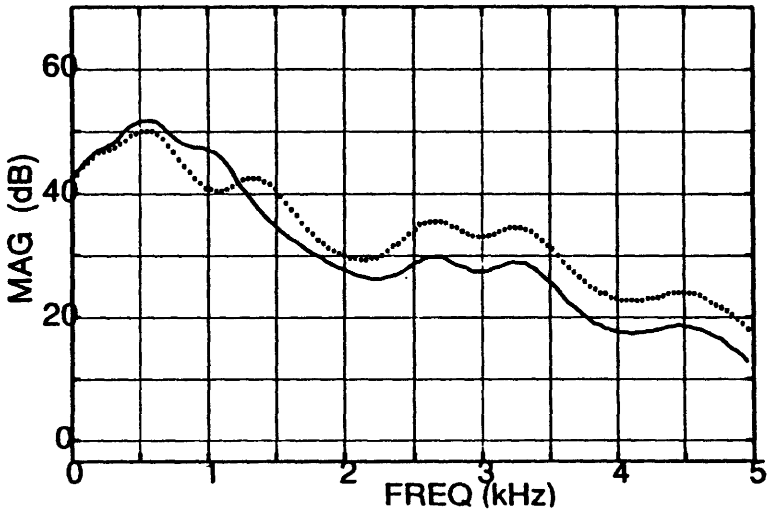


Figure 2.8: Smoothed DFT spectra sampled 16 ms into the vowel. The dashed spectrum was sampled in the synthetic /da/ utterance and the solid spectrum, in the /ba/ utterance. The smoothing was done with a 300 Hz wide filter.

higher F2 value and hence a greater F2-F3 proximity for the alveolar. Likewise, the relative amplitudes of F1 and F2 are higher for /ba/ than they are for /da/ (Figure 2.7 (a,b)) because of a greater F1-F2 proximity. Another illustration of the amplitude differences due to F2 movement is shown in Figure 2.8 where smoothed DFT spectra sampled 16 ms after the vowel onset for /b/ and /d/ are shown. The spectra were smoothed using a 300 Hz wide window.

If we compare the plots of the formant amplitudes for the synthetic and natural utterances (Figure 2.7 vs. 2.3) we notice that the main difference between the two sets is that in natural utterances there is a significant increase in the amplitude of each formant frequency, sometimes as much as 10 dB, from the initial period of voicing to the following pitch period. No similar jump occurs in the synthetic utterances. The reason for the amplitude increase is that in the natural utterances the first glottal pulse appeared to be considerably weaker than the other glottal pulses. In his theoretical study of the characteristics of stop consonants, Stevens (in preparation) showed that the initial glottal pulses following the release of stop consonants are weaker than the steady value of these pulses. Our results are in agreement with his observation and it appeared that the first glottal pulse was considerably weaker (especially in the C/ɑ/ case) than the following pulses.

The amplitude of voicing for the synthetic utterances, on the other hand, started abruptly to enhance the 'stop-like' property of the stimuli because no burst was synthesized. In other words, the synthesis did not account for an initial weak glottal pulse. However, if we consider only the formant amplitudes starting from the second pitch period of the natural utterances in Figure 2.3 and compare those with the formant amplitudes in the synthetic utterances (Figure 2.7) we observe a greater similarity between the two figures.

In open-response tests, ten repetitions of the synthetic /bɑ/ and /dɑ/ stimuli were presented in a random order in quiet to three phonetically-trained subjects who were asked to identify the consonant. The identification scores were 100% for /b/ and /d/. These utterances were considered appropriate as reference utterances for perceptual experiments.

2.2.2 C/ɛ/ Syllables

The synthesis procedure described in Section 2.2 was adopted for the C/ɛ/ syllables. Synthetic, burstless C/ɛ/ utterances were generated using the cascade part of KLSYN88. The synthesis involved mainly specifying the formant trajectories and bandwidths.

Figure 2.9 shows schematized trajectories for the synthetic /bɛ/ (dotted lines) and /dɛ/ (solid lines) utterances. Note that the two syllables differ in the trajectories of

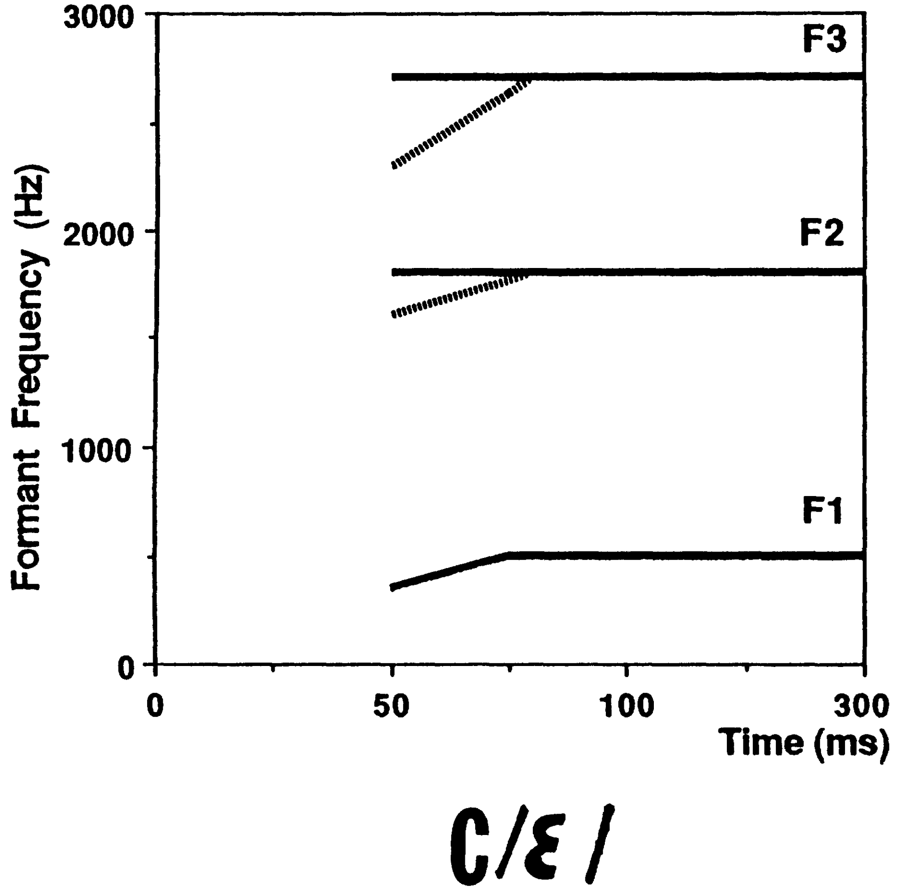


Figure 2.9: Schematized trajectories of the first three formant frequencies for the synthetic /bɛ/ (dashed lines) and /dɛ/ (solid lines) utterances.

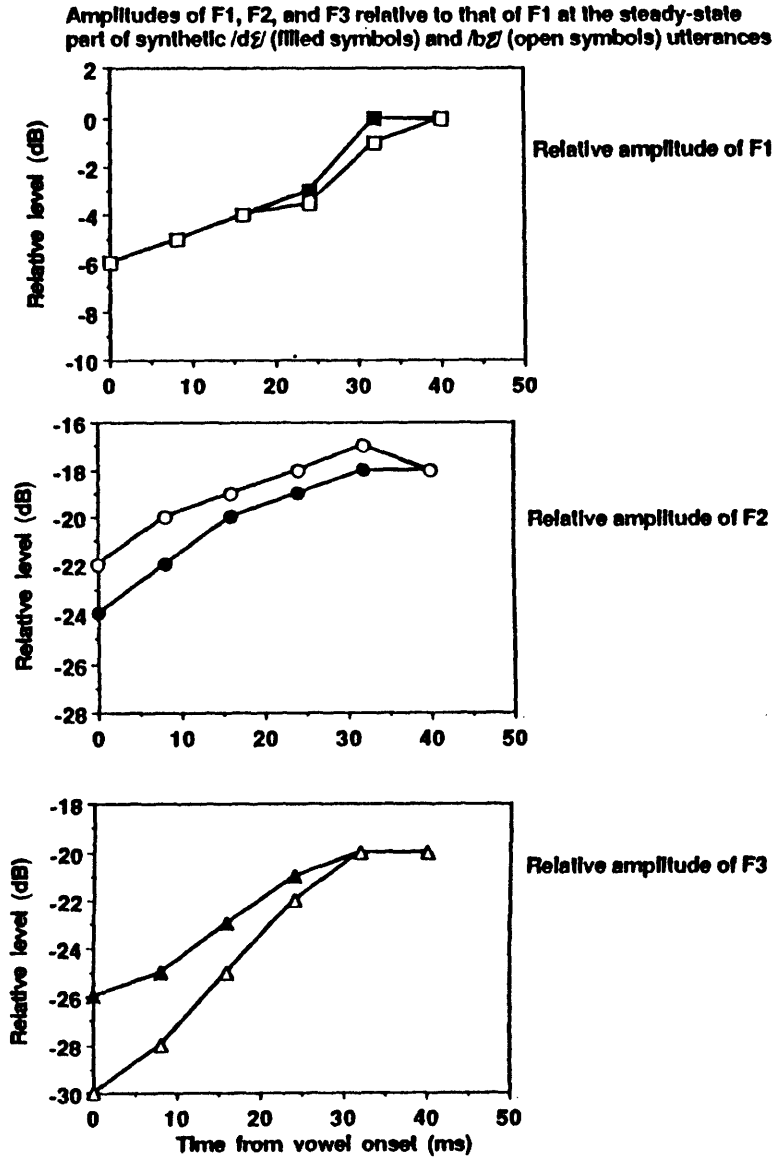


Figure 2.10: Plots of the relative amplitudes in dB of the first three formant frequencies in comparison with that of F1 at the steady-state part of the vowel for the /bɛ/ (open symbols) and /dɛ/ (filled symbols) synthetic utterances. Estimates were based on DFT spectra computed every pitch period.

both F2 and F3. Based on judgements by phonetically-trained subjects, a difference in F2 trajectory alone was not sufficient for the perception of either consonant (unlike the C/ɑ/ case). The onset values of the first three formant frequencies were: 350, 1600, and 2300 Hz for /bɛ/ and 350, 1800, 2700 Hz for /dɛ/. The transition times for F1, F2, and F3 were: 25, 30, and 30 ms, respectively. F4 was kept fixed at 3400 Hz. The steady-state values of the first three formant frequencies were: 500, 1800, and 2700 Hz, respectively. As in the C/ɑ/ case, the choice of optimum (in terms of naturalness) transition times and onset frequencies were determined based on the results of pilot experiments with two phonetically-trained subjects.

The bandwidths of F1, F2, and F3 were 60, 120, and 120 Hz, respectively. This choice of bandwidth values yielded relative amplitudes of formant frequencies in the steady-state part of the vowel which were similar to those observed in natural utterances. Plots of the relative formant amplitudes in comparison with F1 amplitude for the two utterances are shown in Figure 2.10. The synthetic /bɛ/ and /dɛ/ tokens were perfectly identified by two phonetically-trained subjects.

2.3 Summary

In this chapter we have quantified some of the acoustic properties for the stops /b/ and /d/ in the context of a back unrounded vowel /ɑ/ and a front vowel /ɛ/. Based on the results of the analysis, synthetic CV utterances were generated using the formant synthesizer KLSYN88. The synthetic utterances differed in F2 trajectory in the /ɑ/ case, and both F2 and F3 in the /ɛ/ case. We have shown how the different F2 (or F2 and F3) trajectories affect the amplitudes of all formant frequencies.

The following two chapters will examine the perceptual importance of the formant trajectories and their amplitudes in signalling the place distinction for the two consonants in noise.

Chapter 3

Masking of Tones and Formants: Theory and Experiments

As indicated in Chapter 1, one of the goals of this study is to examine the perceptual importance of the formant frequencies in signalling the place of articulation distinction for /b/ and /d/ in noise. The perceptual importance of the formants is investigated via experiments examining the perception of synthetic CV utterances in noise; results are described in Chapter 4. Two noise maskers are used in the experiments: white noise and band-pass noise centered around F2. In the case of a white-noise masker, all formant frequencies are within the bandwidth of the masker; in contrast, in the case of a band-pass noise masker, only F2 is within the bandwidth of the noise whereas F3 and higher formant frequencies are above the cutoff frequency of the masker.

In this chapter, the analytical tools which are developed to predict formant-frequency masking are described. For each case of masking (within-band and above-band) theoretical predictions of tone masking are first made based on relevant literature. Experiments with tones and synthetic vowels are then conducted to address two issues: 1) the validity of the theoretical predictions when applied to tone masking, and 2) the applicability of the theoretical predictions to the masking of formant frequencies in steady-state vowels.

3.1 Theoretical Predictions

3.1.1 Within-band Masking

One of the important results of masking theory is that the masked threshold of a tone in a broad-band noise background can be estimated accurately from the signal-to-noise power ratio in the critical bandwidth centered at the tone frequency.

The critical ratio of a tone can be expressed as (Fletcher, 1940):

$$CR(f) = T(f) - N_0 \tag{3.1}$$

where

$CR(f)$ = critical ratio in dB of a tone at frequency f

$T(f)$ = masked threshold of the tone in dB SPL (re 0.0002 μbar)

N_0 = noise spectrum level in dB SPL (re 0.0002 μbar).[1]

If the bandwidth of the noise masker is bw (in Hz), then Equation 3.1 can be rewritten as:

$$T(f) = [P_n - 10log(bw)] + CR(f) \tag{3.2}$$

where P_n is the overall noise power in dB SPL (re 0.0002 μbar).

As mentioned in Section 1.1.1, the critical ratio was found to be 2.5 times narrower (or 4 dB lower) than direct measurements based on estimates of the critical bandwidth. Hence, Equation 3.2 can be rewritten as:

[1] The noise spectrum level is the noise power density (noise power in a 1 Hz band) expressed in dB SPL (re 0.0002 μbar).

$$T(f) = [P_n - 10log(bw)] + 10log(cb(f)) - 4 \tag{3.3}$$

where $cb(f)$ is the critical bandwidth (in Hz) at frequency f.

If we define the tone to noise power ratio in a critical band ($SNRB$) to be the difference between the power of the tone (P_{tone}) and that of the noise masker in a critical band plus 4 dB, or:

$$SNRB = P_{tone} - [P_n - 10log(bw) + 10log(cb(f))] + 4 \tag{3.4}$$

then, masking occurs when the term $SNRB$ is equal to zero dB. In that case, P_{tone} is the masked threshold of the tone ($T(f)$).

In this study, short-time Discrete Fourier Transform (DFT) power spectra were used to estimate the signal-to-noise power ratio in a critical band ($SNRB$).

The overall power of a signal is proportional to its mean square value. For a finite-length signal x[m] of duration L samples, the overall power is proportional to $\sum_{i=0}^{L-1} |x[i]|^2$.

From Parseval's theorem (Oppenheim and Schafer, 1975) we have:

$$\sum_{i=0}^{L-1} |x[i]|^2 = \frac{1}{L} \sum_{i=0}^{L-1} |X[k]|^2 \tag{3.5}$$

where $|X[k]|$ is the magnitude of the DFT defined as:

$$X[k] = \sum_{i=0}^{L-1} x[i] e^{-j2\pi ki/L} \tag{3.6}$$

Hence, the overall power of the signal can be expressed as:

$$P_{overall} = \Lambda \sum_{i=0}^{L-1} |X[k]|^2 \tag{3.7}$$

where Λ is a proportionality constant. If the DFT for both the tone and the noise masker are computed with the same analysis window then Λ is the same for both. The $SNRB$ can then be estimated as:

$$SNRB = 10log \sum_{i=0}^{L-1} |\tilde{X}[k]|^2 - [10log \sum_{i=0}^{L-1} |\tilde{N}[k]|^2 - 10log(bw) + 10log(cb(f))] + 4 \tag{3.8}$$

where $|\tilde{X}[k]|$ and $|\tilde{N}[k]|$ are the magnitudes of the DFT of the windowed tone and noise, respectively, and bw is the bandwidth of the noise masker. Masking is expected to occur when $SNRB = 0$ dB.

Assuming that masking of formants in vowels is the same as tone masking (an assumption to be verified through experiments which are described in a later section) the level of the noise needed to just mask each formant peak can be calculated. All harmonics within a critical band are assumed to contribute to the overall level of a formant peak. The critical band is assumed to be rectangular.

For example, consider an F2 in a steady-state vowel at 1100 Hz, where the critical band is 190 Hz (Reed and Bilger, 1973). If the fundamental frequency (f0) is at 100 Hz then there will be only one harmonic (the 11th harmonic) in the 190 Hz critical band centered at 1100 Hz. Hence, the level of the 11th harmonic is considered to be a good estimate of the level of F2 for purposes of estimating the masking of the formant peak by broadband noise. If, on the other hand, f0 is at 90 Hz, then two harmonics (the 12th and 13th) would fall into the 190 Hz centered at 1100 Hz; the two harmonics are at 1080 and 1170 Hz. Hence, the overall level of F2 ($F2L$) in dB is:

$$F2L = 10log(10^{x_{12}/10} + 10^{x_{13}/10}) \tag{3.9}$$

where x_i is the level in dB of the *ith* harmonic.

For example, if the level of F2 is estimated to be 30 dB, then we can estimate the level of a white noise with bandwidth 4500 Hz needed to just mask F2. Using equation 3.4, we find that the overall noise level necessary to mask F2 is 48 dB.

The procedures outlined in this section were used to estimate the *SNRB* for a tone or a formant frequency in a steady-state vowel. Consequently, predictions of tone/formant masking were made based on the frequency and level of the tone/formant, the level of the noise masker, and critical band values.

3.1.2 Above-band Masking

As noted in Section 1.1.1, above-band masking refers to masking of tones at frequencies higher than the cutoff frequency of the masker. The reason for our interest in this kind of masking is to determine the level of a band-pass noise masker needed to mask a tone or a formant peak that is above the masker band.

The non-linear growth of above-band masking with masker level is referred to in the literature as 'upward-spread of masking'. A standardized method for predicting upward spread of masking was described as part of a procedure for predicting speech intelligibility in noise (ANSI-S3.5, 1969). The upward spread of masking algorithm, which was based on the experimental results of Bilger and Hirsh (1956) and of Carter and Kryter (1962), predicts the slope of the 'effective' noise spectrum, which extends above the physical spectrum, if the spectrum level and cut-off frequency of the masker are known. In addition, the algorithm specifies a minimum spectrum level, which is 46 dB SPL (re .0002 μbar), for spread of masking to occur. The masked threshold of a tone can then be calculated by adding the effective spectrum level of the masker at that frequency to the critical ratio of the tone.

Another procedure for predicting upward spread of masking was proposed by Ludvigsen (1985). Ludvigsen derived an equation based on the experimental results with band-pass noise maskers published by Zwicker (1963). The equation is as follows[2]:

$$T(f) = N_0 + CR(f) + \alpha_n log_2(\frac{f}{f_m}) \tag{3.10}$$

where

$$\alpha_n = -80 + 0.6[N_0 + 10log(0.231f_m)] \tag{3.11}$$

and f_m is the cutoff frequency of the masker in Hz. $T(f)$, N_0, and $CR(f)$ are the tone masked threshold, noise spectrum level, and tone critical ratio, respectively, as defined in the previous section.

Ludvigsen compared predicted and measured thresholds for hearing-impaired subjects in the presence of a low-pass masker and concluded that the model was a good first-order approximation of spread of masking. Ludvigsen's equation are assumed to be valid at any spectrum level; this is in contrast with the ANSI algorithm which specifies a minimum noise spectrum level for spread of masking to occur.

The two methods (ANSI and Ludvigsen's equation) have not been previously compared for their accuracy in predicting masked thresholds.

In this chapter, both procedures are used and compared to predict masked thresholds in the above-band masking case.

[2] For simplicity, we have eliminated a term from the original equation which is only necessary when simulating hearing loss.

3.2 Tone and Formant Detection Experiments: Methodology

In this section the experimental methodology of tone-in-noise and vowel-in-noise experiments is presented. The goal of the tone experiments was to measure the masked thresholds of tones in both white noise (for measurement of the subjects' critical bands) and band-pass noise. The goal of the vowel experiments was to find out whether the results of masking theory established primarily for pure tones could be applied to formant frequencies of vowels in the presence of either a white noise or a band-pass noise masker.

3.2.1 Paradigms

Procedures common to tone-in-noise and vowel-in-noise experiments

Two-interval 2AFC (two-alternative forced choice) experiments with no feedback were used. In this type of experiment, subjects were presented with two observation intervals, asked a question, and then forced to choose the interval (1 or 2) thought to contain the information asked for. The information was equally likely to occur in either interval. The time between observation intervals was chosen to be 500 ms because a time less than 500 ms or greater than 2 sec is known to cause performance degradation in a two-alternative test (Tanner, 1961). The time between trials was 3 sec and there was a 4.5 sec pause after each block of ten trials.

The signal duration was 320 ms and the noise masker was 520 ms. Both signal and masker were generated digitally and were gated on and off with a 10-ms ramp. The signal was centered in the middle of the noise waveform. The reason for the onset delay between the signal and the masker was to allow for adaptation to the characteristics of the noise waveform and, hence, avoid the overshoot effect which has been observed for short onset delay times (Zwicker and Fastl, 1990). Four frozen noise waveforms (drawn from the same stationary noise process) were used in the experiments. The signal-to-noise ratio was varied by attenuating the signal level while keeping the noise level fixed. All signal processing necessary for stimulus preparation (noise generation, signal attenuation, mixing, etc.) was done with the aid of a block signal-processing scheme implemented using the *MITSYN* language (Henke, 1989).

Stimuli were repeated, randomized, converted into analog signals using a 12-bit D/A converter and recorded onto TDK-AD60, Type I cassette tapes. There were 20 repetitions for each signal level. The stimuli were checked to ensure that no clipping or distortion occurred. The subjects listened binaurally to the stimuli over Sennheiser HD-430 headphones in a sound-treated room. Headphones were calibrated using a probe-microphone calibration technique (Villchur and Killion, 1975). A Ballantine voltmeter was used to measure the voltage supplied to the headphones. The calibration curves were then used to convert the voltage to dB SPL. For the 'within-band' experiments, subjects were free at the onset of each experiment to adjust the volume on the tape deck to a comfortable level as long as the adjustment would not result in sound levels exceeding 90 dB SPL. Most subjects adjusted the volume such that the overall noise level in each experiment was between 70 and 80 dB SPL. Training periods, lasting between 1/2 h to 1 h depending on the subject, preceded each listening session. Listening sessions lasted 30-60 minutes. After 30 minutes of listening, there was a rest period of a few minutes. For each subject, experiments were conducted twice on two different days. Threshold estimates were taken to be the 76% correct level (Green and Swets, 1966).

Tone-in-noise experiments

In these experiments, one of the observation intervals contained noise only and the other contained noise plus a tone. The observers were forced to choose the interval thought to contain the tone.

Vowel-in-noise experiments

The signals in these experiments were either one-formant or 'multi-formant' synthetic vowels. The vowels were synthesized using the parallel part of the formant synthesizer KLSYN88. The one-formant vowels were presented in one of the intervals of the 2AFC tests and the subjects were forced to choose the interval thought to contain the vowel (same as the tone-in-noise experiments). For the 'multi-formant' vowel experiments, subjects were presented with two pairs of vowels mixed with noise: one pair had identical vowels (/V/, /V/), and the second vowel in the other pair had a missing second formant frequency (/V/, /VnoF2/). The same noise waveform was used for both pairs of vowels on each trial and the time between members of the same pair was 500 ms. The 'different' pair was equally likely to be in each position. The observers were forced to choose the interval thought to contain the different pair of vowels.

3.2.2 Signals

Two sets of signals were used in the experiments: the /ɑ/ set and the /ɛ/ set. The /ɑ/ set consisted of either a tone at 1100 Hz, a synthetic one-formant vowel at 1100 Hz, or two synthetic /ɑ/-like vowels. One of the /ɑ/-like vowels had no second formant frequency (which would otherwise be at 1100 Hz). The /ɛ/ set consisted of either a tone at 1800 Hz, a synthetic one-formant vowel at 1800 Hz, or two synthetic /ɛ/-like vowels. One of the /ɛ/-like vowels had no second formant frequency (which would otherwise be at 1800 Hz).

Tones were generated digitally and the vowels were synthesized using the parallel part of the formant synthesizer KLSYN88[3]; the default voicing source model (KLGLOTT88) was used. The synthesis of 'pathological' vowels (for example, a one-

[3] For a brief description of the synthesizer, the reader is referred to Section 2.2.

formant vowel or an /ɑ/ vowel with no F2) required the individual control of each formant amplitude, and hence the parallel part of the synthesizer was used. Both the tones and the synthetic vowels were of duration 320 ms and were gated on and off by a 10-ms ramp.

One-formant vowels

Two one-formant vowels (V11 and V18) were synthesized. Figure 3.1 shows DFT spectra sampled at the midpoint of the two vowels. The formant frequency for the V11 vowel was at 1100 Hz and for V18, it was at 1800 Hz. Both vowels were synthesized with a constant fundamental frequency of 100 Hz and a formant bandwidth of 60 Hz.

The /ɑ/ vowels

Two synthetic /ɑ/-like vowels were used in the experiments. One of the vowels was synthesized with the first four formant frequencies located at: 625, 1100, 2700, 3250 Hz, respectively. The bandwidths of the formant frequencies were set at the default values of the synthesizer: $B1 = 60$ Hz, $B2 = 90$ Hz, $B3 = 150$ Hz, and $B4 = 200$ Hz. The relative amplitude of the formant peak at F2 and F3 in comparison with that of F1 was $-$ 6 and $-$ 20 dB, respectively.

The second vowel (hereafter referred to as 'ɑnoF2') was identical to the /ɑ/ vowel except that it had no second formant frequency. Figure 3.2 (a,b) shows DFT spectra sampled at the mid-point of the two vowels (/ɑ/, and /ɑnoF2/), and Figure 3.2c shows an overlay of the smoothed spectra of the two vowels.

In order to keep the amplitudes of all formant frequencies, except F2, the same between the two vowels (Figure 3.2c), the amplitude of each formant frequency was manipulated to match the target spectrum levels using the parallel part of the formant

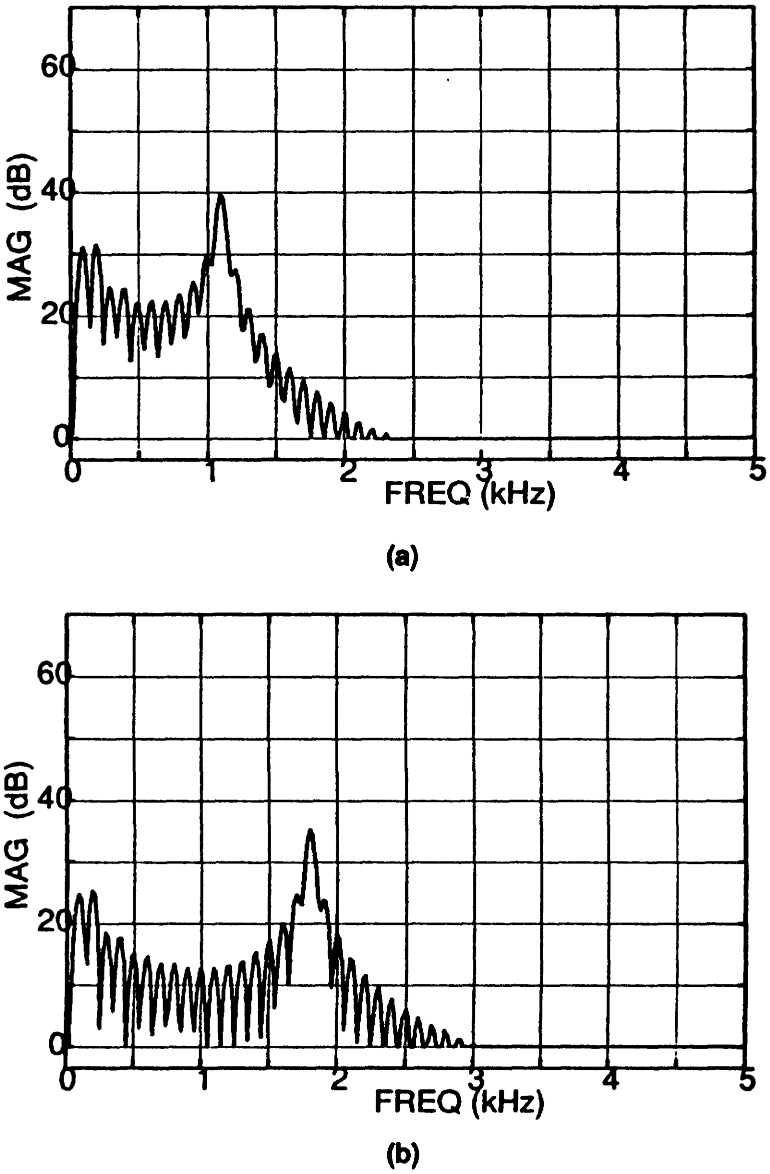
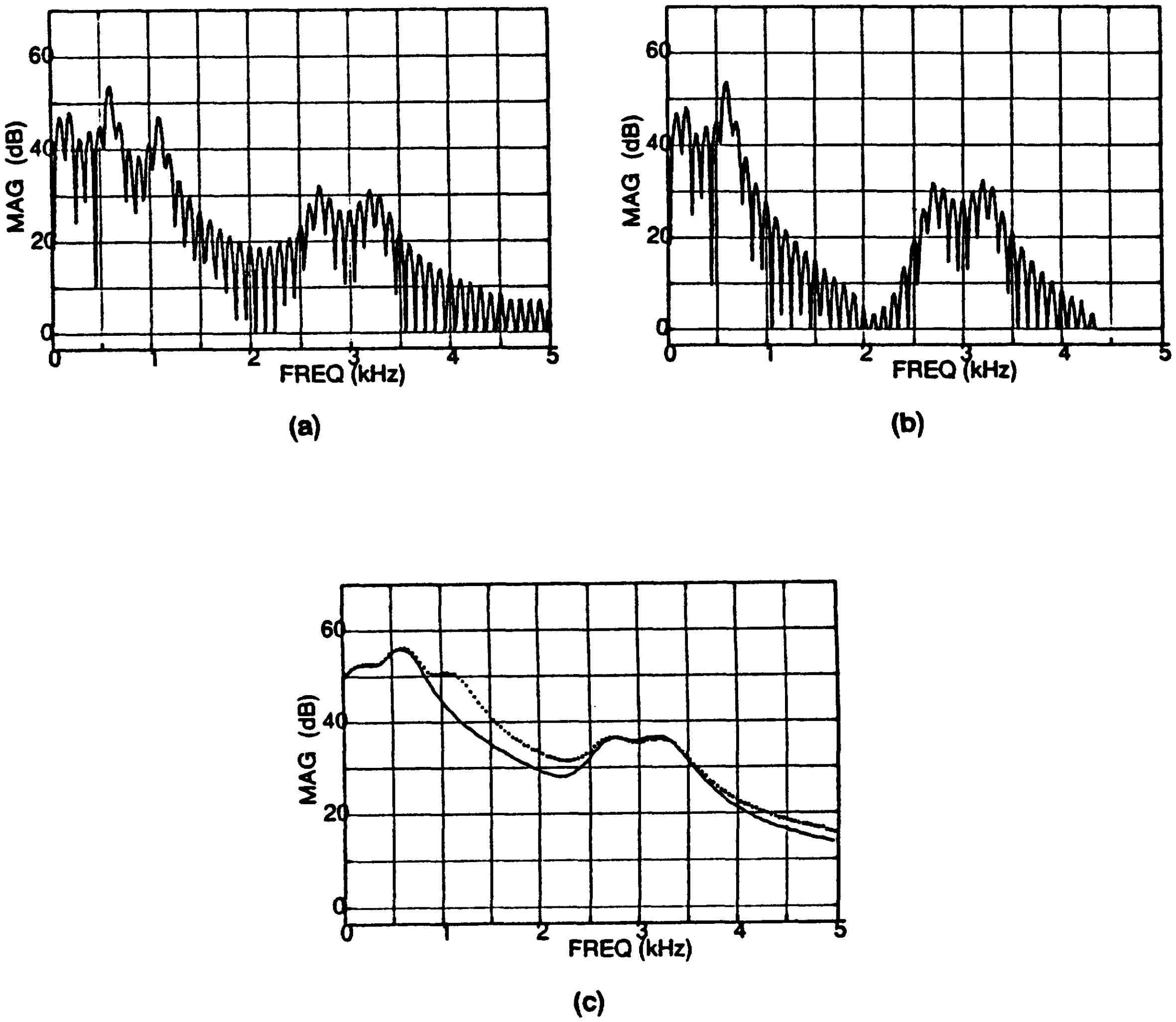


Figure 3.1: Discrete Fourier transform spectra sampled at the mid-point of two steady-state synthetic vowels with one formant frequency only: (a) formant at 1100 Hz (V11), and (b) formant at 1800 Hz (V18). Spectra were computed using a Hamming window of duration 25.6 ms. The signals were not preemphasized.



(a)

(b)

(c)

Figure 3.2: Discrete Fourier transform spectra sampled at the mid-point of two synthetic vowels: (a) a full-formant /ɑ/ vowel and (b) a missing-F2 vowel (/ɑnoF2/). An overlay of the smoothed spectra of the two vowels is shown in (c). The dashed line spectrum is the smoothed spectrum for /ɑ/ and the solid line, for /ɑnoF2/. Smoothing was done with a 300 Hz filter. All spectra were computed using a Hamming window of duration 25.6 ms. The signals were not preemphasized.

synthesizer KLSYN88. The fundamental frequency for the vowels was set at 100 Hz and was constant throughout the vowel.

The /ɛ/ vowels

Two /ɛ/-like vowels were synthesized using the parallel part of the synthesizer KLSYN88. One of the vowels was a synthetic /ɛ/ vowel with the first four formant frequencies located at: 500, 1800, 2700, 3500 Hz, respectively. The bandwidths of the formant frequencies were set at the default values of the synthesizer. The location and relative amplitudes of the formant frequencies were chosen to be similar to those obtained from the analysis of natural utterances. The relative amplitude of both formant peaks at F2 and F3 in comparison with that at F1 was – 16 dB.

The second vowel (hereafter referred to as 'ɛnoF2') was identical to the /ɛ/ vowel except that it had no second formant frequency. Figure 3.3 (a,b) shows DFT spectra sampled at the mid-point of the two vowels (/ɛ/, and /ɛnoF2/), and Figure 3.3c shows an overlay of the smoothed spectra of the two vowels. The fundamental frequency for the /ɛ/ vowels was set at 125 Hz and was constant throughout the vowel.

3.2.3 Noise Maskers

Frozen noise waveforms were used as maskers in the perceptual experiments. Figure 3.4 shows a block diagram of the signal-processing scheme used to generate the noise signals. White-noise waveforms were generated by passing a random sample through a digital low-pass filter with a cut-off frequency of .47 times the sampling frequency (4700 Hz for a sampling rate of 10000 Hz) (Figure 3.4a). Band-pass noise signals were generated by digitally filtering white-noise signals (Figure 3.4b). Butterworth digital filters were implemented based on the McClellan et al. (1973) algorithm.

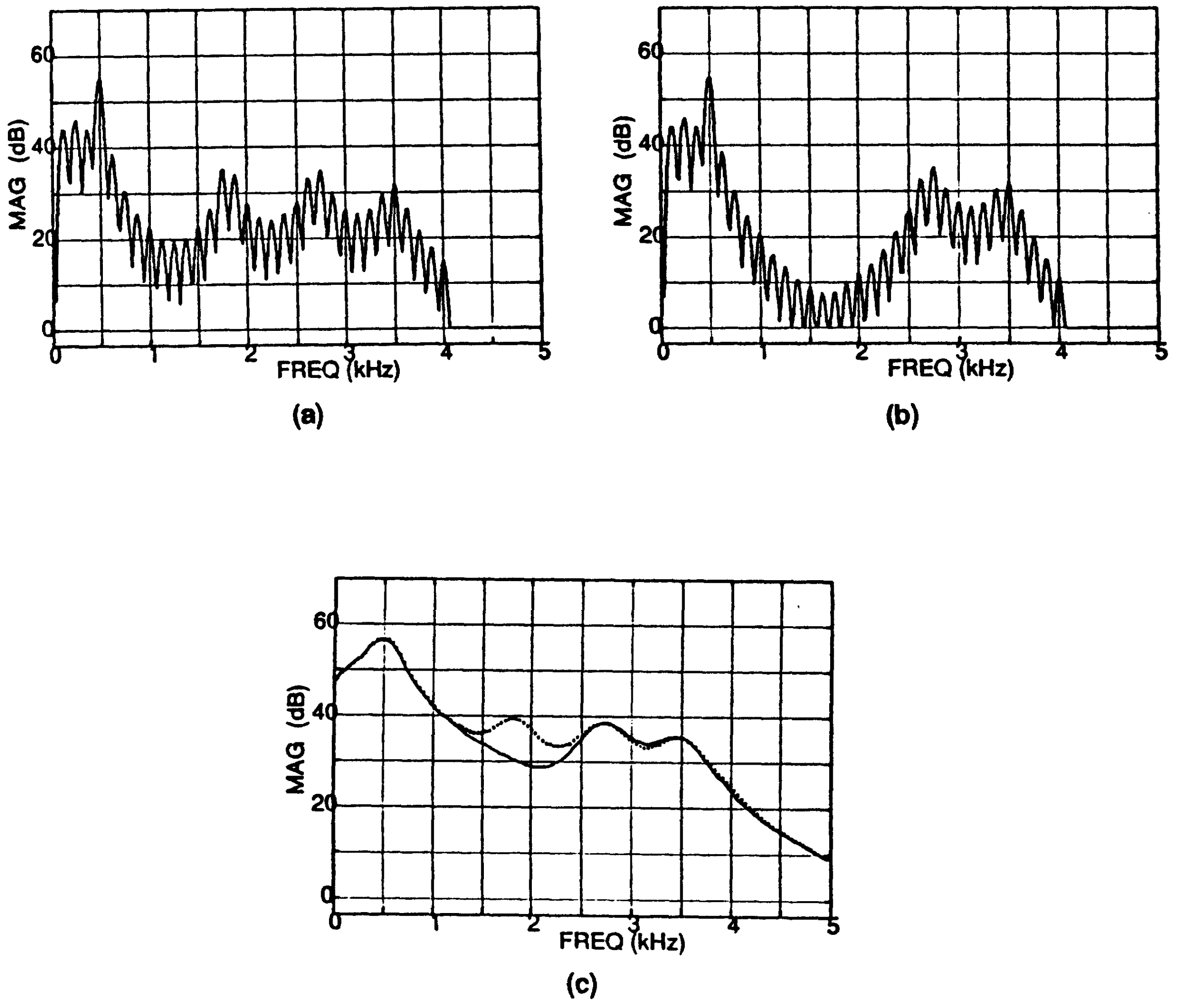


Figure 3.3: Discrete Fourier transform spectra sampled at the mid-point of two synthetic vowels: (a) a full-formant /ε/ vowel and (b) a missing-F2 vowel (/εnoF2/). An overlay of the smoothed spectra of the two vowels is shown in (c). The dashed line spectrum is the smoothed spectrum for /ε/ and the solid line, for /εnoF2/. Smoothing was done with a 300 Hz filter. All spectra were computed using a Hamming window of duration 25.6 ms. The signals were not preemphasized.

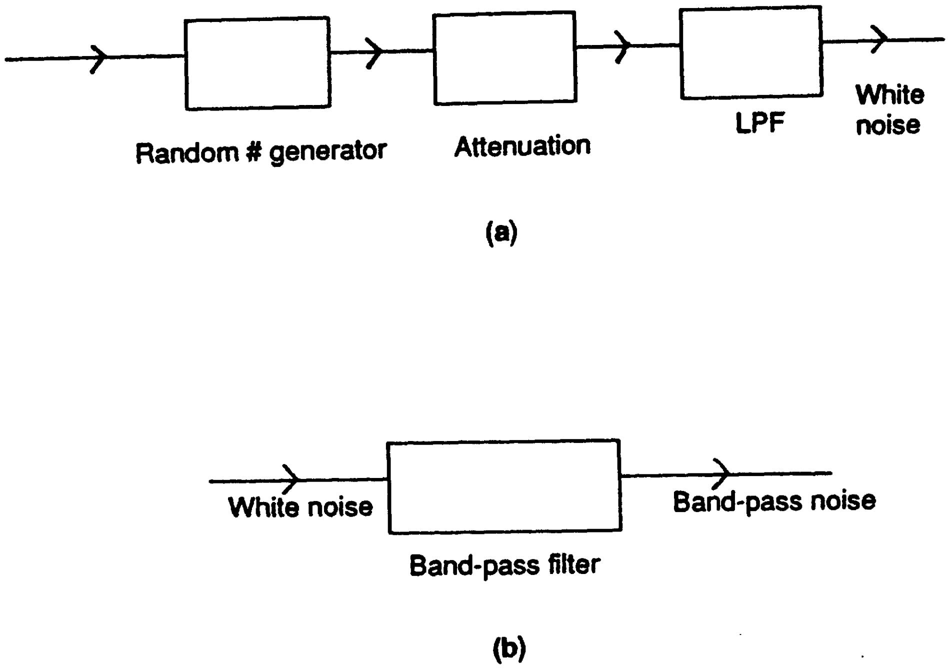
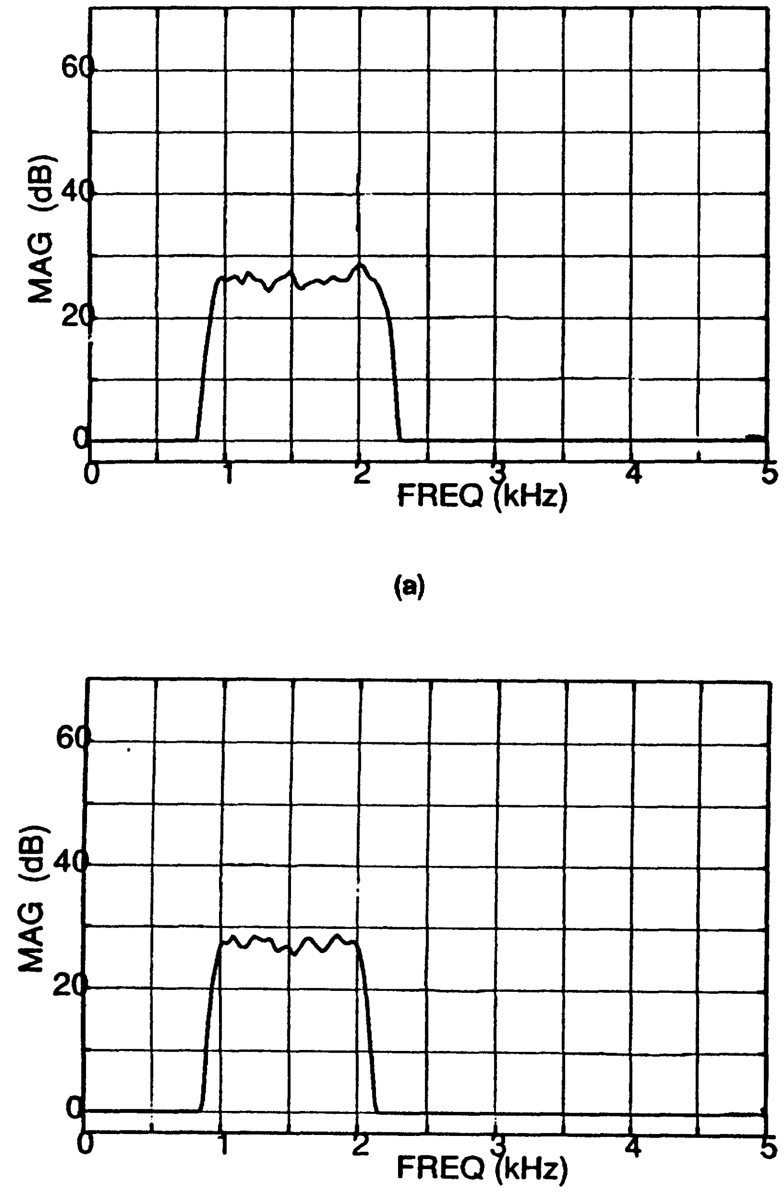


Figure 3.4: Block diagram of the signal-processing scheme used to generate (a) white noise and (b) band-pass noise. For white-noise generation, a random sample was first generated, attenuated, and low-pass filtered (digitally). Band-pass noise was generated by filtering a white-noise signal with the appropriate digital filter.

Figure 3.5 shows spectra of the two band-pass noise maskers (n12, n10) used in the 'band-pass' experiments. The spectra were computed by averaging 12 overlapping 25.6 ms segments. Both band-pass maskers were centered at 1550 Hz but differed in bandwidth: 'n12' was 1200 Hz wide, and 'n10' was 1000 Hz wide. The masker 'n12' was used for the within-band experiments whereas both maskers 'n12' and 'n10' were used in the above-band masking experiments. The sharp skirts of the band-pass noise rolled off at 100 dB/ 1/3 octave. The noise waveforms were 520 ms in duration and were gated on and off by a 10-ms ramp.

3.2.4 Subjects

Five subjects participated in the /ɛ/ experiments (CB, JK, JM, JW, TR), and the same subjects, with the exception of subject (TR), participated in the /ɑ/ experiments. Subject (TR) was unavailable for the one-formant vowel experiments. Subject (MM) whose initial appears on some of the graphs participated in a few of the experiments. All subjects were native speakers of English ranging in age from 18-44 years old. Only subjects JK, JM, and MM were paid for their services. None had any known speech or hearing problems.



(a)

(b)

Figure 3.5: Discrete Fourier transform average spectra of two band-pass noise maskers centered at 1550 Hz and were: (a) 1200 Hz wide (n12) and (b) 1000 Hz wide (n10). Both spectra were computed by averaging 12 overlapping 25.6 ms segments. No preemphasis was used.

3.3 Experimental Results: Within-Band Masking

Representation of the results in the figures in this section was in terms of percent correct (P(c)) versus signal-to-noise ratio in a critical band ($SNRB$) centered around a tone (for the tone-in-noise experiments), or around F2 of the steady-state vowel (for the vowel experiments). The function that represents the relationship between P(c) and $SNRB$ is referred to as a *psychometric* function. Each data point in the psychometric functions were based on forty judgements by each subject. Estimation of the $SNRB$ was based on critical-band values published by Reed and Bilger (1973) and were 190 and 280 Hz for tones at 1100 and 1800 Hz, respectively.

3.3.1 White-noise Masker

Tones

In these experiments the masked thresholds of two tones (1100 Hz and 1800 Hz) in white noise were measured. The procedure described in Section 3.2.1 was used in these experiments. The overall level of the noise was between 70-80 dB SPL, depending on the subject.

The averaged results of these experiments are shown in Figures 3.6a and 3.7a, and individual psychometric functions are shown in Figures 3.6b and 3.7b for the 1100 and 1800 Hz tones, respectively. On average, the masked threshold, which is the 76% correct point, corresponded to an $SNRB$ of 0 dB. Subject CB seemed to have the widest critical band of all of the subjects and subject JM, the narrowest.

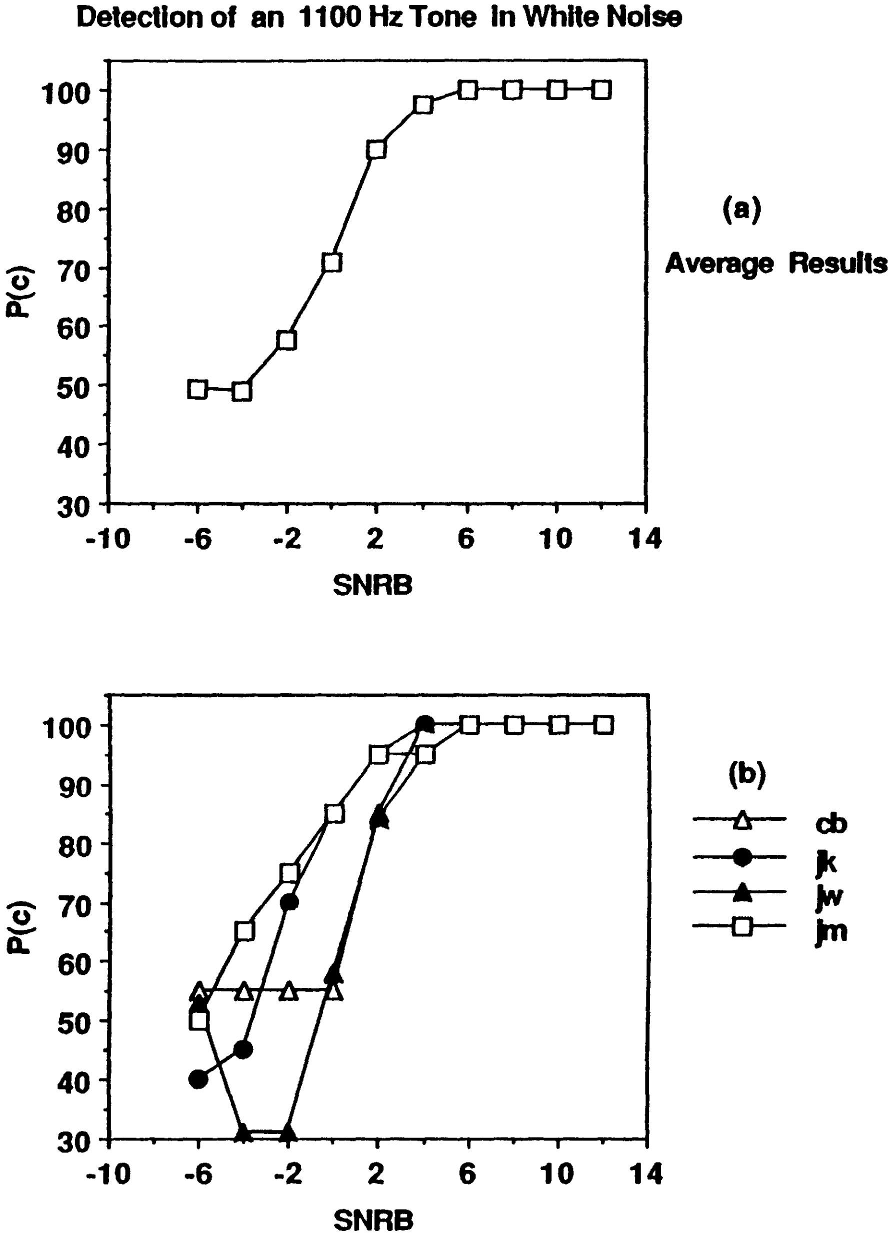


Figure 3.6: Psychometric functions of a tone-in-noise experiment. The tone was at 1100 Hz and the masker was white noise. Average results are shown in (a) and individual functions are shown in (b).

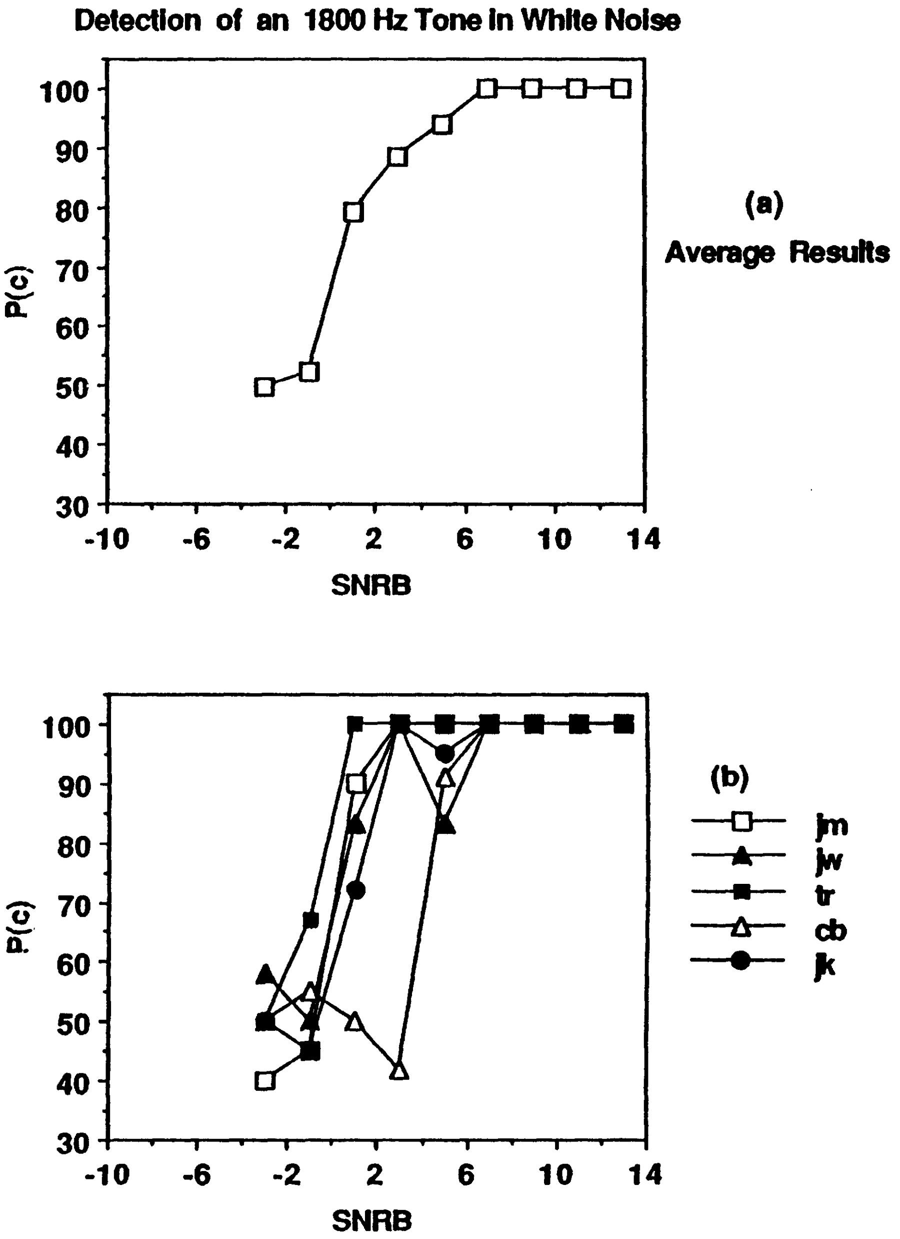


Figure 3.7: Psychometric functions of a tone-in-noise experiment. The tone was at 1800 Hz and the masker was white noise. Average results are shown in (a) and individual functions are shown in (b).

One-formant vowels

As described in Section 3.2, two one-formant synthetic vowels (V11 and V18) with formant frequencies at 1100 Hz and 1800 Hz, respectively, were mixed with noise and presented to subjects in one of the intervals of a 2AFC test. The subjects then chose the interval thought to contain the signal. In pilot experiments with the one-formant vowels, subjects reported hearing the vowel-like sound as a tone at low $SNRB$ and as a 'buzzer' at high $SNRB$. Hence, the instructions for these experiments did not specify what the signal was but rather asked the subject to report the interval within which the signal was thought to reside.

Average results are shown in Figure 3.8.a and 3.9.a and individual psychometric functions are shown in Figures 3.8.b and 3.9.b for the vowels V11 and V18, respectively. On average, the 76% correct point for the vowel (V11) occurred at the same $SNRB$ as that of the tone at 1100 Hz, whereas the threshold for the tone at 1800 Hz was 2 dB higher than that of the vowel (V18). That is, the listeners detected the vowel with a formant at 1100 Hz in a way which is similar to that of detecting a tone at the same frequency, whereas they did 'better' in detecting the vowel with a formant at 1800 Hz than they did with a tone at that frequency. In addition, the psychometric functions for the vowel experiments were shallower than those from the tone experiments.

Multi-formant vowels

As described in Section 3.2.1, the two /ɑ/-like vowels were mixed with white noise and presented to the listeners in pairs ([/ɑ/,/ɑ/; /ɑ/,/ɑnoF2/] or [/ɑ/,/ɑnoF2/; /ɑ/,/ɑ/]). The listeners reported which pair (first or second) had different vowels. The noise masker was presented at overall levels between 72 and 84 dB SPL, depending on the subject. Similarly, the /ɛ/-like vowels were presented to subjects in pairs (with and without F2). The noise masker was presented at overall levels of 70-85 dB SPL. The hy-

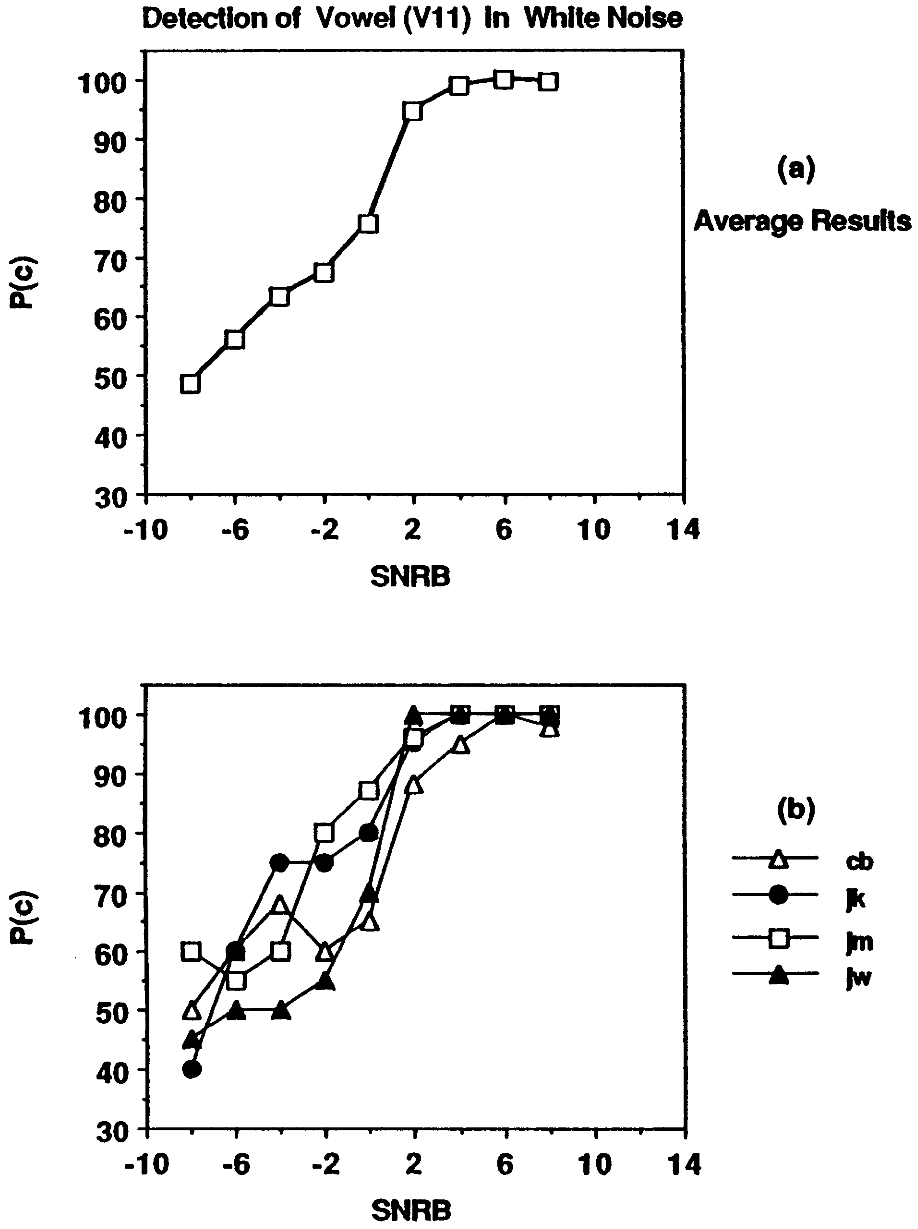


Figure 3.8: Psychometric functions of a vowel-in-noise experiment. The vowel (V11) had one formant frequency at 1100 Hz and the masker was white noise. Average results are shown in (a) and individual functions are shown in (b).

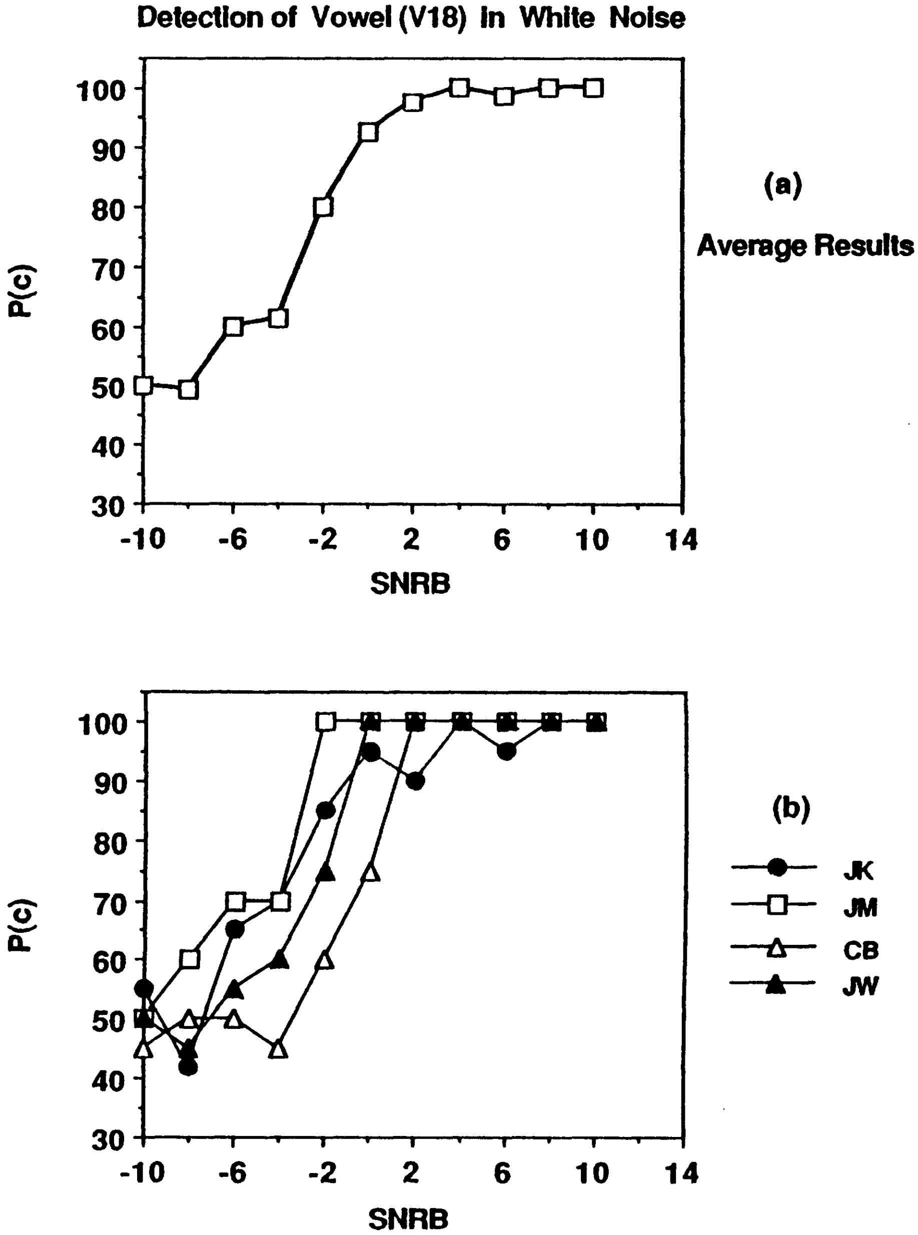


Figure 3.9: Psychometric functions of a vowel-in-noise experiment. The vowel (V18) had one formant frequency at 1800 Hz and the masker was white noise. Average results are shown in (a) and individual functions are shown in (b).

pothesis here was that if masking theory holds for formant frequencies in vowels, then the vowels /V/ and /V noF2/ would sound the same when F2 in the 'full-formant' vowel /V/ is predicted to be masked.

Results of the vowel experiments are shown in Figures 3.10 and 3.11. Figures 3.10a and 3.11a show average results, and Figures 3.10b and 3.11b show individual results, for the /ɑ/-like vowel and the /ɛ/-like vowel, respectively. The results show a consistent trend among listeners: when the $SNRB$ in a critical band around F2 is 0 or less, then the vowel pairs /ɑ/, /ɑnoF2/ and /ɛ/, /ɛnoF2/ sound similar. On average, the 76% correct point corresponded to an $SNRB$ of -1 dB for /ɑ/ (slightly better than tone detection) and 2 dB for /ɛ/ (slightly worse than tone detection). Subjects reported that the discrimination task in these experiments was more difficult than the detection task in the tone and one-formant vowel experiments.

3.3.2 Band-pass Noise Masker

The noise masker for the band-pass experiments was 1200 Hz wide (n12) and centered around 1550 Hz. A spectrum of the noise masker was shown in Figure 3.5a. The signals used in these experiments included tones and multi-formant vowels. Equation 3.8, which was used to estimate the $SNRB$, assumed that the background noise was broadband. This assumption was reasonable for the band-pass masker used in these experiments since the masker width was wider than the critical band of the tones and formants in question.

Tone-in-noise experiments

In these experiments, the masked thresholds of two tones at 1100 Hz and at 1800 Hz were measured in the presence of a band-pass masker. The experiments were repeated

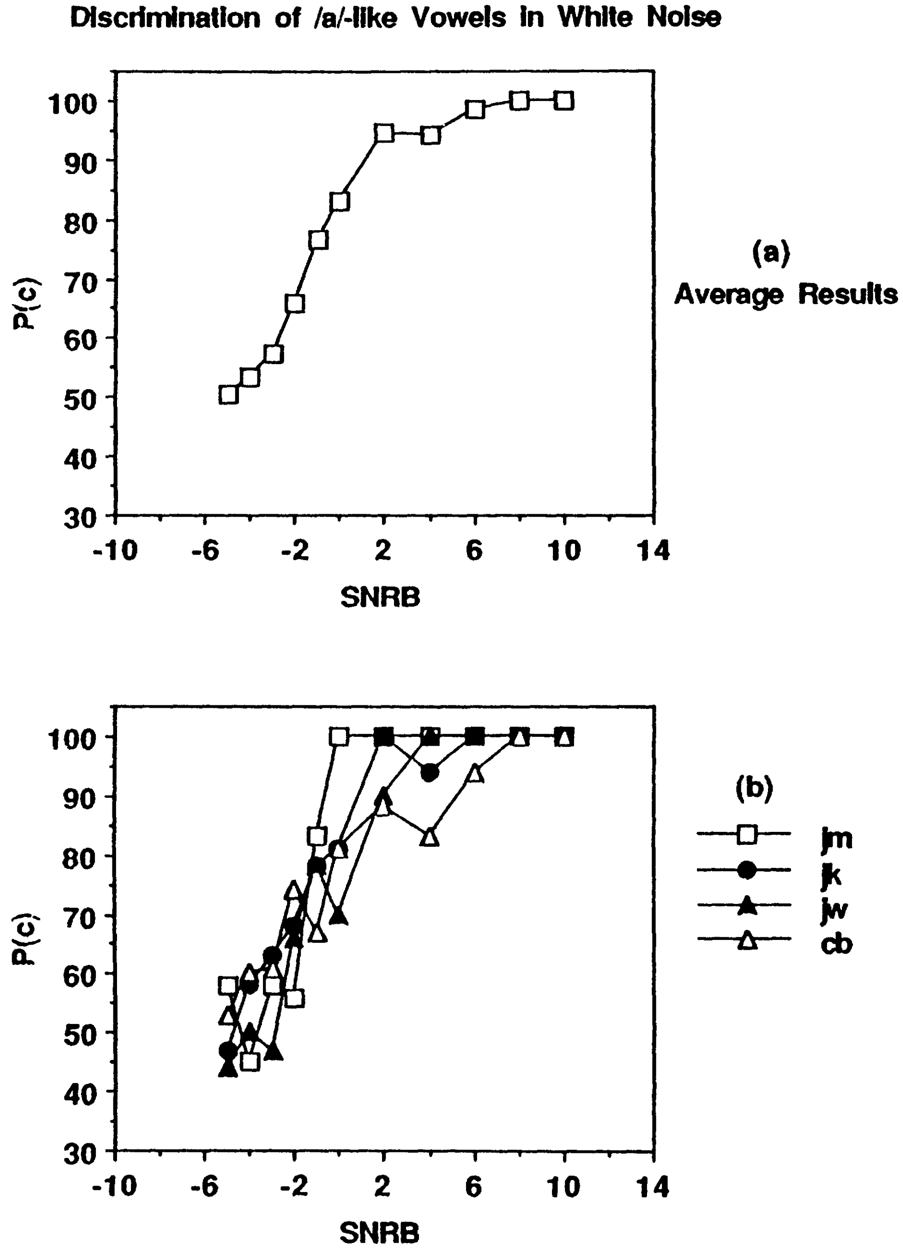


Figure 3.10: Psychometric functions of a vowel-in-noise experiment. Two /ɑ/-like vowels were used in this experiment: one was synthesized with all appropriate formants (/ɑ/) and the other was synthesized with a missing F2 (/ɑnoF2/). The masker was white noise. P(c) here reflects the listener's ability to discriminate correctly between the two vowels. Average results are shown in (a) and individual functions are shown in (b).

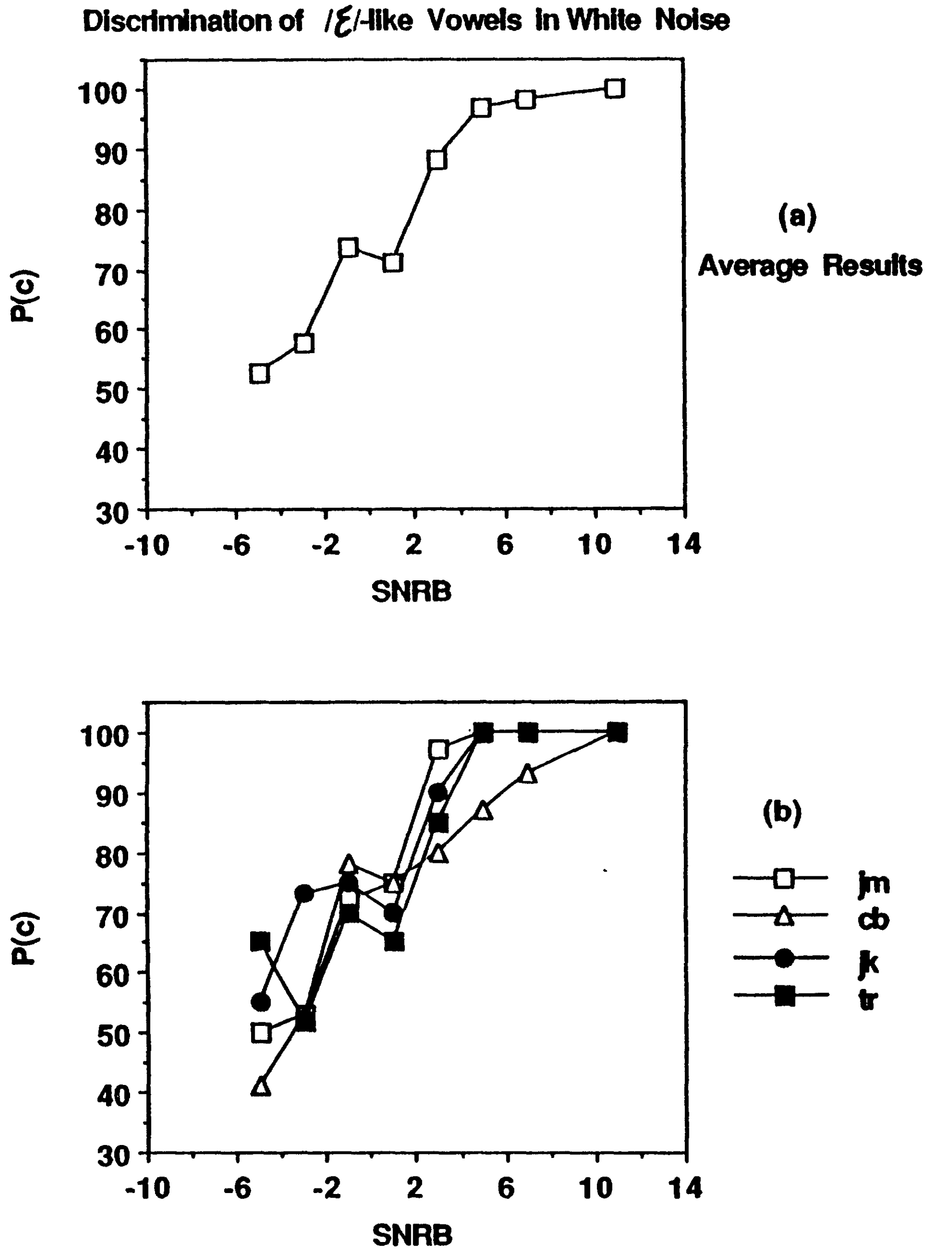


Figure 3.11: Psychometric functions of a vowel-in-noise experiment. Two /ɛ/-like vowels were used in this experiment: one was synthesized with all appropriate formants (/ɛ/) and the other was synthesized with a missing F2 (/ɛnoF2/). The masker was white noise. P(c) here reflects the listener's ability to discriminate correctly between the two vowels. Average results are shown in (a) and individual functions are shown in (b).

twice for each subject on different days. The overall noise level presentations were between 70-80 dB SPL, depending on the subject. The subjects reported that the band-pass experiments were more difficult than experiments with white noise because the band-pass noise had a 'pitch', the perception of which interfered with the tone detection task.

Average results are shown in Figures 3.12a and 3.13a and individual results are shown in Figures 3.12b and 3.13b. The data show a larger variability than the white-noise case (Figure 3.6 and 3.7) and the curves are less steep. However, on average, the masked threshold corresponded to an $SNRB$ of +1 dB for the 1100 Hz case and +2 dB for the 1800 Hz case. The thresholds in this case were slightly higher than those in the presence of a white-noise masker. Subject CB again exhibited wider critical bands than the other subjects.

Vowel-in-noise experiments

The procedure outlined in Section 3.2.1 was used in these experiments with a band-pass noise masker. In one experiment, the synthetic vowels /ɑ/, /ɑnoF2/ were used , and in the other, /ɛ/, /ɛnoF2/. The overall presentation level of the noise was 65-75 dB SPL, depending on the subject.

Results of the /ɑ/ experiments are shown in Figure 3.14 and results of the /ɛ/ experiment are summarized in Figure 3.15. As in the tone-in-noise experiments, there is a larger variability among listeners' responses in comparison to the white-noise case. However, on average, if the $SNRB$ was 0 dB or less then masking of F2 occurs, resulting in perceptual similarity between the full-formant vowel and the 'no-F2' vowel.

Subject JK was able to detect the difference between the full-formant /ɛ/ vowel and the missing F2 /ɛ/ vowel (/ɛnoF2/) even for an $SNRB$ which is well below 0 dB.

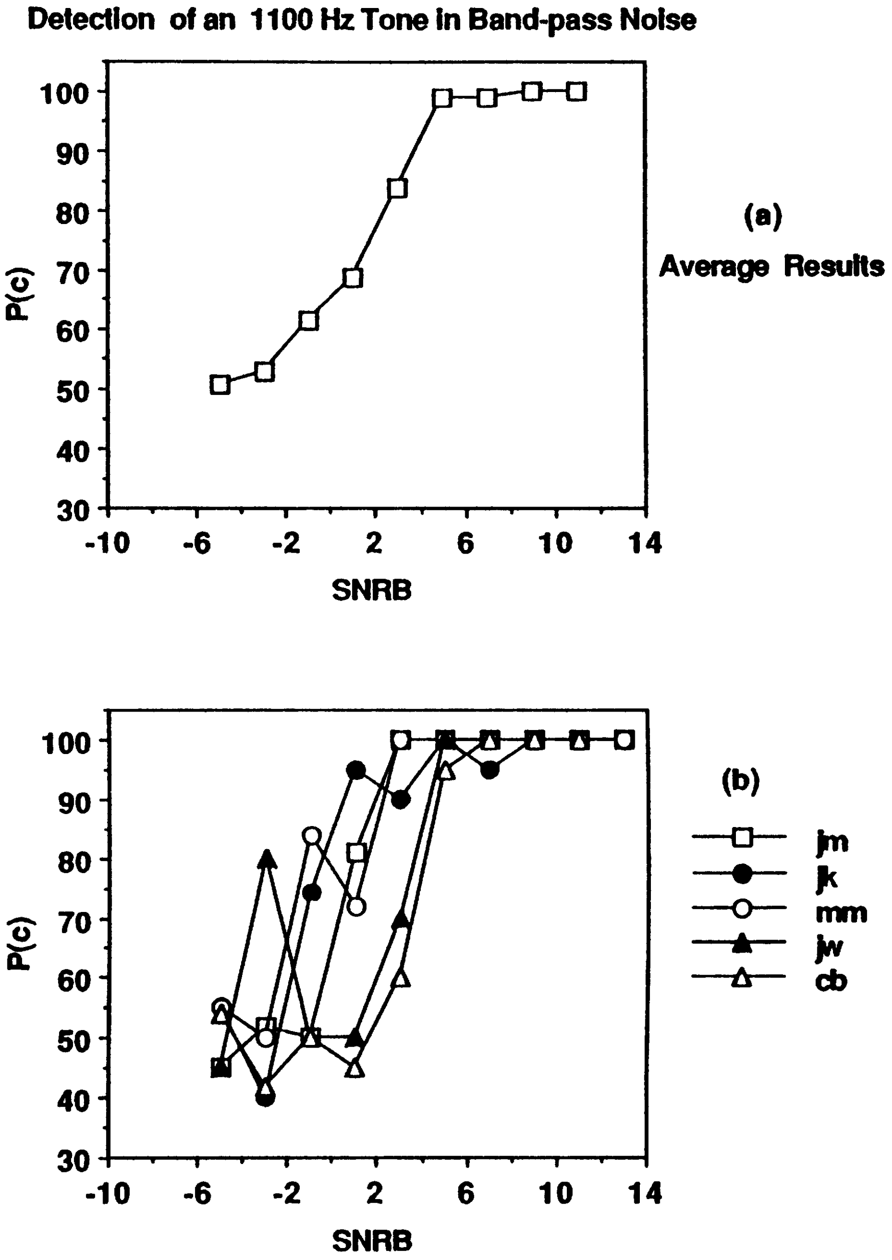


Figure 3.12: Psychometric functions of a tone-in-noise experiment. The tone was at 1100 Hz and the masker was a band-pass noise masker centered at 1550 Hz and was 1200 Hz in bandwidth. Average results are shown in (a) and individual psychometric functions are shown in (b).

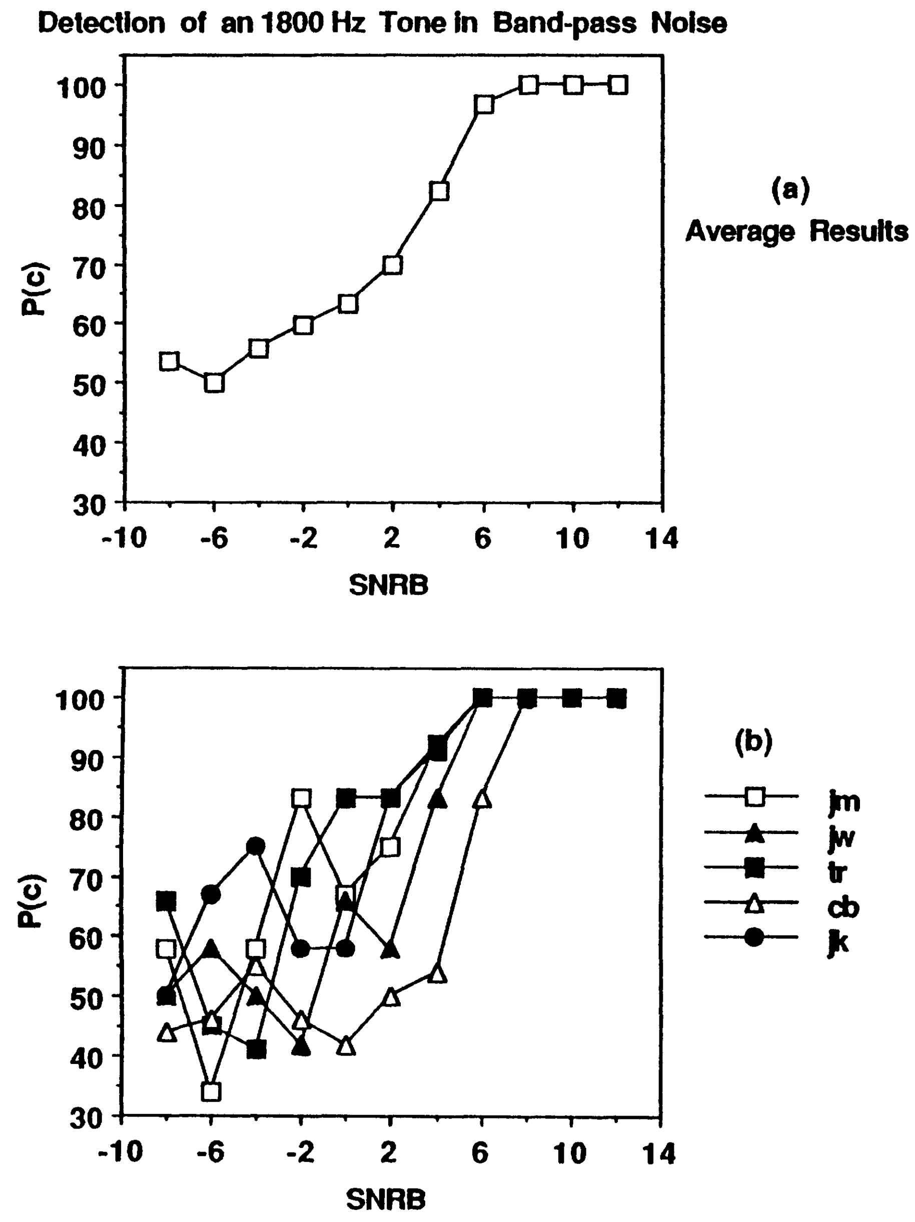


Figure 3.13: Psychometric functions of a tone-in-noise experiment. The tone was at 1800 Hz and the masker was a band-pass noise masker centered at 1550 Hz and was 1200 Hz in bandwidth. Average results are shown in (a) and individual psychometric functions are shown in (b).

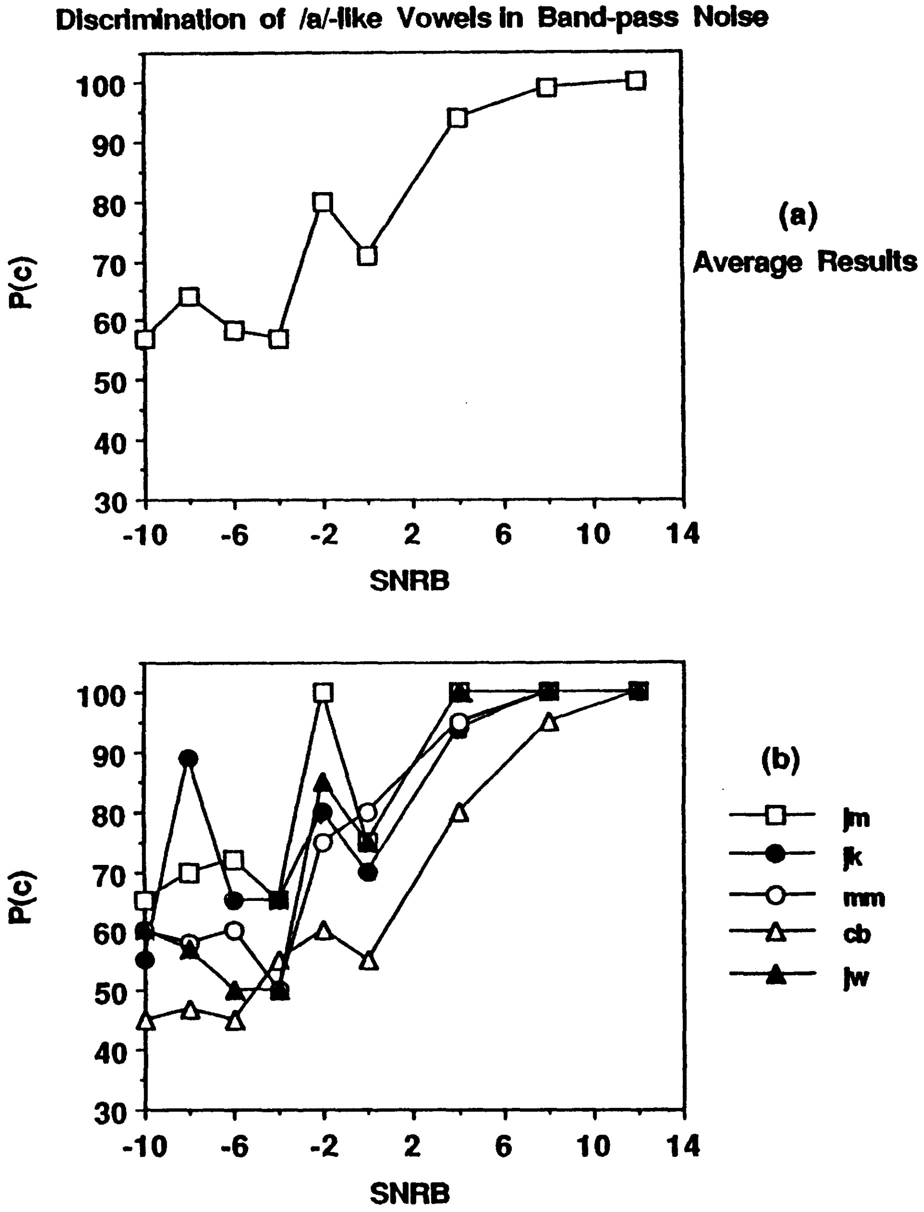


Figure 3.14: Psychometric functions of a vowel-in-noise experiment. Percent correct here reflects the ability of the subject to discriminate correctly between the two synthetic /a/ vowels which differed only in the presence of F2 (/a/) or absence of F2 (/anoF2/). The masker was band-pass noise centered at 1550 Hz and was 1200 Hz in bandwidth. Average results are shown in (a) and individual results are shown in (b).

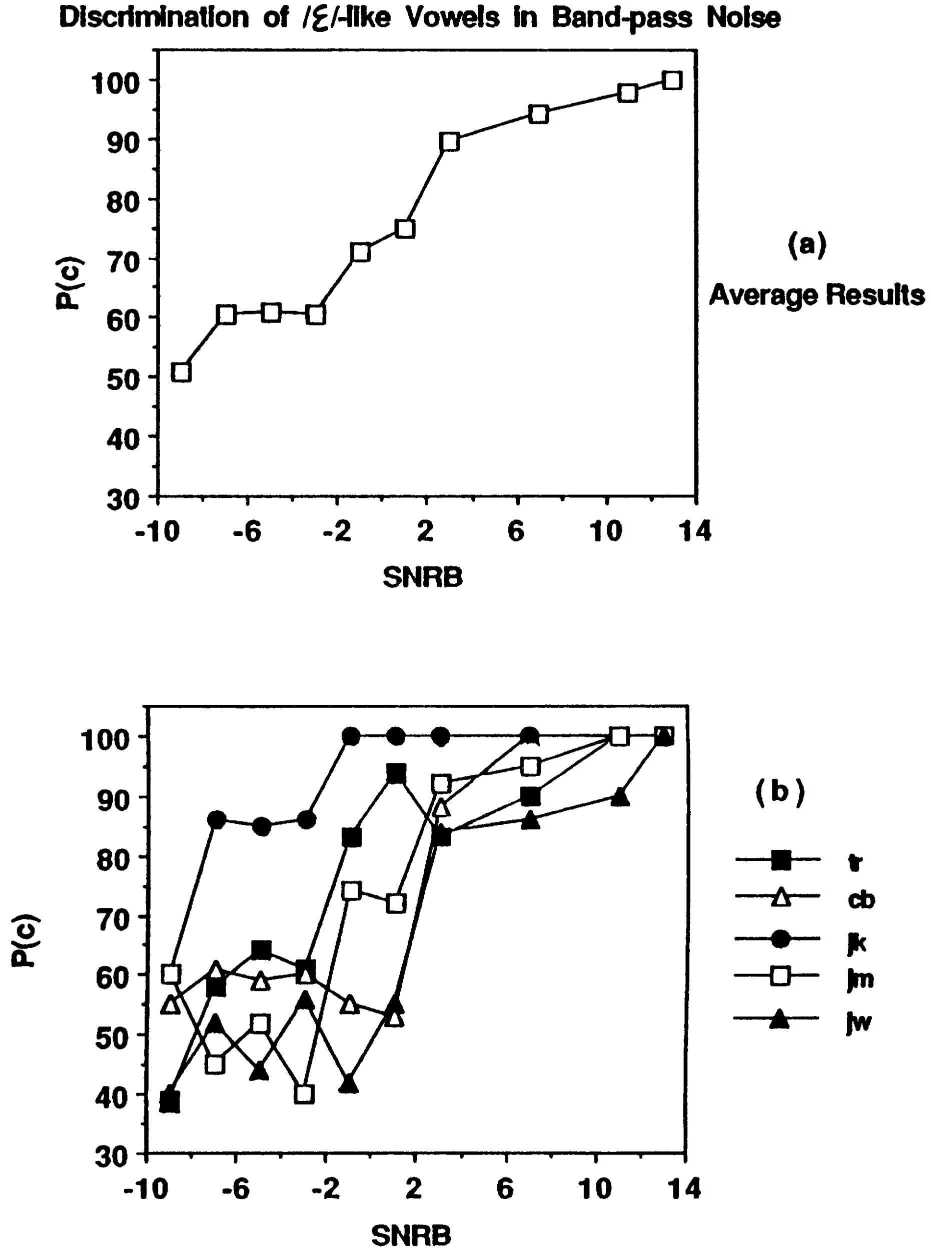


Figure 3.15: Psychometric functions of a vowel-in-noise experiment. Percent correct here reflects the ability of the subject to discriminate correctly between the two synthetic /ε/ vowels which differed only in the presence of F2 (/ε/) or absence of F2 (/εnoF2/). The masker was band-pass noise centered at 1550 Hz and was 1200 Hz in width. Average results are shown in (a) and individual results are shown in (b).

This subject's performance might be due to her ability to use amplitude differences in the incompletely masked harmonics in the two vowels. If this was the case, then the subject used amplitude differences in harmonics in the frequency region from 2150 Hz (this frequency is the upper edge of the noise spectrum) and 2500 Hz (above this frequency, harmonics in both spectra have the same amplitude, Figure 3.3). An alternate explanation is that the subject was using an artifact in one of the signals as a cue to discriminate between the two vowels.

3.4 Experimental Results: Above-band Masking

In these experiments, we measured the masked thresholds of tones at frequencies appropriate for F3 and F4 for the vowels /ɑ/ and /ɛ/ in the presence of a band-pass masker centered around a region appropriate for F2 for the two vowels. The goal was to develop a predictive measure of the masking of F3 and F4 when these formants are above the passband of a noise masker.

Two band-pass maskers centered at 1550 Hz were used. One of the maskers was 1200 Hz in width (n12) and was later used in C/ɑ/ experiments described in Chapter 4. The thresholds of tones at 2700 and 3250 Hz, which were appropriate frequency locations for F3 and F4, respectively, for the vowel /ɑ/, were measured in the presence of the masker (n12). The other band-pass masker used in these experiments was 1000 Hz in width (n10) and was later used in C/ɛ/ experiments. The tones tested in this case were at 2300, 2700, and 3500 Hz. The tones at 2300 and 2700 Hz were chosen because these values were the extreme values in the F3 range used in later experiments with synthetic C/ɛ/ utterances. The tone at 3500 Hz was close to the frequency location of F4 in the C/ɛ/ utterances.

Three subjects participated in these experiments: JK, CB, who participated in the previous experiments, and the author. The experiments were conducted at two overall levels of noise: 70 and 77 dB SPL.

Average results for the 'n12' and the 'n10' cases are shown in Figures 3.16 and 3.17, respectively. Also shown of the figures are predicted values using the ANSI procedures (S3.5, 1969) (referred to as PA), Equation 3.10 (Ludvigsen, 1985) (referred to as PL)[4], and predictions based on a modified ANSI procedure to be described below (referred to as P). Parts (a) and (b) in each figure show measured and predicted thresholds at overall masker levels of 70 and 77 dB SPL, respectively. Notice that the 7 dB increase in the masker level does not necessarily result in the same increase in the tone thresholds.

As seen in the figures, there are two problems with the original ANSI predictions: not predicting any spread of masking at 70 dB SPL, and overestimating thresholds at 77 dB SPL. At an overall level of 70 dB SPL the original ANSI procedures do not predict upward-spread of masking because the spectrum level in this case is less than the minimum spectrum level needed for the procedures to be effective. The minimum spectrum level specified by ANSI to be 46 dB SPL might have been chosen simply because the minimum spectrum level used in Carter and Kryter's study (1962) was 50 dB SPL. Hence, ANSI predicts that the above-band tone thresholds in Figures 3.16a and 3.17a would be the tone audibility thresholds, which are around 10 dB SPL in this case (ANSI-S3.6, 1969).

The difference between the experimental data and the predictions of the original ANSI standards for the overall noise level of 77 dB SPL might be due to differences in the skirts of the noise maskers used in our experiments versus those used in the original experiments on which the ANSI standards were based. For example, the noise skirts in our experiments rolled off at 100 dB / 1/3 octave whereas the masker skirts in the Bilger and Hirsh (1956) experiments rolled off at 36 and at 54 dB/octave, and in Carter and Kryter's study (1962) the skirts rolled off at 40 dB/octave. Bilger and Hirsh noticed differences in masked thresholds due to differences in the skirts of the noise masker. However, the ANSI standards do not accommodate for differences in

[4] For further details on these procedures, see Sections 1.1.1 and 3.1.2.

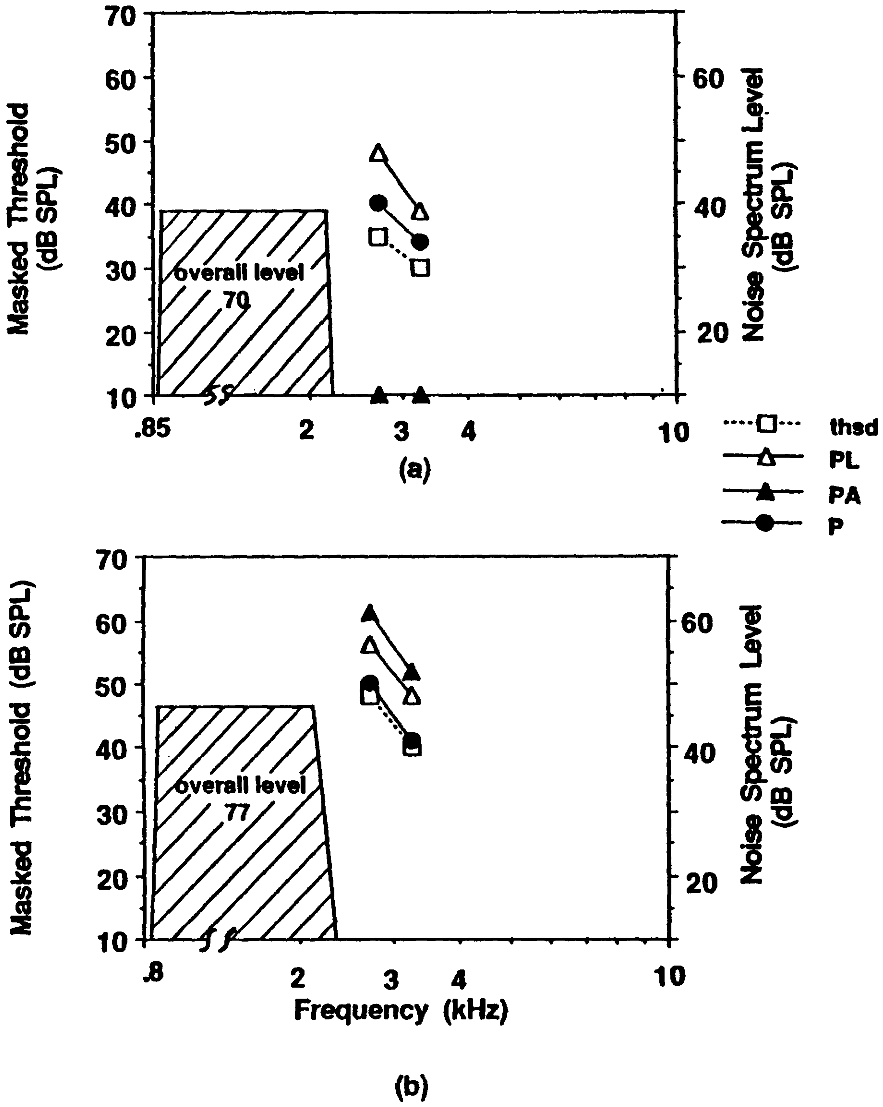


Figure 3.16: Thresholds of tones located at 2700 Hz and 3250 Hz. Thresholds are average values from three subjects. The masker was 1200 Hz wide and was centered at 1550 Hz. Two overall masker levels were used: (a) 70 dB SPL and (b) 77 dB SPL. Dashed lines refer to measured values and solid lines refer to predicted values using: ANSI (PA), Ludvigsen (PL), and a modified ANSI algorithm (P). ANSI did not predict spread of masking at 70 dB SPL, hence, audibility thresholds of the tones were used.

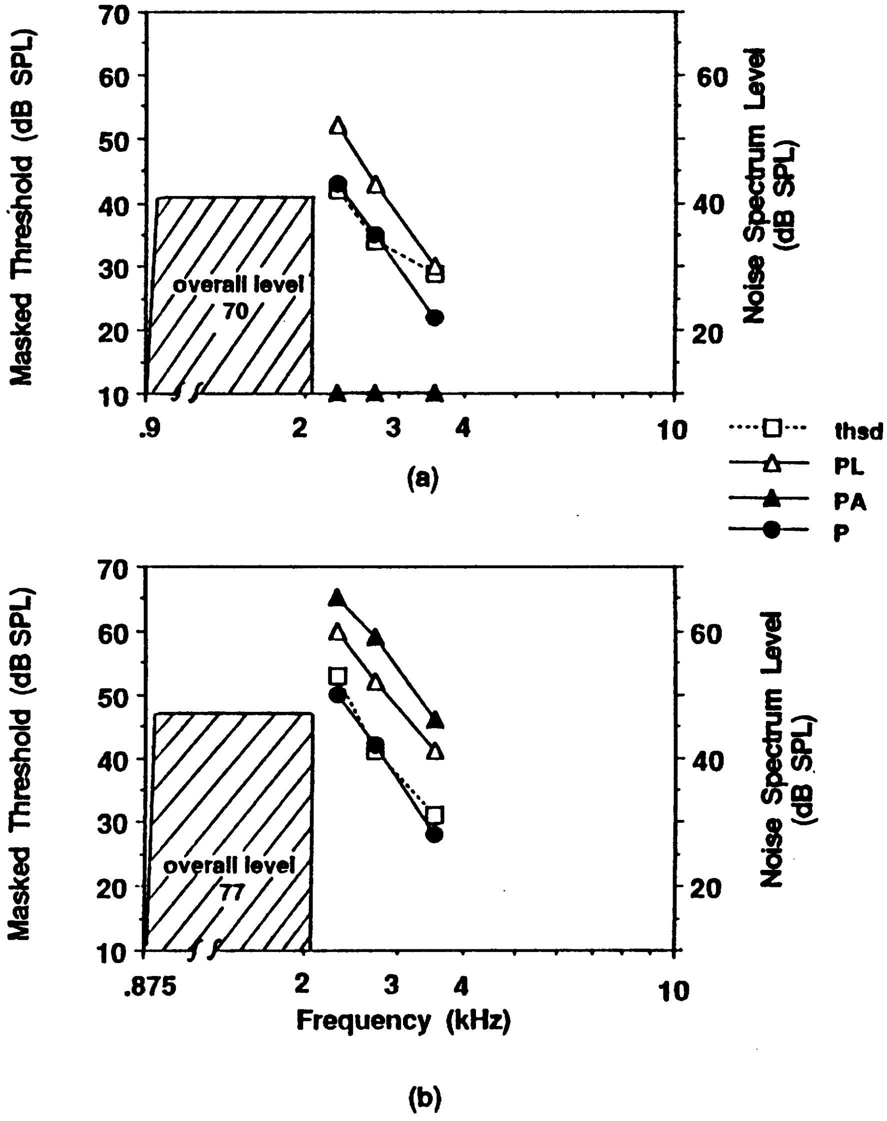


Figure 3.17: Thresholds of tones located at 2300, 2700, and 3500 Hz. Thresholds are average values from three subjects. The masker was 1000 Hz wide and was centered at 1550 Hz. Two overall noise levels were used: (a) 70 dB SPL and (b) 77 dB SPL. Dashed lines refer to measured values and solid lines refer to predicted values using: ANSI (PA), Ludvigsen's equation (PL), and a modified ANSI algorithm (P). ANSI did not predict spread of masking at 70 dB SPL, hence, audibility thresholds of the tones were used.

skirt characteristics.

Ludvigsen's predictions were closer to the experimental results than were those derived from the ANSI standards. Ludvigsen (1985) compared calculated and measured thresholds from hearing-impaired subjects. He calculated the thresholds using Equation 3.10 with a modification to account for hearing loss. In one case, the difference between the measured and calculated thresholds was as high as -9 dB (the calculated values being higher). Hence, it is not clear that his equation could be used to predict masked thresholds of normal-hearing subjects accurately.

The modified ANSI procedure, which was empirically derived to account for the experimental data, addressed the problem of overestimating masking by modifying the 'starting point' of the effective spectrum of the noise masker. The original ANSI procedures specified the starting point of the effective spectrum to be the frequency which was 3 dB below the maximum point in the noise spectrum. The starting point we used was 0.4 octave below the cut-off frequency, instead of the 3 dB down point; this modification reduced masking overestimation. The second modification involved extending the lower limit of the noise spectrum level which causes spread of masking from 46 dB SPL, as specified by the original ANSI standards, to 36 dB SPL. This change allowed us to account for above-band masking at 70 dB SPL.

In order to determine if our method of estimating masked thresholds is valid for data reported elsewhere, we examined experimental data on upward-spread of masking by Gagné (1988). In Gagné's paper, data on the masking effects of three noise maskers of different widths and for overall levels between 50 and 100 dB SPL were shown. Here, we will only examine the thresholds of some of the frequencies masked by a low-pass noise masker of width 1175 Hz because of the similarity to the masker bandwidth used in our experiments, and at overall levels of 67 and 77 dB SPL. The upper cutoff slope of the masker was greater than 90 dB/octave. Figure 3.18 shows the experimental data from Gagné's study along with the theoretical predictions using the original ANSI standards (PA), predictions using Equation 3.10 (PL), and our modified ANSI algorithm (P). Again, we notice that the 'modified ANSI' predictions were the closest to the experimental data.

While our modified algorithm provides a reasonably good prediction of both our data and those of Gagné, different noise maskers might require a different modification to predict above-band masked thresholds accurately. However, the ANSI procedures and Ludvigsen's equation seem to be a good framework for the improved predictive methods. Iterative procedures could be used to modify and optimize predictive functions to accommodate different experimental conditions. It appears that, in the case of above-band masking, and consequently upward-spread of masking, several characteristics of the noise masker should be taken into consideration: bandwidth, overall level, cutoff frequency, and cutoff slope of the masker.[5]

[5] The conclusion that bandwidth might be a factor in determining spread of masking, is based on pilot experiments not reported here.

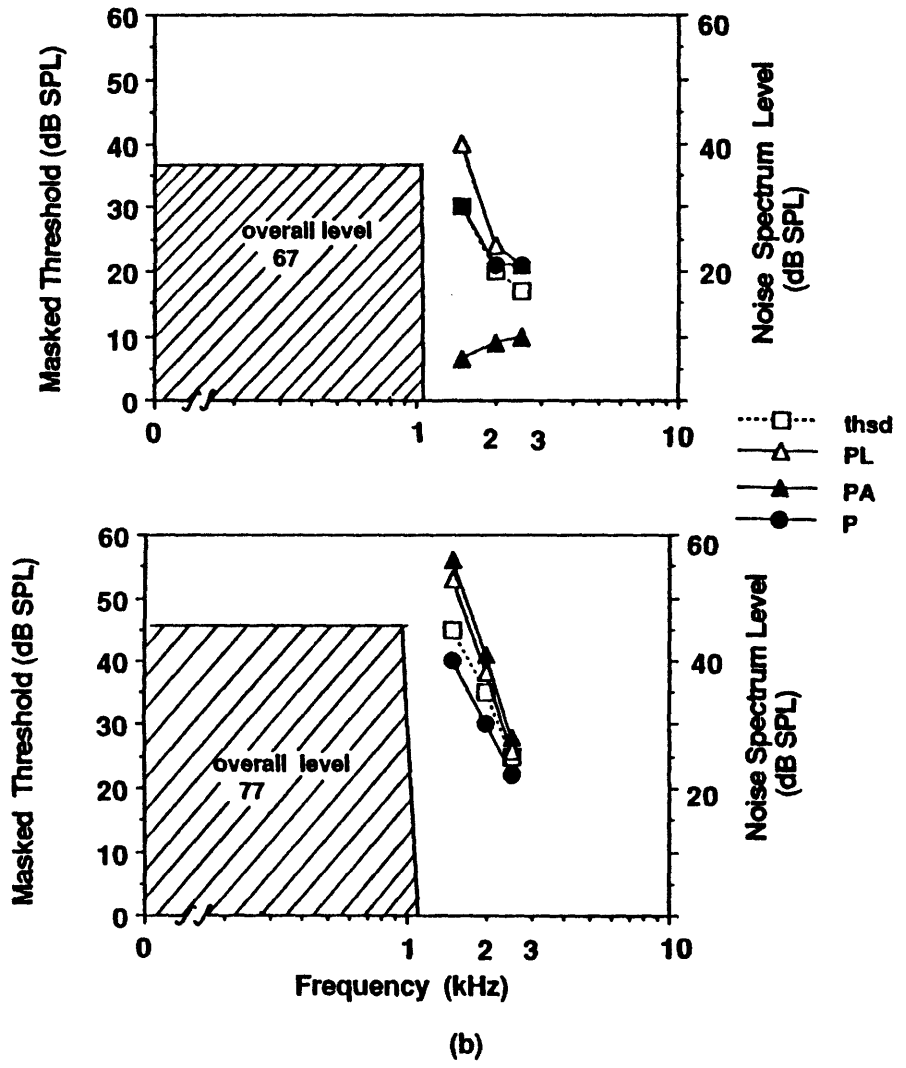


Figure 3.18: Thresholds of tones located at 1500, 2000, and 2500 Hz (from Gagné, 1988). The low-pass masker was 1175 Hz wide. Data for two overall noise levels are shown: (a) 67 dB SPL and (b) 77 dB SPL. Dashed lines refer to measured values and solid lines refer to three predictions: ANSI (PA), Ludvigsen (PL), and a modified ANSI algorithm (P). ANSI did not predict spread of masking at 67 dB SPL, hence, audibility thresholds of the tones were used.

3.5 Summary

In this chapter, theory and experiments of within-band masking (tone/formant within the masker's bandwidth) and above-band masking (tone/formant above the masker's bandwidth) were discussed. For the within-band masking case, experiments involved: 1) Tone-in-noise and vowel-in-noise detection tasks; the vowels in this case had one formant frequency, and 2) Vowel discrimination tasks where the perceptual similarity of two vowels was examined in the presence of noise. One vowel was synthesized with all formant frequencies and the other, with all formants except for F2. The vowels used in these experiments were /ɑ/-like and /ɛ/-like. Two maskers were used: white noise, and band-pass noise. In both cases, F2 for the vowels was within the bandwidth of the noise masker.

Experiments on above-band masking involved tone-in-noise detection tasks. The tones were located at frequencies appropriate for F3 and F4 for the /ɑ/ and /ɛ/ vowels and two band-pass maskers, both centered at 1550 Hz, were used: one of the maskers was 1000 Hz in width (n10) and the other, 1200 Hz in width (n12).

It was found that the signal-to-noise ratio in a critical band ($SNRB$) was a good predictive measure of average results of tone and formant masking for the 'within-band' case. On the average, an $SNRB$ value of zero dB or less corresponded to the masking of a tone or formant frequency. The intrasubject variability in the band-pass masker case was larger than that with white noise. The larger variability was attributed by the subjects to interference from the perceived 'pitch' of the band-pass noise. Subjects also reported that the vowel discrimination task was more difficult for them than the tone or vowel detection tasks. The increased difficulty with the discrimination task was attributed by two phonetically-trained subjects to the fact that they were attempting to label the vowels phonetically rather than treating them as 'abstract' entities. At high $SNRB$, the 'multi-formant' vowels were heard as clear /ɑ/'s or /ɛ/'s. However, as the $SNRB$ decreased, the phonetic identity of the vowel became ambiguous, resulting in larger uncertainty in the subjects' responses.

Results of above-band masking could not be predicted well from either the ANSI procedures or from the equation proposed by Ludvigsen (1985) (Section 3.1.2). A modified ANSI procedure was then developed to account for the experimental results. One explanation for the difference between the experimental results and predictions based on the above-mentioned procedures was that the shape of the noise masker used in the earlier studies was different than that used in this study. In addition, the ANSI procedures, although modified to account for tone masking by band-pass noise maskers, were originally intended to measure an index related to speech intelligibility in noise and not to account precisely for tone thresholds.

Based on the results of this chapter, a metric was developed to predict the masking of formant frequencies in synthetic CV utterances. This metric is given in the following chapter. For the 'within-band' masking case, the $SNRB$ is used as a predictive measure. The modified ANSI procedure is used to predict the masking of F3 and F4 when these higher formants are above the passband of a masker centered at F2.

Chapter 4

Masking the Place of Articulation Distinction for /b/ and /d/

In this chapter, theoretical predictions of formant frequency masking in synthetic CV syllables are discussed. The predictions are based on the analytical tools developed in Chapter 3. Results of perceptual experiments in which the stimuli are synthetic CV syllables mixed with noise are then presented. The consonant is either /b/ or /d/, the vowel, /ɑ/ or /ɛ/, and the masker is either white noise or band-pass noise centered around the F2 region. The goal of the experiments is to investigate the perceptual importance of F2 and of higher formant frequencies in distinguishing between the consonants /b/ and /d/ in noise.

4.1 Theoretical Predictions

In this section, the method used for predicting masking of formant frequencies in synthetic CV utterances is described.

For the 'within-band' masking case, which includes masking of all formant frequencies by a white-noise masker and masking of F2 by a band-pass masker centered around the F2 region, the method used to predict formant masking involves: 1) estimating the level of the *ith* formant peak (A_i) and 2) estimating the noise level in a critical band centered at the *ith* formant (N_{ci}). The *ith* formant peak is assumed to be masked if N_{ci} is greater than $A_i + 4$ dB or, equivalently, the signal-to-noise ratio in the *ith* critical band ($SNRB$) is 0 dB (Equation 3.8). As mentioned in Chapter 3, the level of a formant peak (A_i) is considered to be the sum of the power of the harmonics in the *ith* critical band, whereas (N_{ci}) is the overall level of the noise minus 10 log (ratio of the noise bandwidth to the critical band centered at the *ith* formant frequency). Estimation of A_i and N_{ci} is based on Discrete Fourier Transform (DFT) spectra computed using a 256 point (or 25.6 ms at a 10 kHz sampling rate) Hamming window. The analysis window is centered at the closed phase of a glottal pulse when analyzing the utterances.[1] The same analysis window is used for analyzing both the signal and the noise masker. It is necessary to update the calculations of A_i and N_{ci} during the transition period of the formants because a change in the location of a formant frequency results in a change in the value of the critical bandwidth (hence, affecting the calculations of both N_{ci} and A_i), and also results in a change in the formant amplitude and the amplitudes of all other formant peaks (Equation 2.1). In other words, a change in a formant frequency results in a change in the $SNRB$ at each formant peak. In this study, the calculations are updated every pitch period. Based on these calculations, the time interval within which each formant peak is masked is determined.

Figure 4.1 illustrates these computations for a white-noise masker at a particular level for which the F2 transition in the synthetic /dɑ/ utterance is partially masked. In the figure, the term $\acute{A}_i$ equals $A_i + 4$ dB. As shown in Figure 4.1, masking theory

[1] At the onset of the study DFT spectra were computed using three Hamming windows of different durations: 12.8 ms, 25.6 ms, and 51.2 ms. It was found that spectral estimation using the shortest window (12.8 ms) was highly sensitive to where in the glottal pulse the waveform was sampled, and as a result, the computations, especially those of the time course of the formant amplitudes, were prone to errors. DFT spectra computed with the largest window (51.2 ms) did not illustrate clearly spectral changes occurring in each pitch period during the transition period (because of greater smoothing in time in comparison with the other windows). Hence, a window of 25.6 ms was chosen for our purposes. If our objective was to analyze other aspects of the signal, such as a stop release, then a shorter window might be adequate. In contrast, a larger window might be desired when analyzing the steady-state part of the vowel.

predicts that F1 in the synthetic utterance is never masked ($N_{c1} < \acute{A}_1$), F3 is always masked ($N_{c3} > \acute{A}_3$), and only the first two pitch periods of the F2 transition are masked ($N_{c2} > \acute{A}_2$ in the first two pitch periods only). Note that the amplitude of the spectral peak of F2 changes by about 9 dB during the transition period. Figure 4.2 summarizes the calculations of Figure 4.1 by showing on a schematized /dɑ/ spectrogram the time interval in which each formant peak is predicted to be masked. Similar calculations are done for all other synthetic utterances.

In the 'above-band' masking case, which includes the masking of F3 and F4 due to a band-pass masker centered at the F2 region, the level of a formant peak is estimated with the same procedure as that described in the within-band case. The formant levels and the spectrum level of the noise masker (N_0) are converted to dB SPL using the calibration curves of the headphones. The effective spectrum level of the band-pass masker is then estimated from the modified ANSI procedure described in Section 3.4. The threshold of a formant peak (in dB SPL) is predicted to be the sum of the effective spectrum level and the critical ratio at the frequency of the formant. If the formant level is below the threshold value then the formant is masked, otherwise, it is not.

Figure 4.3 illustrates an example where both within-band and above-band masking calculations are used to predict formant masking. The Figure shows a short-time spectrum sampled 8 ms into the vowel in a /dɑ/ utterance along with a schematized spectrum of a band-pass masker. The spectrum level of the masker in this case is 46 dB SPL, which corresponds to an overall level of 77 dB SPL. The solid-line spectrum is the physical spectrum of the masker, and the dashed-line spectrum is the effective spectrum of the masker as predicted from the modified ANSI algorithm. Formant levels are indicated by triangles and the predicted thresholds of these formants are indicated by circles. Masked thresholds are computed by adding the noise spectrum level to the critical ratio at that frequency. In the F2 case, the physical spectrum level is used to predict thresholds (since F2 is within the bandwidth of the masker), and in the F3 and

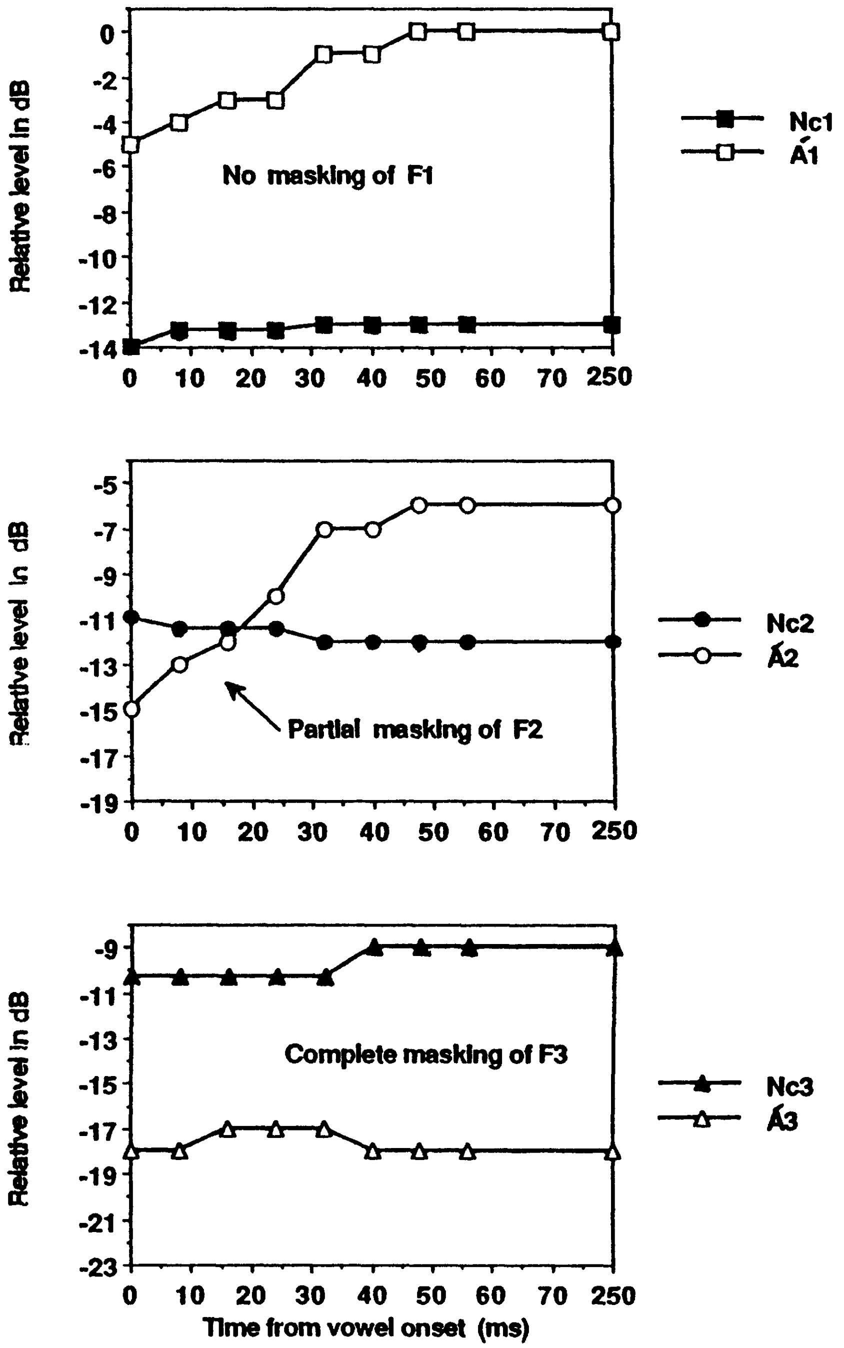


Figure 4.1: Plots of the levels in dB of the amplitudes of the first three formant frequencies plus 4 dB ($\acute{A}_i$), along with the noise levels in critical bands centered at each formant frequency (N_{ci}). Both $\acute{A}_i$ and N_{ci} levels are relative to the level of F1 in the steady-state part of the vowel. Masking is predicted to occur when $N_{ci} > \acute{A}_i$. The masker is white noise. Estimates of the levels are based on DFT spectra and are computed every pitch period in the synthetic /da/ utterance.

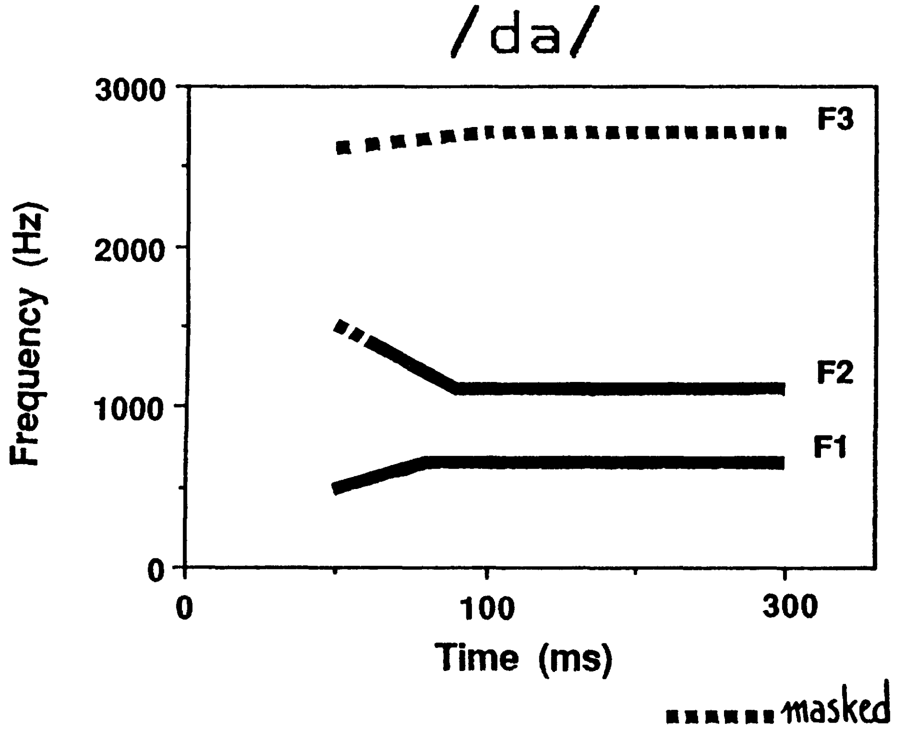


Figure 4.2: Schematized spectrogram of the synthetic /da/ utterance in white noise. The signal-to-noise ratio is, according to theory, such that F1 is not masked, only the first two pitch periods of F2 are masked, and F3 is masked completely. The time interval in which each of the three formant peaks is predicted to be masked is indicated by dashed lines. This figure summarizes the calculations shown in Figure 4.1.

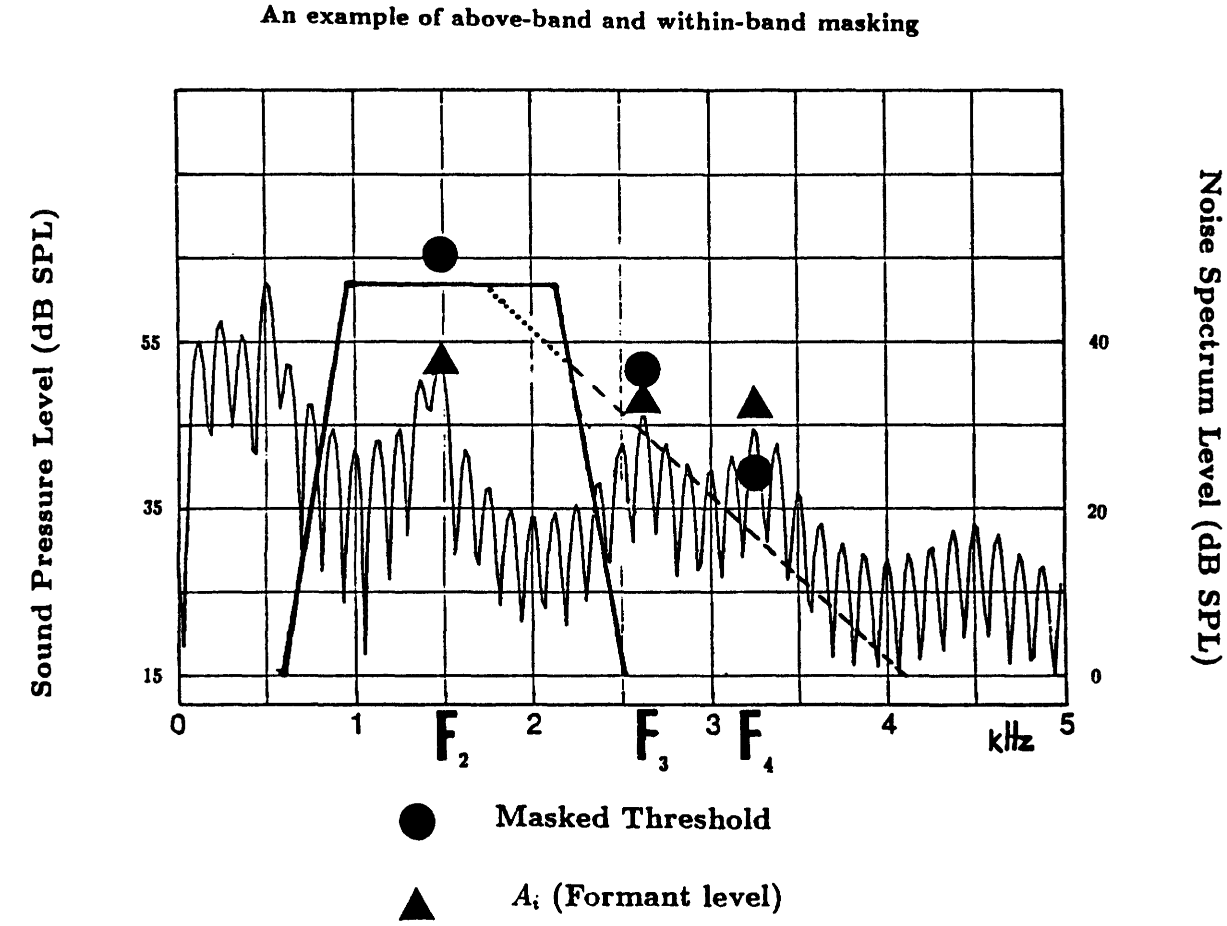


Figure 4.3: Short-time spectrum sampled 8 ms into the /da/ utterance and the spectrum of a band-pass masker. The solid-line spectrum is the physical spectrum of the masker and the dashed-line spectrum is the effective spectrum of the masker (derived from the modified ANSI algorithm). Triangles are estimated formant levels (A_i) and dots represent predicted thresholds. According to the predictions, F2 and F3 are masked ($A_2, A_3 < \textit{threshold}$) and F4 is not ($A_4 > \textit{threshold}$). Thresholds of F3 and F4 were predicted by adding the effective spectrum level of the noise and the critical ratio at that frequency. F2 threshold is calculated using within-band masking predictions.

F4 case, the effective spectrum level is used (since both F3 and F4 are above the passband of the masker). F2 and F3 are predicted to be masked (A_2, and $A_3 < \textit{threshold}$) whereas F4 is not. Similar predictions are done every pitch period for all synthetic utterances used in the experiments.

4.2 Experimental Methodology

4.2.1 Stimuli

Stimuli consisted of synthetic CV syllables mixed with noise at different signal-to-noise ratios (SNR). The SNR was varied by changing the signal level in 1, 2 , or 4 dB steps, depending on the experiment, while keeping the noise level constant. The range of SNR chosen for each experiment was such that the stimuli ranged from clearly identifiable to non-distinguishable. Pilot experiments with phonetically trained subjects were used to determine which stimuli were correctly identifiable and which were not. The consonant in these utterances was either /b/ or /d/ and the vowel /ɑ/ or /ɛ/. Synthesis of the utterances was done using the cascade part of the formant synthesizer KLSYN88 (1990) and was described in detail in Chapter 2. Spectrograms of the synthetic utterances are shown in Figure 4.4. Frozen noise waveforms were used as maskers. The noise masker, which was 500 ms in duration, was either white noise or band-pass noise. The band-pass maskers were centered at 1550 Hz and were 1200 Hz in width, for the C/ɑ/ experiments, and 1000 Hz in width (n10), for the C/ɛ/ experiments. Section 3.2.3 included a description of the noise generation procedure and Figure 3.5 showed spectra of the band-pass maskers used in the experiments. Each CV utterance, which was 250 ms in duration, was centered in the middle of a noise waveform.

The reason for choosing nonsense syllables rather than meaningful words in the perceptual experiments was to make sure that lexical effects, such as word frequency, did not bias subjects' responses.

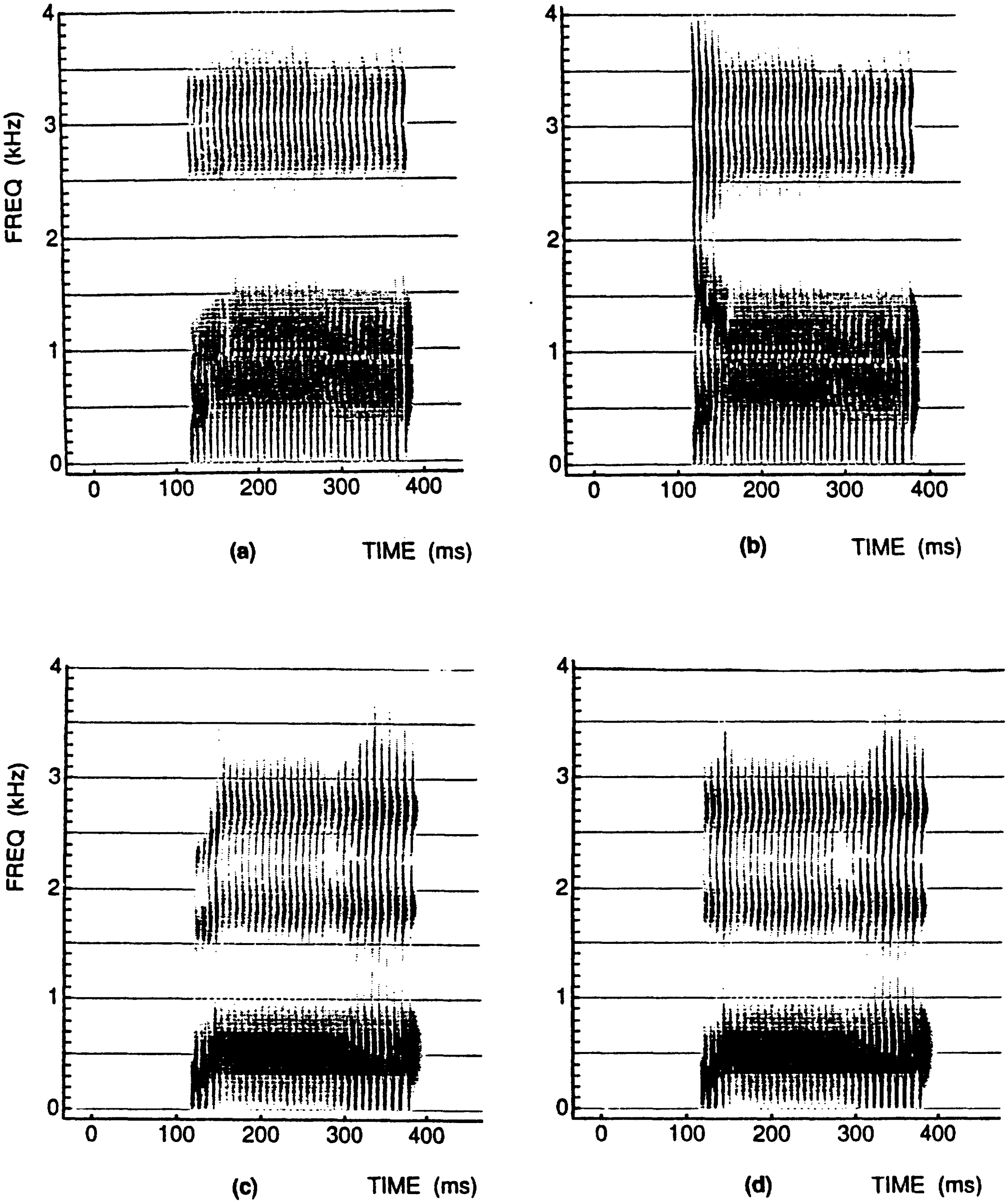


Figure 4.4: Spectrograms of the synthetic utterances a) /bɑ/ and /dɑ/ and b) /bɛ/ and /dɛ/. The spectrograms were computed with a 6.4 ms Hamming window.

4.2.2 Paradigm

A paradigm similar to that described in Section 3.2.1, with the exception of the type of task used, was adopted here. The task for the experiments described in this chapter was forced choice identification. Subjects were presented with stimuli and were forced to identify the consonant they heard as either /b/ or /d/. The subjects had the option of writing next to their answer and in parentheses any consonant other than /b/ or /d/ they might have heard. If they didn't hear anything at all, they were instructed to write down an (x) next to their answer.

Each experiment was designed such that half of the stimuli sounded like /b/, and the other half like /d/. This procedure was done to eliminate bias. Determining whether an utterance sounded /b/-like or /d/-like in noise was based on pilot experiments. The stimuli were repeated 20 times, recorded in random order onto TDK AD60 cassette tapes and presented to subjects binaurally over Sennheiser headphones in a sound treated room. The inter-stimulus duration was 3 s and there was a longer pause of 4.5 s after each block of ten trials. In the beginning of each 'white-noise' experiment, subjects were free to adjust the volume on the tape deck to a comfortable level as long as the maximum level resulting from that adjustment did not exceed 90 dB SPL. In the band pass case, the subjects could choose between two overall noise levels: 70 or 77 dB SPL. There was a separate cassette prepared for each level to ensure that the criterion that half the stimuli sounded /b/- or /d/-like was met in both cases. Training sessions of 1/2 - 1 hour, aimed at familiarizing the subjects with the computer-generated speech and with the task, preceded each experiment.

4.2.3 Subjects

Subjects (CB, JM, JK, JW, TR) who participated in the tone and vowel experiments described in Chapter 3 participated in the CV experiments. Subject TR did not participate in the C/ɑ/ experiments. Other subjects (KNS, SSH, NN), whose initials appear on some of the figures, participated in some of the experiments. All subjects, with the exception of subjects CB, KNS, and SSH, had no previous exposure to synthetic speech sounds. None of the subjects had any known speech or hearing problems.

4.2.4 Results

Subjects' responses are represented in terms of identification functions plotted as a function of the signal-to-noise ratio in a critical band centered around F2 at the steady state part of the vowel ($SNRB$). The critical band at F2 for /ɑ/ is 190 Hz and for /ɛ/, 280 Hz (Reed and Bilger, 1973). Responses to inaudible stimuli are not included in the plots.

4.3 Experimental Results: White-Noise Masker

In these experiments, synthetic CV utterances were mixed with white noise and presented to subjects in identification tests.

4.3.1 The C/ɑ/ Case

As seen in Figure 4.4a, the two synthetic C/ɑ/ utterances differed in the F2 trajectory; /dɑ/ had a falling F2 trajectory and /bɑ/, a slightly rising F2 trajectory. Prior to conducting the experiment it was anticipated that if the noise masks the F2 peak at the onset of a stimulus, then confusions between the two consonants would occur. F3 and higher formant frequencies were predicted to be masked in all the stimuli used in this experiment whereas F1 was not predicted to be masked in any of the stimuli.

All subjects, with the exception of subject JK, whose responses will be reported separately, responded to the /bɑ/ stimuli correctly at all noise levels. Figure 4.5a shows average responses for the /dɑ/ stimuli and individual responses are shown in Figure 4.5b. Shown on the top axis in Figure 4.5a is the time interval in which the F2 peak was estimated to be masked for selected stimuli. For example, an $SNRB$ of 4 dB was estimated to mask 25 ms of the F2 transition, whereas an $SNRB$ of 0 dB was estimated to mask all of the F2 transition time.

The identification functions in Figure 4.5 show an abrupt shift from /dɑ/ to /bɑ/ responses as the $SNRB$ decreased. The average 50% correct point corresponded to the stimulus where calculations indicated that 25 ms of the F2 transition was masked.[2] Recall that the total F2 transition time was 40 ms.

[2] One of the phonetically trained subjects who participated in the pilot experiments reported that some of the consonants she labelled as /b/ could have been also labelled as /p/, /f/, or /v/. Likewise, some of the consonants labelled as /d/ could have been labelled as /t/, /s/, /z/, or /n/.

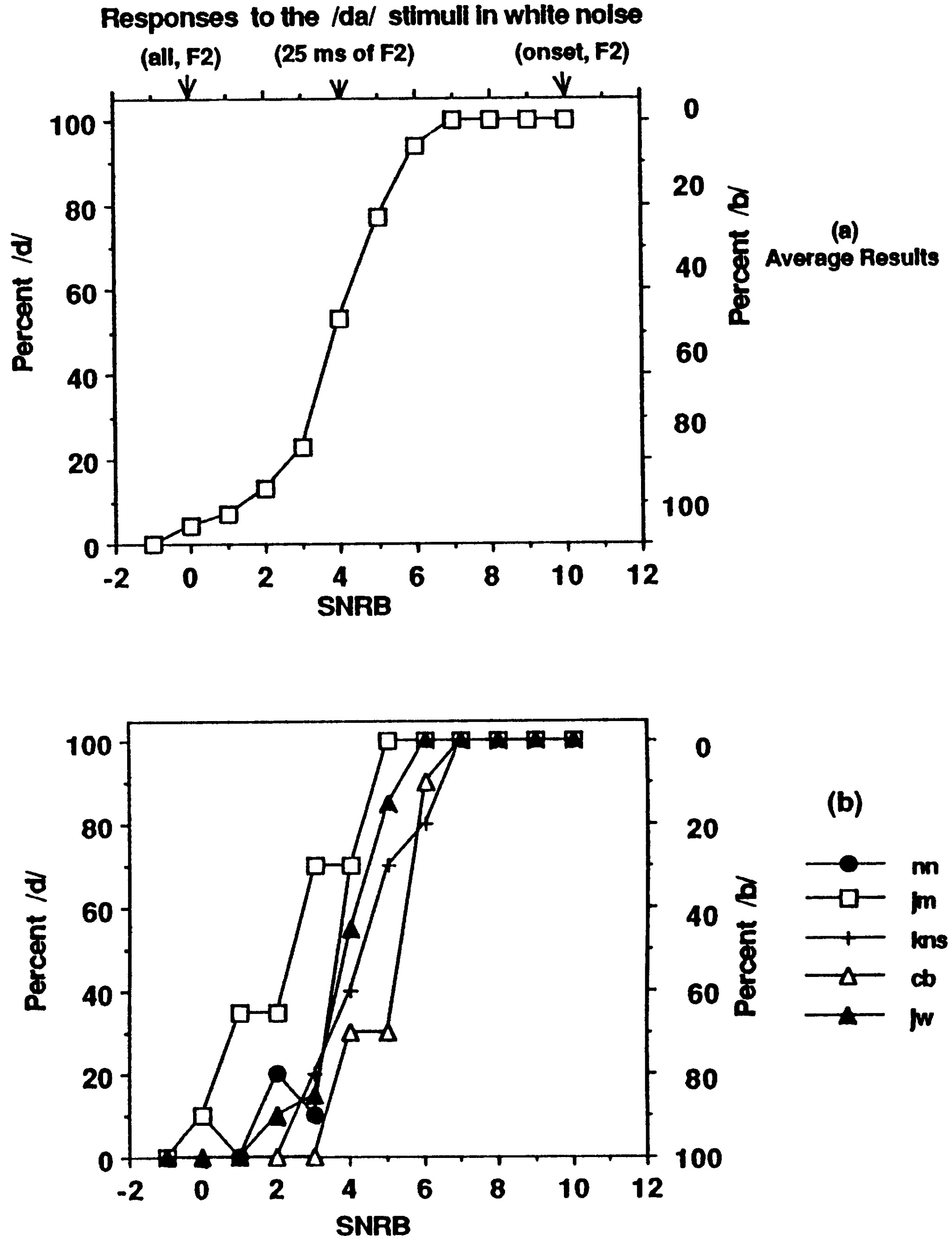


Figure 4.5: Plots of the subjects' responses to the /da/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the top axis in part (a) indicate the time interval in which F2 was predicted to be masked for selected stimuli. For example, (onset, F2) indicates that only the onset value of F2 was, theoretically, masked. According to predictions, F3 and higher formants were masked in all stimuli used in the experiment, whereas F1 was not masked in any of the stimuli. The masker was white noise.

Subjects differed somewhat in their choice of F2 transition time needed to identify /d/ correctly. For example, the responses of subject CB shifted to /b/ after 20 ms of the F2 transition was masked. Subject JM, on the other hand, shifted his responses to /b/ when at least 26 ms of the F2 transition was masked.

Subject JK's response pattern was different from that of the other subjects. Figure 4.6 shows the responses for JK. As shown in the figure, JK perceived the /bɑ/ and /dɑ/ stimuli correctly until most of the F2 transition in both consonants was masked. When most or all of the F2 transition was masked, the subject did not default to either consonant but, rather, her responses were close to being random.

There are two possible explanations for the subjects' responses in the current experiment; the first is that there was simply a *bias* towards /b/ responses; that is, subjects defaulted to /b/ when the signal level was too weak to identify either consonant. An alternative explanation derives from the observation that in the context of the vowel /ɑ/, the labial feature for /b/ was signalled by an almost flat F2 trajectory whereas the perception of the alveolar /d/ was signalled by a falling F2 trajectory. If the $SNRB$ was at a value where part or most of the F2 transition in the /dɑ/ utterance was masked, then subjects heard a 'flat' trajectory and identified the stimulus as /b/. Recall that F2 moved, during the transition time, by 130 Hz in the /bɑ/ utterance, whereas it moved by 400 Hz in /dɑ/.

As mentioned earlier, the only exception was subject JK who did not identify a 'flat' F2 trajectory as /b/-like; she seemed to identify /b/ only if the F2 trajectory was rising.

Perceptual experiments with synthetic C/ε/ syllables in noise were conducted to provide some insights into distinguishing between the two explanations offered above. Preceding the vowel /ε/, the consonants /b/ and /d/ have different formant trajectory shapes than those observed with /ɑ/ (see Chapter 2 for further details). F2 and F3

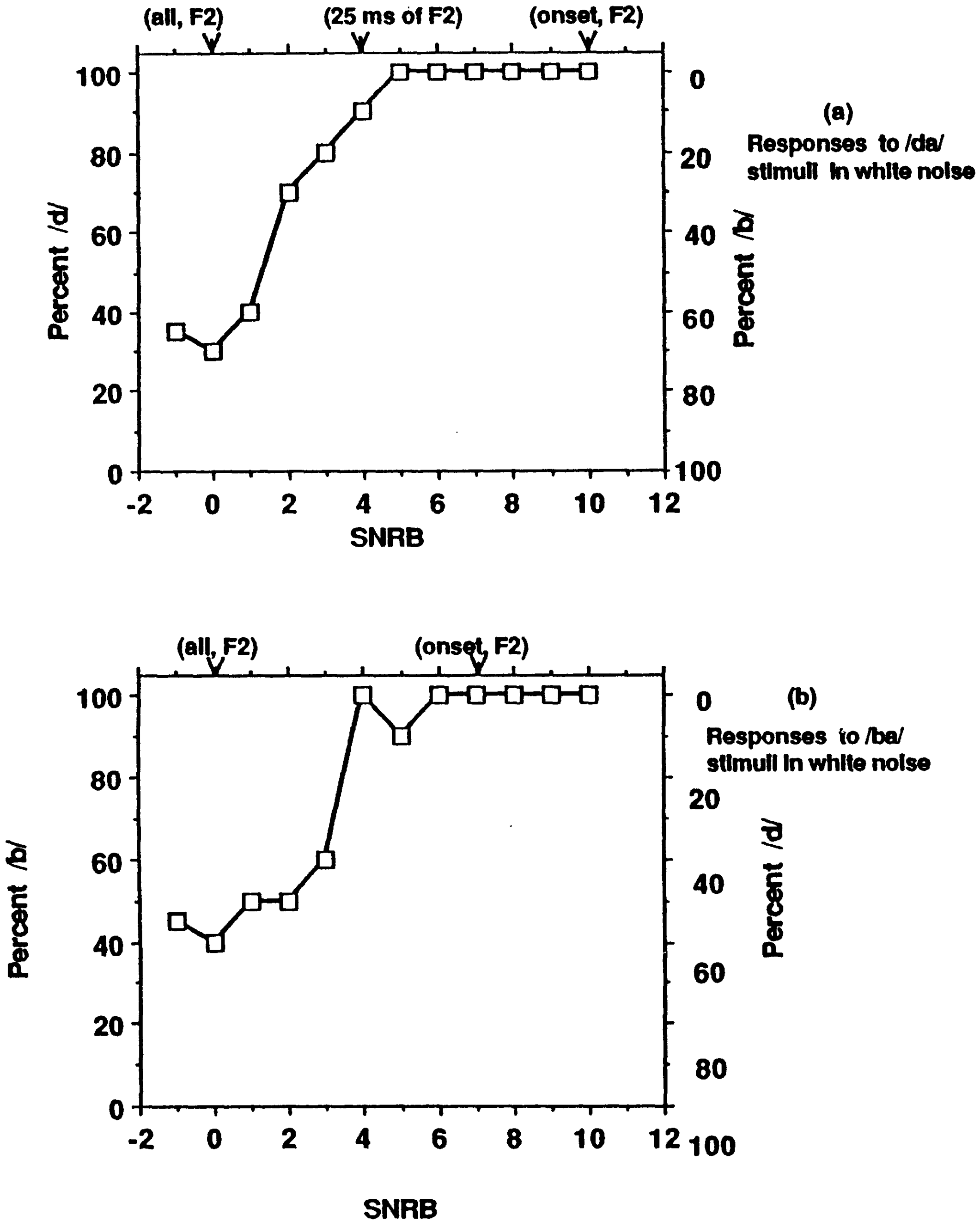


Figure 4.6: Plots of the responses for subject JK to: (a) /da/ stimuli and (b) /ba/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Arrows on the top axes indicate the time interval in which F2 was predicted to be masked for selected stimuli. The masker was white noise.

trajectories are relatively flat for /dɛ/, whereas these formants rise steeply into the vowel for /bɛ/. If there is simply a bias towards /b/ responses, then the responses would probably shift from /d/ to /b/ as the *SNRB* decreases. If, on the other hand, the perception of either consonant is signalled by the shape of the formant trajectories, then the opposite shift should occur. That is, as the signal level degrades, the rising formant trajectories for /bɛ/ would be masked and the 'flat' trajectories should lead to the perception of /dɛ/. The following section summarizes the results of the C/ɛ/ experiment.

4.3.2 The C/ɛ/ Case

In this experiment, synthetic C/ɛ/ syllables were mixed with noise at various *SNR* and presented to subjects in identification tests. As shown in Figure 4.4b, the two synthetic utterances (/bɛ/ and /dɛ/) differed in the shape of the F2 and F3 trajectories because a difference in only the F2 trajectory did not result in the perception of a clear /b/ or /d/ (see Section 2.2.2).

Results of this experiment show that the /dɛ/ stimuli were identified correctly at all noise levels. Responses for the /bɛ/ stimuli shifted from /b/ to /d/ as the signal was degraded by noise.[3] Figure 4.7a shows average responses for the /bɛ/ stimuli and individual responses are shown in Figure 4.7b. The time intervals in which calculations indicate that F2 and F3 in selected stimuli are masked are shown on the bottom and top axes, respectively, in Figure 4.7a. For example, an *SNRB* of 6 dB was estimated to mask 16 ms of the F3 transition and not to mask F2 whereas an *SNRB* of 1 dB was estimated to mask 24 ms of the F2 transition and all of the F3 transition. The total transition times for both F2 and F3 were 30 ms and these formants moved by 200 and

[3] It should be noted here that the two phonetically trained subjects who participated in pilot experiments reported that the phonetic quality of the vowel in the C/ɛ/ syllables changed from /ɛ/ to /o^w/ at low *SNRB* values. The consonant at these low *SNRB* values was still heard as /d/.

400 Hz, respectively, for /bɛ/ whereas /dɛ/ had flat F2 and F3 trajectories. F1 was not masked in any of the stimuli used in the experiment.

The identification functions show that both F2 and F3 play an important role in identifying the consonants with the vowel /ɛ/. This result is not entirely surprising since it has been shown that front vowels can be approximated well by two-formant synthetic vowels, the two formants being at F1 and $\acute{F2}$ (Carlson et al., 1970, 1975). In their study, Carlson et al. found that the effective second formant frequency ($\acute{F2}$) can be predicted from a set of equations which was empirically derived. We used Carlson et al. (1975) equations to predict $\acute{F2}$ for our data at the steady-state part of the vowel ($F_1 = 500, F_2 = 1800, F_3 = 2700, F_4 = 3600$ Hz) and found that $\acute{F2}$ is predicted to be at 2300 Hz, which is closer to F3 than it is to F2. It is not clear that the same set of equations proposed by Carlson et al. would hold for English vowels, but their conclusions do suggest that the effective second formant frequency ($\acute{F2}$) is somewhere between F2 and F3 for front vowels. Hence, if the place of articulation of a consonant is identified mainly by the formant transitions in the adjacent vowel and the vowel were a front vowel, then either F3 alone or both F2 and F3 would be important in signalling the place information.

In summary, the results show that when the transition intervals of F3 and F2 in the /bɛ/ utterance were masked, /dɛ/ was perceived. These results validate the second explanation proposed in the previous section which hypothesizes that the shapes of the formant trajectories are important in signalling the place of articulation for /b/ and /d/. If the $SNRB$ is at a value where most of the formant transitions are masked, then flat trajectories are perceived, leading to the perception of /b/ preceding the vowel /ɑ/, and to the perception of /d/ preceding the vowel /ɛ/.

A final remark on the experiments with a white-noise masker: subjects agreed

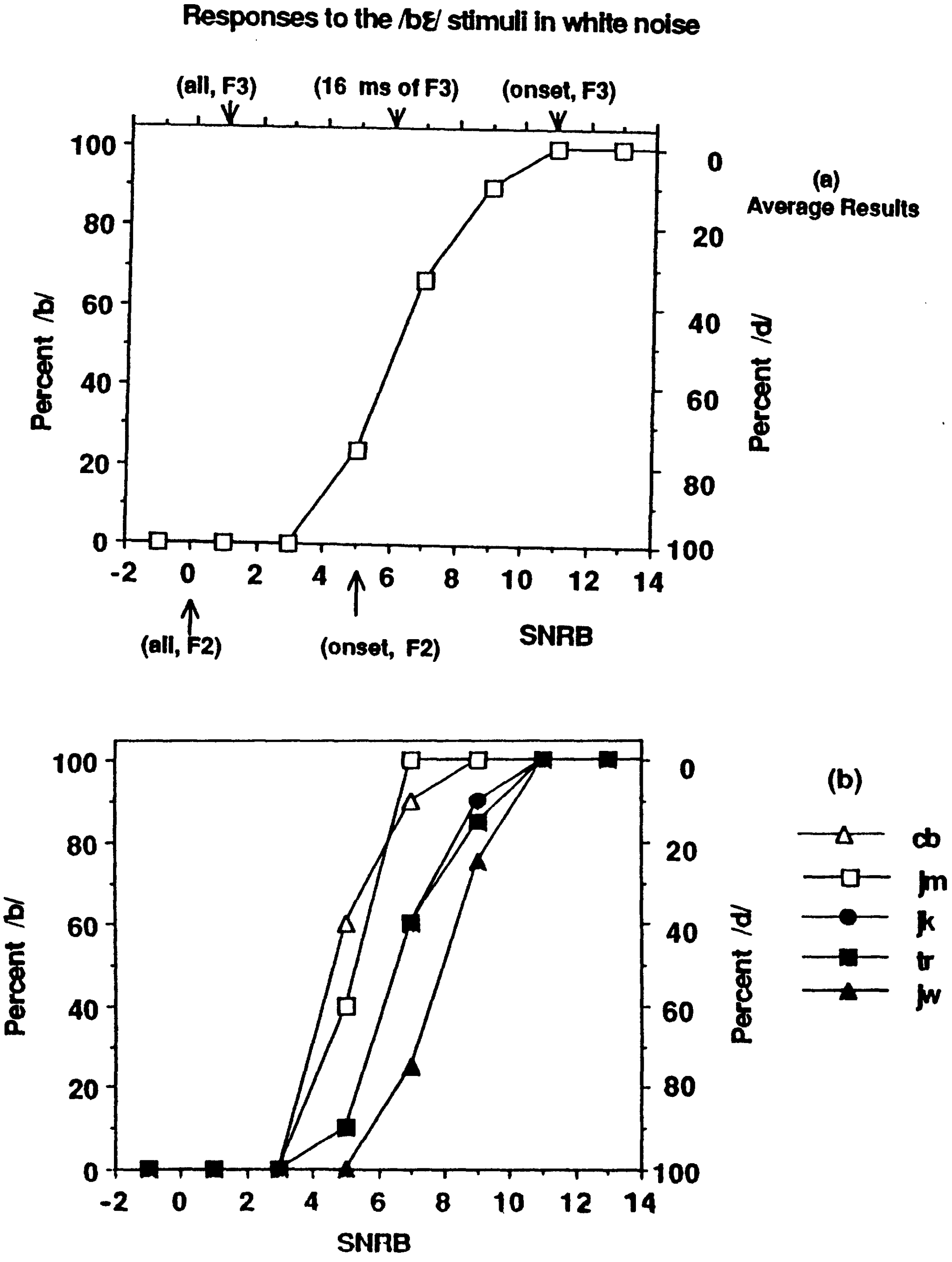


Figure 4.7: Plots of the subjects' responses to the /bɛ/ stimuli as a function of the signal-to-noise ratio in a critical band (*SNRB*) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the axes indicate the time intervals in which F2 and F3 were predicted to be masked in selected stimuli. For example, an *SNRB* of 6 dB was predicted to result in the masking of 16 ms of F3 transition and in no masking of F2. F1 was not predicted to be masked in any of the stimuli. The masker was white noise.

unanimously that the synthetic CV utterances sounded more natural in noise at high $SNRB$ than in quiet; this subjective judgement was especially true in the C/ɛ/ case. One explanation for this preference is that listeners expect to hear a burst at the onset of each syllable; in quiet, the absence of the burst is noticeable whereas in noise the listeners might be assuming that there *is* a burst but it is masked (hence, inaudible) and therefore they judge the stimuli to be more natural.

To examine further the perceptual role of F2 in signalling the place of articulation distinction for the labial and the alveolar stops, experiments in which F2 was selectively masked by a band-pass noise masker were conducted. The following section describes these experiments.

4.4 Experimental Results: Band-pass Noise Masker

In these experiments, synthetic CV utterances were mixed with a band-pass noise masker and presented to subjects in identification tests. The masker was centered at the F2 region.

4.4.1 The C/ɑ/ Case

As mentioned earlier, the two synthetic C/ɑ/ utterances differed in the F2 trajectory. Consequently, the amplitudes of all formant peaks were different for the two consonants (Equation 2.1). Amplitude differences between the onset and the steady-state part of the vowel for the synthetic /bɑ/ were 10 and 11 dB, for F3 and F4, respectively, whereas these differences were 0 and 2 dB, for the synthetic /dɑ/.

In this experiment, F2 was predicted to be completely masked by a band-pass noise masker such that the only acoustic cues in the formant frequencies that could signal the place of articulation for the two consonants were the amplitude differences in the F3 and F4 regions.

Results of the experiments are summarized in Figures 4.8, 4.9, and 4.10. The overall noise level chosen by subjects (CB, JM, JK, SSH) was 70 dB SPL, and for subject JW, it was 77 dB SPL. Figures 4.8 and 4.9 show experimental results for the /dɑ/ and /bɑ/ stimuli, respectively, at an overall noise level of 70 dB SPL. Average results are shown in part (a) and individual results are shown in part (b). Figures 4.10a and 4.10b show results for subject JW for the /bɑ/ and /dɑ/ stimuli, respectively, at an overall noise level of 77 dB SPL. Arrows on the bottom axes indicate the signal level at which the entire F2 trajectory is predicted to be masked, whereas arrows on the top axes indicate the signal level at which F3 and F4 peaks are predicted to be masked. Predicting the masking of F3 and F4 is based on above-band masking calculations whereas masking predictions of F2 are based on within-band masking calculations (Section 4.1).

The results show that even with the F2 peak completely masked, subjects were still able to identify the consonants correctly. For example, percent correct identification for /b/ and /d/ at an $SNRB$ of $-$ 10 dB was at least 90% (Figures 4.8, 4.9, and 4.10). Recall that F2 peak is predicted to be masked at an $SNRB$ of 0 dB. It was not until both F3 and F4 were predicted to be completely masked that the confusions between the two consonants occurred. These results suggest that the differences in the relative amplitudes of F3 and F4 between the onset of the vowel and that at the steady state part could be used as a cue for place information. Confusions in the responses to the /dɑ/ stimuli occurred at a lower $SNRB$ than for the /bɑ/ stimuli. This difference is probably due to the fact that F3 and F4 were higher in amplitude for /d/ than they were for /b/.

The use of amplitude differences in F3 and F4 in C/ɑ/ utterances as place cues

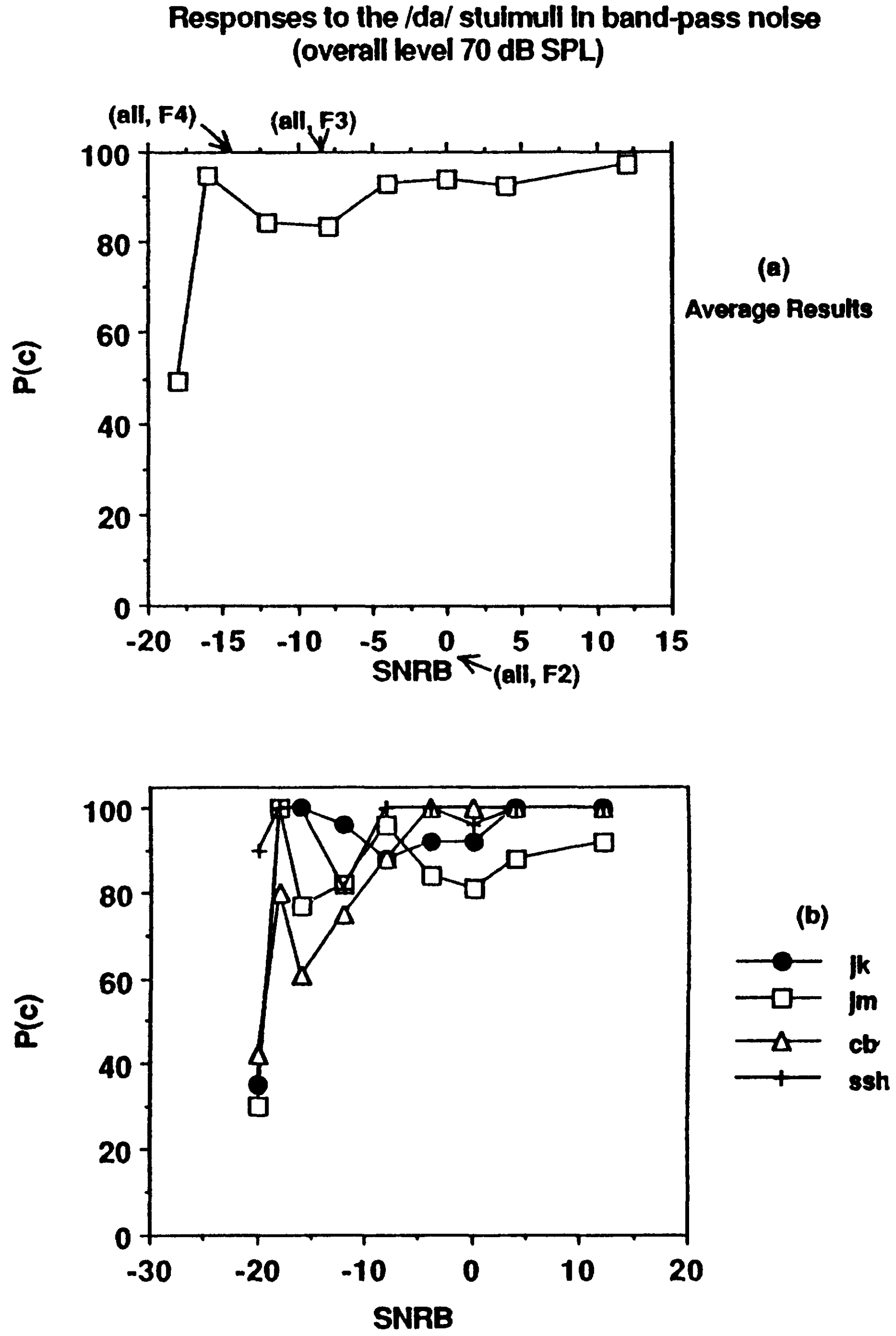


Figure 4.8: Plots of the subjects' responses to the /da/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the axes show the time interval in which a formant was masked as predicted from models of within-band and out-of-band masking. The band-pass masker was 1200 Hz in width, centered at F2, and its overall level was 70 dB SPL.

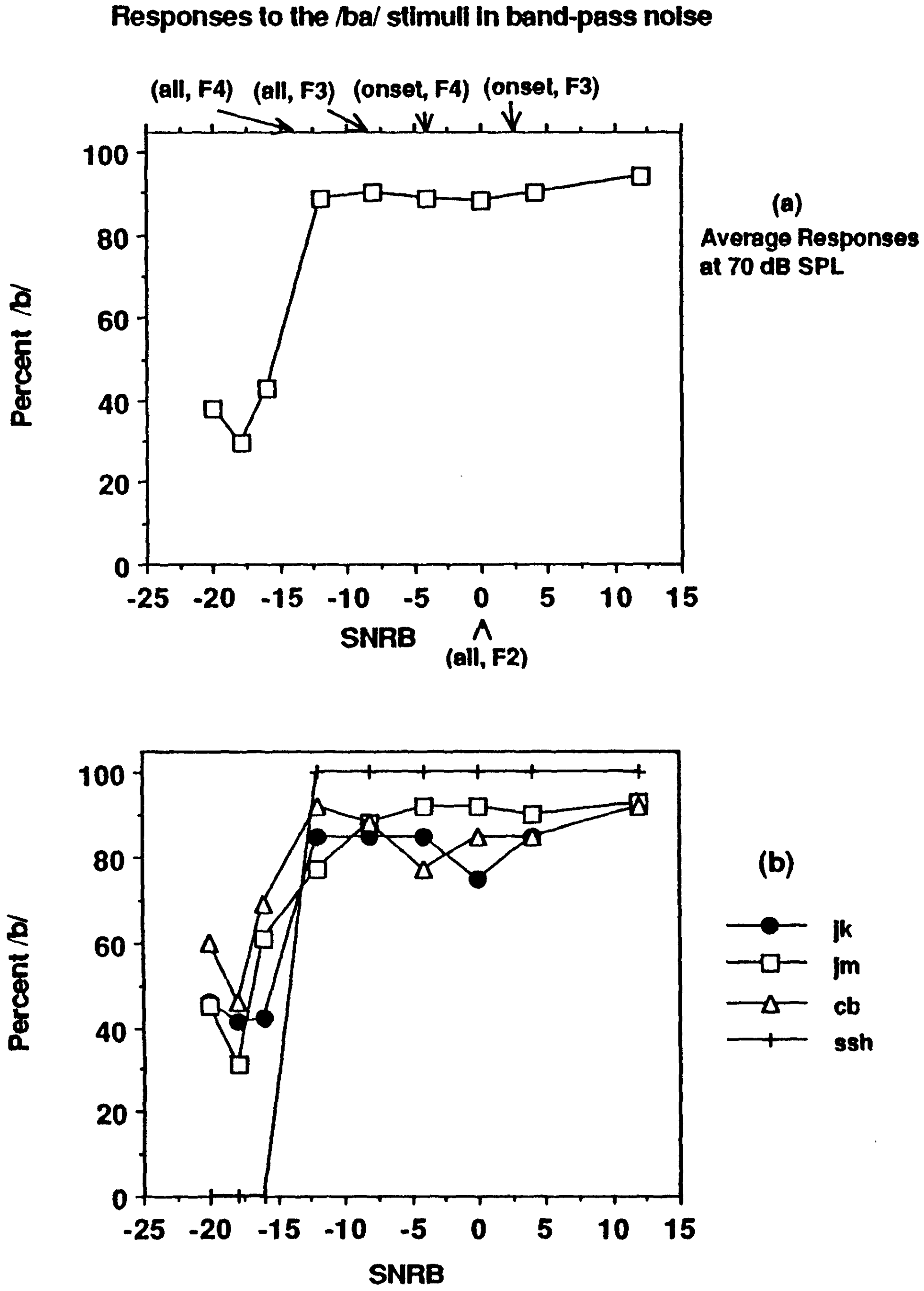


Figure 4.9: Plots of the subjects' responses to the /ba/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the axes show the time interval in which a formant was predicted to be masked. The band-pass masker was 1200 Hz, centered at F2, and its overall level was 70 dB SPL.

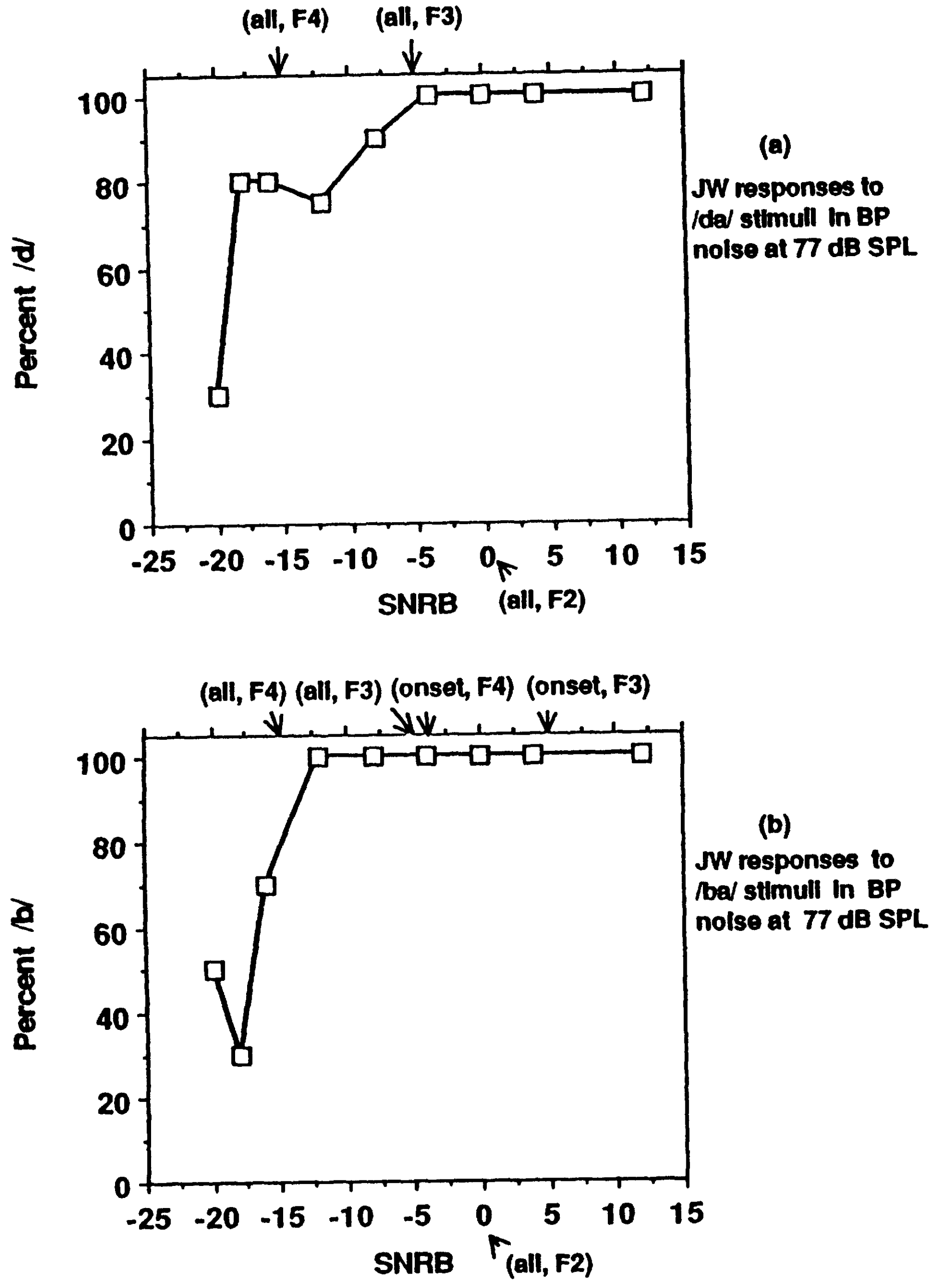


Figure 4.10: Plots of the responses for subject JW to (a) /da/ stimuli and (b) /ba/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Arrows on the axes show the time interval in which a formant was predicted to be masked. The band-pass masker was 1200 Hz, centered at F2, and its overall level was 77 dB SPL.

might be due to either the listeners' ability to use all relevant formant cues in the time-varying spectra to identify the consonants or that the listeners were extrapolating from the amplitude differences that a /d/ burst or a /b/ burst was present (since the bursts for these consonants differ in the energy concentrations in the F3 and F4 regions, Figure 2.2).

4.4.2 The C/ɛ/ Case

Experiment I

The results of the experiments with C/ɛ/ syllables in white noise (Section 4.3.2) showed that transitions of both F2 and F3 signal place of articulation for the two consonants /b/ and /d/.

In the experiment described in this section, F2 was masked by a band-pass noise masker centered around the F2 region. For low SNRB, the higher formant frequencies were predicted to be masked due to above-band masking. The goal of this experiment was to find out if the listeners could rely on cues other than F2 to distinguish between /b/ and /d/ in noise.

All /dɛ/ stimuli were identified correctly in this experiment. That is, /d/ responses were made independent of the signal level. The responses for the /bɛ/ stimuli are summarized in Figures 4.11 and 4.12 for overall noise levels of 70 and 77 dB SPL, respectively. Average results are shown in part (a) and individual results are shown in part (b) of each figure. Arrows on the top and bottom axes in part (a) indicate the time interval in which a formant peak is masked.

The results at both noise levels were similar to those in the white-noise case: when the formant transition was masked (in this case it was F3 transition) 'flat' trajectories were perceived and the responses shifted to /d/.

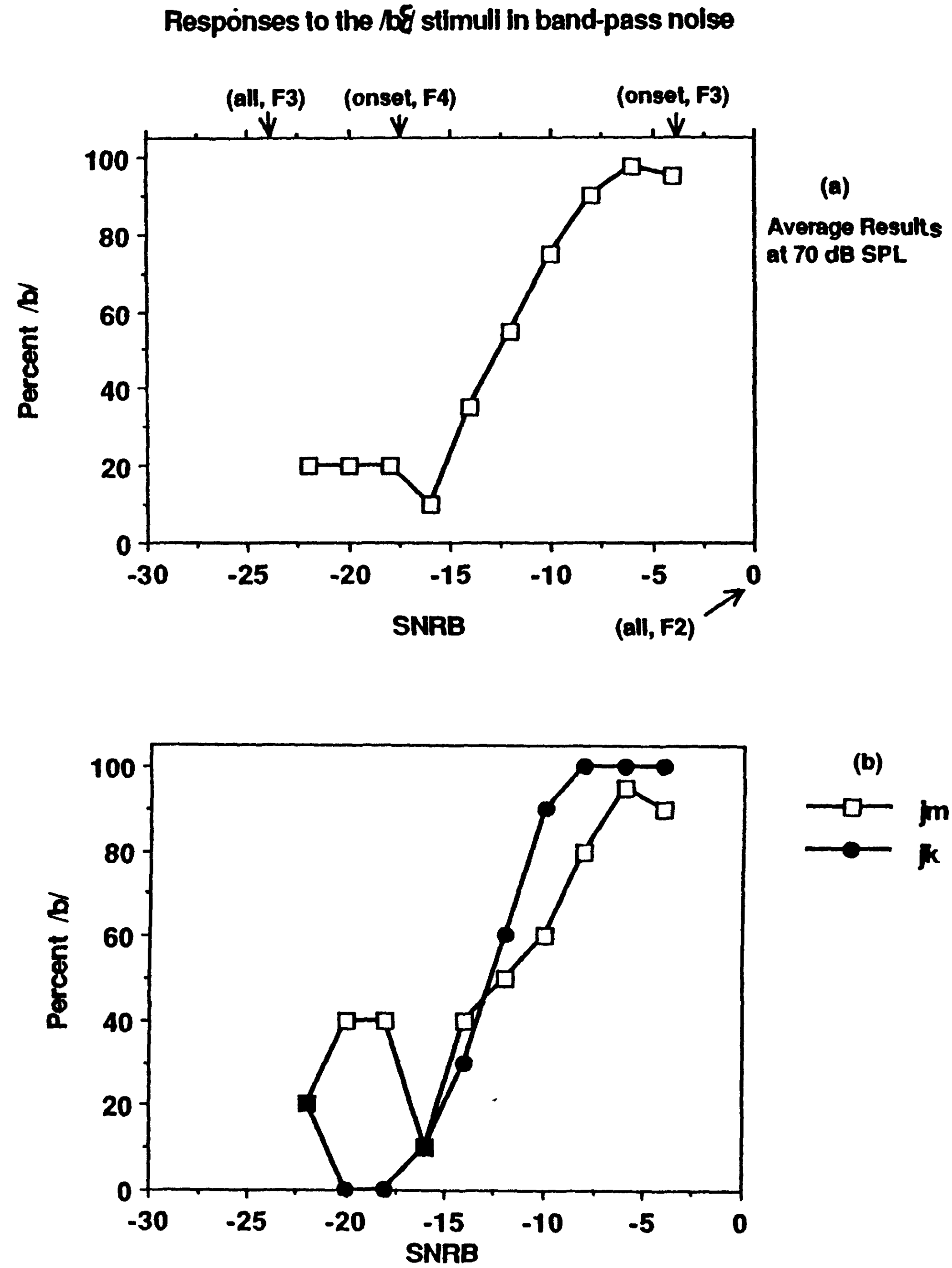


Figure 4.11: Plots of the subjects' responses to the /bɛ/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the axes show the time interval in which a formant was predicted to be masked from models of within-band and out-of-band masking. The band-pass masker was 1000 Hz in width, centered at F2, and its overall level was 70 dB SPL.

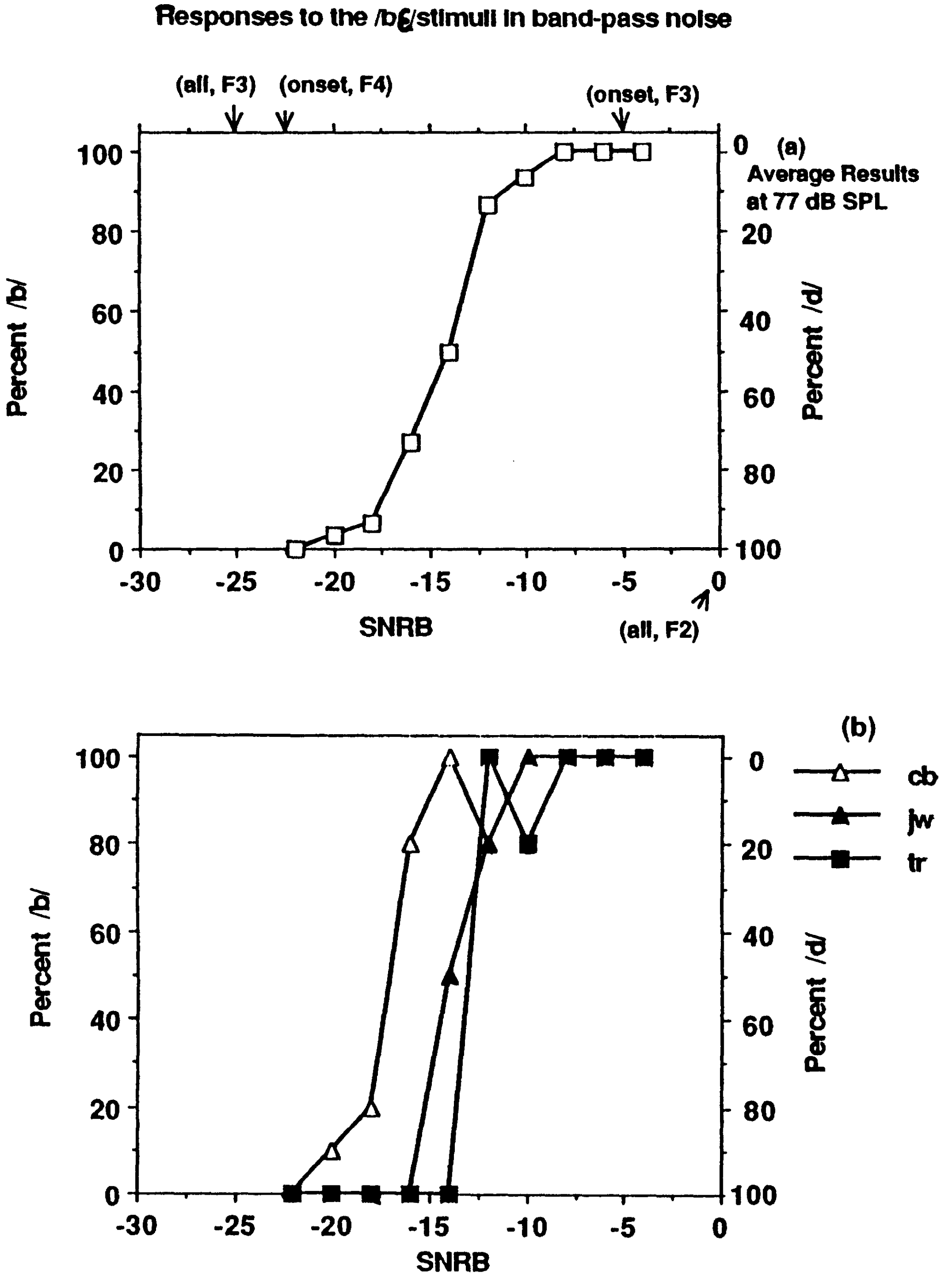


Figure 4.12: Plots of the subjects' responses to the /bɛ/ stimuli as a function of the signal-to-noise ratio in a critical band (*SNRB*) centered at F2 in the steady state part of the vowel. Average results are shown in (a), and individual results are shown in (b). Arrows on the axes in (a) show the time interval in which a formant was predicted to be masked. The band-pass masker was 1000 Hz in width, centered at F2, and its overall level was 77 dB SPL.

It seemed that the subjects could rely entirely on the F3 transition to identify the consonant, since /b/ was correctly identified when the entire F2 trajectory was masked (at an *SNRB* of 0 dB). This result led us to another question: would the responses to the C/ε/ stimuli change if the onset value of F3 was at a value intermediate between the values chosen for /bε/ and /dε/? To answer this question, the following experiment was conducted.

Experiment II

In this experiment, the synthetic utterances used were similar to those described in the previous sections except for F3: F3 was the same for both /b/ and /d/ and was rising from a value of 2500 Hz at the onset of the vowel to a value of 2700 Hz at the steady state part of the vowel. This choice of the F3 onset value was halfway between the onset values used previously, which were 2300 and 2700 Hz for /bε/ and /dε/, respectively. F2 was different, as it was in the previous experiments, for the two consonants: rising for /b/ and 'flat' for /d/. The utterances used in this experiment will be referred to as /b̃ε/ and /d̃ε/. Spectrograms of these utterances are shown in Figure 4.13. Two phonetically trained subjects identified both syllables in quiet as /bε/; however, the subjects reported that the C̃/ε/ utterances sounded less natural than the C/ε/ utterances used in the previous experiments. Their judgements were presumably due to the relatively high F3 onset value for /b/.

Results of the experiment show that the perception of both /C̃ε/ utterances shifted to /d/ in a systematic way: when the F3 transition is masked, /d/ is perceived. Figure 4.14 shows average results, across three subjects, for the /C̃ε/ stimuli at overall noise levels of 70 and 77 dB SPL. Recall that a band-pass noise masker centered around the F2 region was used in this experiment and that an *SNRB* of 0 dB predicts masking of

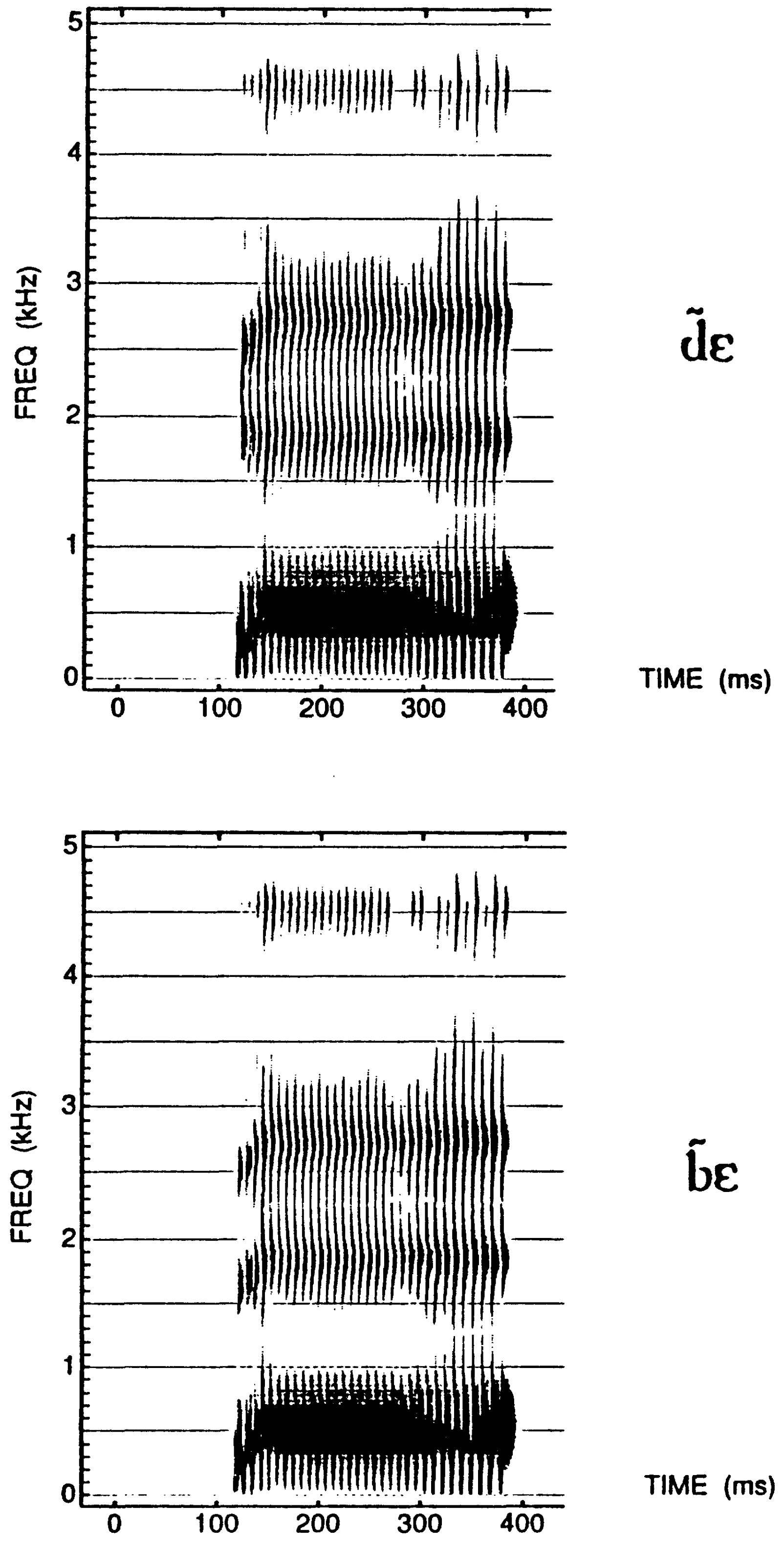


Figure 4.13: Spectrograms of the synthetic /d̃ɛ/ and /b̃ɛ/ utterances. The spectrograms were computed with a 6.4 ms Hamming window.

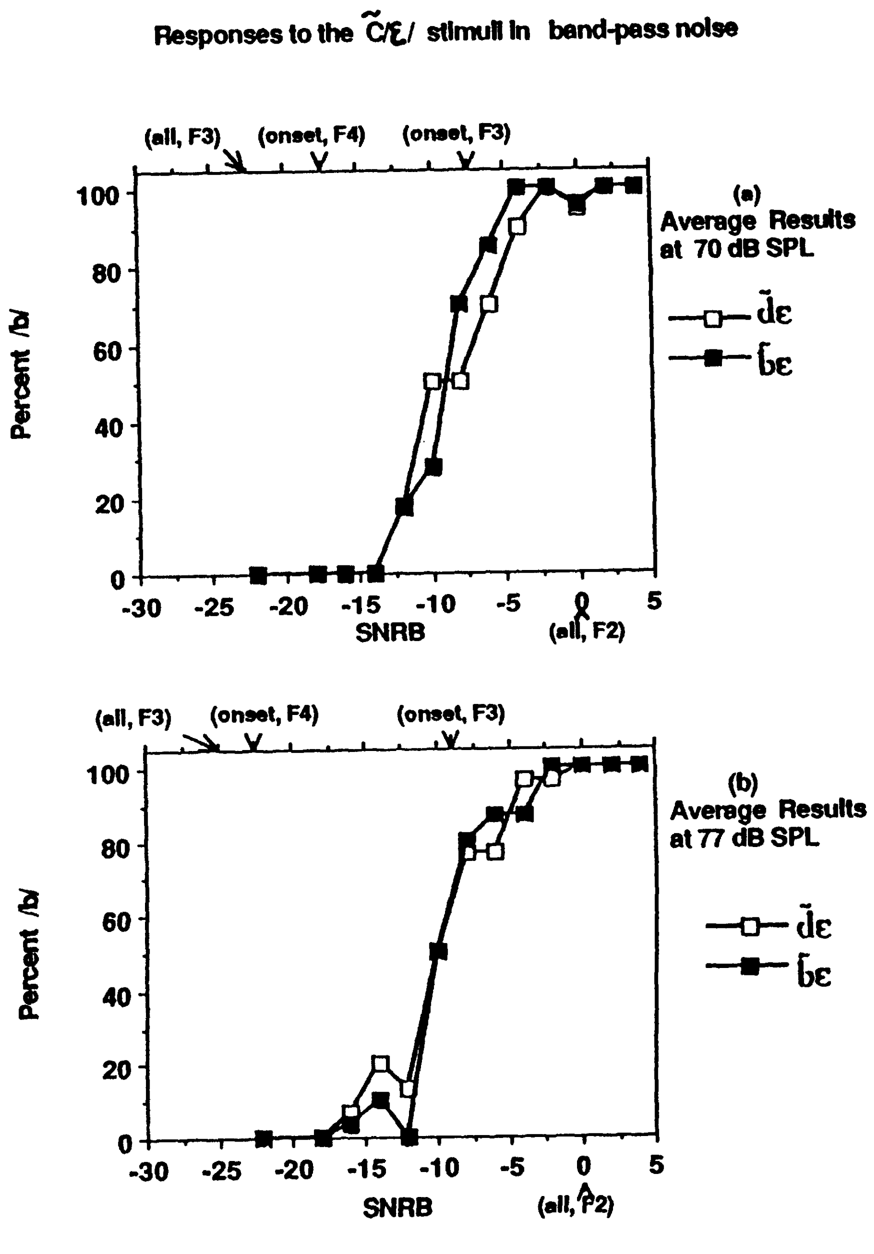


Figure 4.14: Plots of the subjects' responses to the /b̃ε/ and /d̃ε/ stimuli as a function of the signal-to-noise ratio in a critical band ($SNRB$) centered at F2 in the steady state part of the vowel. Part (a) shows average results at an overall noise level of 70. dB SPL and part (b) shows average results at 77 dB SPL. Arrows on the axes show the time interval in which a formant was predicted to be masked. The band-pass masker was 1000 Hz in width and was centered at F2.

the entire F2 trajectory.

The results here indicate, yet again, that the shape of the F3 trajectory was used to identify the consonants, with a 'flat' trajectory being perceived as /d/. This finding is in agreement with the results of Experiment I.

4.5 Summary and discussion

In this chapter, we used the metric developed in Chapter 3 to predict masking of formant peaks by a noise masker for analyzing results of identification tasks of synthetic CV utterances in noise. The consonant was either /b/ or /d/, the vowel /ɑ/ or /ɛ/, and the masker, white noise or band pass noise.

Results of the experiments are summarized as follows:

- The shape of the F2 trajectory was a sufficient cue for signalling the place of articulation for the consonants in the C/ɑ/ case. When most of the F2 transition was masked in the /dɑ/ stimuli, resulting in a remaining trajectory that was almost 'flat', /b/ was perceived. There was one exception (subject JK) where the subject identified the consonant as /b/ only if there was a rise in the F2 trajectory.

- When F2 was selectively masked in the C/ɑ/ stimuli, using a band-pass masker centered at the F2 region, then differences in the amplitudes of F3 and F4 between the onset and the steady state part of the vowel appeared to be used as perceptual cues for place information.

- In the context of the vowel /ɛ/, the shape of the trajectories of F2 and F3 was an important perceptual cue to signal the place of articulation for the consonants. When most of the F3 and F2 trajectories were masked, the consonant was perceived as /d/. When F2 was selectively masked, the subjects relied on the shape of the F3 trajectory to identify the consonants. This result was tested for two different onset values of F3.

Overall, it appears that, when all possible cues are taken into account, it is possible to predict when the /b-d/ distinction will be masked based on a simple psychoacoustical model.

How do these results compare with the classic data of Miller and Nicely (hereafter, referred to as MN) data? A direct comparison is difficult because of differences in the control strategy, type of task, and stimuli between the two studies. The task in our experiments was a two-alternative-forced choice task, and it was controlled such that half of the stimuli sounded /b/-like and the other half, /d/-like; no such control strategy was used in the MN study and their task was multiple choice in which the listeners could answer any one of the 16 consonants chosen for their study. Also, MN experiments were conducted using natural utterances spoken by five female talkers, whereas our study used synthetic utterances based on the speech of a male talker. These two differences would yield greater variability in the MN data than in ours. Nevertheless, there were similar tendencies in the subjects' responses. The comparison will be restricted to C/ɑ/ utterances in white noise since it was the only condition common to both studies, but first we have to translate the term $SNRB$ to the overall SNR.

F2 was calibrated with a tone at the same frequency and level using a Ballantine voltmeter (MN study used a VU meter for signal measurements). The signal level, taken to be the peak value in the C/ɑ/ syllable, corresponded to a value which was 6 dB above the F2 level. The overall noise level at an $SNRB$ of 0 dB is equivalent to the level of F2 plus 18 dB (Equation 3.8). Hence, at an $SNRB$ of 0 dB, the signal-to-noise ratio (SNR) is -12 dB.

Our experimental results for the C/ɑ/ utterances in white noise were shown for $SNRB$ values ranging from 0 to 10 dB (Figure 4.5), or the SNR range was from -12 to -2 dB. The average percent correct score for the /dɑ/ stimuli in our study was 0, 50, and 100, respectively, for SNR values of -12, -8, and -2 dB, respectively. The /bɑ/ responses were 100% at all SNR values. MN data show that the average percent correct scores for the place of identification for /b/ were 57, 78, and 88, and for /d/, 34, 46, and 76, respectively, at SNR values of -12, -6, and 0 dB.[4] Hence, subject responses in the two studies showed that the place of articulation feature /+*labial*/ was

[4] Correct answers for place of articulation for /b/ in the MN data were considered to be: /p, b, f, v, m/ and for /d/, they were /t, d, s, z, n/.

preserved at an SNR lower than that needed for the feature /+*alveolar*/. However, it is unclear why the labial responses in MN study did not stay at 100% as it did in this study. One explanation is that the subjects had four choices of place of articulation, instead of two in our study, which included, besides *labial* and *alveolar*, *velar* and *dental* places which would result in differences in subjects' responses.

Chapter 5

Conclusion

5.1 Summary

This study represents an effort to integrate knowledge of the acoustic properties which signal phonetic contrasts of speech sounds with that of auditory masking theory and to use the integrated knowledge to analyze and predict perceptual confusions of speech sounds in noise. The focus of the study was on perceptual confusions of the stop consonants /b,d/ in CV syllables where the vowel was either /ɑ/ or /ɛ/. The feature specifications for the two consonants differ only in the place of articulation features: /b/ is + *labial* and /d/ is + *alveolar*. The difference in the place of articulation for the two consonants is manifested acoustically mainly by differences in the formant trajectories from the consonant to the adjacent vowel and by differences in the spectral shape of the bursts at the release of the consonants. Our study focused on the perceptual role of the formant trajectories in signalling the place of articulation for these consonants when these consonants are heard in noise.

A metric was developed to predict the level of noise needed to mask a formant peak. The metric was based on a combination of theoretical and empirical results. Two kinds of masking were studied: within-band masking (where the formant is within the bandwidth of the masker) and above-band masking (where the formant is above the upper cut-off frequency of the masker). Results of auditory masking theory, which was established primarily for pure tones, were used successfully to predict within-band masking of formant peaks. The applicability of the results of masking theory to formant peaks was tested by conducting a series of discrimination and detection experiments with synthetic, steady-state vowels.

While there is a well-defined theory to account for within-band masking, no such theory exists for above-band masking, and, in particular, for upward spread of masking. Hence, an empirical algorithm was developed to account for the experimental results. The algorithm was a modification of the ANSI procedures (S3.5-1969) which predict general masking patterns associated with upward spread of masking.

Identification experiments were then conducted using as stimuli synthetic CV utterances mixed with noise. The consonant was either /b/ or /d/, and the vowel, /ɑ/ or /ɛ/. In the context of the vowel /ɑ/, the two synthetic syllables differed in the F2 trajectory: the F2 trajectory was falling for /dɑ/ and relatively flat for /bɑ/. In the C/ɛ/ context, both F2 and F3 trajectories were different for the consonants: F2 and F3 were flat for /d/, whereas they were rising for /b/. The spectral prominences associated with F2 and F3 always have a lower amplitude during the transitions from the consonant than in the steady-state vowel, so that it is possible, using a steady-state noise, to mask portions of a formant transition without masking the formant peak in the vowel. Subjects' responses were analyzed with the perceptual metric described above. Results show that the shape of the F2 trajectory in the C/ɑ/ case, and the shape of the F2 and F3 trajectories in the C/ɛ/ are sufficient cues for identifying place information for these consonants. In addition, it was found that masking a rising or falling formant trajectory results in the perception of a 'flat' trajectory which leads to the perception of /b/ preceding /ɑ/ and the perception of /d/ preceeding /ɛ/.

It was also found that when only the F2 trajectory is masked, achieved by selectively masking F2 with a band-pass noise masker centered at the F2 region, then amplitude differences in the F3 and F4 regions could be used as cues for place information in the C/ɑ/ case even though the trajectories of these higher formants did not differ for the two consonants.

5.2 Contributions, Limitations and Implications

The main contributions and limitations of this study can be summarized as follows:

- Detection and discrimination experiments with synthetic, steady-state vowels showed that results of masking theory can be applied to predict within-band masking of spectral prominences of complex signals such as vowels. The vowels examined here were synthetic /ɑ/ and /ɛ/ vowels. In order to validate the applicability of the results to formant peaks in vowels with different formant patterns and different fundamental frequencies, further experiments are needed.

- Results of above-band masking experiments showed that the existing procedures (ANSI (S3.5-1969), and Ludvigsen (1985)) do not predict accurately masked thresholds of tones at frequencies higher than the upper edge of the masker. A modified ANSI algorithm was developed to account for the experimental results. However, the algorithm is appropriate for the particular noise masker used in the experiments and it is necessary to determine how well the algorithm predicts masking caused by noise maskers with different spectral shapes and bandwidths.

- Experimental results with CV syllables showed that masking a falling or rising trajectory leads to the perception of 'flat' trajectories and that amplitude differences in F3 and F4 in the C/ɑ/ case can be used as cues for place information. A psychoacoustic model was used successfully to predict the masking of acoustic attributes which led to confusions in phonetic distinctions. The limitation of the experiments is that the synthetic utterances for the two consonants varied only in few acoustic dimensions (only in F2 or in F2 and F3 trajectories) while the other synthesis parameters were kept fixed. Further experiments in which the utterances vary along several dimensions, such as formant transition durations or F1 onset values, are needed.

What are the implications of the results of this study for the way we think about speech perception? The results show that spectral changes between the onset and the steady-state of the vowel are sufficient cues for consonant identification in noise. These spectral changes are demonstrated by formant motions or by differences in formant amplitudes. Consider, for example, the case where all formant frequencies were masked except for F1. In that case, listeners defaulted to the consonant which demonstrated the least spectral changes with respect to the vowel: /b/ in the C/ɑ/ case and /d/ in the C/ɛ/ case.

Our study agrees with the results of Sussman et al. (1991) on the importance of the relational property between the onset and mid-point vowel values of F2 in identifying place for stops; the F2 values of our synthetic stimuli fall right on the locus-equation regression lines shown in Sussman et al. study. Furthermore, our results verify the speculation in that study that F3 might be perceptually important in the context of front vowels. However, if formant frequency motions are predicted to be masked then cues in other parts of the spectrum (such as amplitude differences in the F3 and F4 regions in the C/ɑ/ case) can be used to identify place. The latter result is more in line with the theory of Lahiri et al. (1984) on the importance of the distribution of spectral energy between burst release (vowel onset in our case) and a short period following onset of voicing (three glottal pulses in Lahiri et al. study versus the formant transition time in ours) as an invariant acoustic cue.

The analytical approach presented in our study can be used to examine the perceptual relevance of all acoustic cues which signal phonetic contrasts. Of particular interest are 'unsuspected' secondary cues in the speech signal. If the perceptual relevance of all possible acoustic manifestations of phonetic features are understood then a prediction, for example, of word identification in noise could be made. Such a prediction is based on the assumption that words are stored in memory as arrays of features, and that inability to identify particular features due to noise masking can lead to confusions in word identification. This quantitative approach can ultimately be used as a measure for speech intelligibility in noise and it is different from other methods that predict overall percent correct scores for words or that determine confusion matrices experimentally.

5.3 Future Work

The perceptual metric used in this study was implemented by calculating the signal-to-noise ratio in a critical band around each formant frequency ($SNRB$). The $SNRB$ for each formant was calculated in 25.6 ms frames and the threshold of a formant peak was assumed to be similar to the threshold of a tone at the same frequency. However, tone thresholds were measured for durations greater than 200 ms. The effects of duration on the masked thresholds of tones are not entirely understood. The masked thresholds of pure tones have been shown to increase for durations less than 200 ms (Zwicker, 1965). No such effect has been observed, however, for complex tones, such as those studied in profile-analysis experiments, in which components are widely spaced (Dai and Green, 1991). Further psychophysical experiments with complex stimuli varying in duration are needed to determine the durational effects on the masked thresholds of tonal components. When results on durational effects are more conclusive, the perceptual metric should be refined to incorporate these effects.

The study also showed that the phenomenon of upward spread of masking is not well understood. Several factors seem to affect upward spread of masking: bandwidth, cutoff frequency, overall level and skirt characteristics of the masker. Studying each of these factors in a systematic way would lead to a method which could account accurately for masked thresholds.

Experiments examining the perceptual role of other acoustic cues in noise, such as the spectral shape of stop bursts, should be conducted. It is worth mentioning here that pilot experiments in which /bɑ/ and /dɑ/ utterances with bursts were synthesized and presented to subjects in white noise showed no effects of the burst on the identification of these consonants. This result is not surprising since the stop burst is masked at higher SNR values than those required to mask the F2 peak in the adjacent vowel. The stop burst would probably be of perceptual importance in situations where the F2 trajectory is selectively masked. A natural extension to our study would be conducting similar experiments with naturally-produced utterances. We have conducted pilot experiments with natural C/ɑ/ utterances and found confusions similar to those found in the synthetic case. A cross-language study examining stops with various places of articulation would reveal whether perceptual strategies are language specific, or universal.

Our study was restricted to examining place features in syllable-initial stop consonants. However, the methodology proposed in the study can be adopted to examine and analyze perceptual confusions of other acoustic features in noise and in different contexts. Confusions in the manner of articulation features (for example, /b/ and /m/) or confusions in voicing (/b/ versus /p/) are but a few examples of the types of confusions which could be studied with the proposed methodology. Other types of noise maskers should also be studied. Of particular interest is a noise masker whose spectrum resembles a long-term average speech spectrum (French and Steinberg, 1947); this kind of masker is the most common noise background encountered in normal communication.

Finally, similar experiments could be conducted with hearing-impaired subjects to find out differences in listening strategies, if any, between those used by subjects with hearing impairments and those used by normal-hearing subjects in noise. The hope would be that, with adequate understanding of psychoacoustic capabilities and knowledge of acoustic properties necessary for identification, speech intelligibility for the hearing impaired can be predicted.

Bibliography

Alwan, A. (1986). "Acoustic and perceptual correlates of pharyngeal and uvular consonants," unpublished S.M. thesis, MIT.

ANSI S3.5-1969, "Methods for the calculation of the articulation index," (American National Standards Institute, NY).

ANSI S3.6-1969, "Standard specifications for Audiometers," (American National Standards Institute, NY).

Bilger, R.C., and Hirsh, I.J. (1956). "Masking of tones by bands of noise," *J. Acoust. Soc. Am.*, 28, 623-630.

Blumstein, S., and Stevens, K.N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Am.*, 66, 1001-1017.

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.

Brandenburg, K., and Johnston, J. D. (1990). "High-quality wideband audio coding," *J. Acoust. Soc. Am.*, 87, (Supp. 1), E6.

Carlson, R., Granström, B., and Fant, G. (1970). "Some studies concerning perception of isolated vowels," STL-QPSR, No. 2-3, 19-35.

Carlson, R., Fant, G., and Granström, B. (1975). "Two-Formant Models, Pitch, and Vowel Perception," in *Auditory Analysis and Perception of Speech*, edited by G. Fant, and M.A. Tatham, Academic Press, London, 55-82.

Carroll, J. D., and Wish, M. (1974). "Models and methods for three-way multidimensional scaling," in *Contemporary Developments in Mathematical Psychology*, edited by D.H. Krantz, R.C. Atkinson, R.D. Luce, and P. Suppes. Freeman, San Francisco, Vol. II, 57-105.

Carter, N.L., and Kryter, K.D. (1962). "Masking of pure tones and speech," *J. Auditory Research*, 2, 66-98.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English.* Harper and Row, New York.

Dai, H., and Green D. (1991)."Spectral-shape discrimination as a function of stimulus duration," *J. Acoust. Soc. Am.*, 90 (Supp. 2), 2PP2.

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.

Fant, G. (1960). *Acoustic Theory of Speech Production.* 'S-Gravenhage: Mouton.

Farrar, C.L., Reed, C.M., Ito, Y., Durlach, N.I., Delhorne, L.A., Zurek, P.M., and Braida, L. (1987). "Spectral-shape discrimination. I. Results from normal-hearing listeners for stationary broadband noises," *J. Acoust. Soc. Am.*, 81, 1085-1092.

Fletcher, H. (1940). "Auditory Patterns," *Review of Modern Physics*, 12, 47-65.

French, N.R., and Steinberg, J.C. (1947). "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, 19, 90-119.

Gagné, J.P. (1988). "Excess masking among listeners with a sensorineural hearing loss," *J. Acoust. Soc. Am.*, 83, 2311-2321.

Green, D., and Swets, J. (1966). *Signal Detection Theory and Psychophysics.* John Wiley and Sons Inc, New York.

Greenwood, D.D. (1961). "Critical bandwidth and the frequency coordinates of the basilar membrane," *J. Acoust. Soc. Am.*, 33, 1344-1356.

Hawkins, J.E., and Stevens, S.S. (1950). "The masking of pure tones and of speech by white noise," *J. Acoust. Soc. Am.*, 22, 6-13.

Henke, W.L. (1989). *MITSYN Languages, Language Reference Manual*, V7.0, WLH, 133 Bright Rd. Belmont, MA 02178.

Humes, L.E., Dirks, D.D., Bell, T.S., Ahlstrom, C., and Kincaid, G.E. (1986). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," *J. Speech Hear. Res.*, 29, 447-462.

Jackobson, R., Fant, G., and Halle, M. (1963). *Preliminaries to Speech Analysis.* MIT Press, Cambridge, Mass.

Johnson, S.C. (1967). "Hierarchical clustering schemes," *Psychometrika*, 32, 241-254.

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

Kewley-Port, D., Pisoni, D., and Studdert-Kennedy, M. (1983). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Am.*, 73 , 1779-1793.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, 67, 971-995.

Klatt, D. H. (1984). *M.I.T. Speechvax User's Guide.*

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.

Klatt, D. H. (Book in preparation).

Lahiri, A., Gewirth, L., and Blumstein, S. (1984). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," *J. Acoust. Soc. Am.*, 76, 391-404.

Ludvigsen, C. (1985). "Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners" *J. Acoust. Soc. Am.*, 78, 1271-1280.

Miller, G.A., and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, 27, 338-352.

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

McClellan, J.H., Parks, T.W., and Rabiner, L.R. (1973). "A computer program for designing optimum FIR linear phase digital filters," *IEEE Trans. Audio Electro.*, AU-21, 506-526.

Oppenheim, A.V., and Schafer, R.W. (1975). *Digital Signal Processing.* Prentice-Hall, New Jersey.

Rankovic, C.M., Freyman R., and Zurek, P.M. (to be published, Jan. 1992). "Potential benefits of adaptive frequency-gain characteristics for speech reception in noise," *J. Acoust. Soc. Am.*.

Reed, C. M., and Bilger, R. C. (1973). "A comparative study of S/N and E/N," *J. Acoust. Soc. Am.*, 53, 1039-1044.

Searle, C. L., Jacobson, J. Z., and Rayment, S. G. (1979). "Stop consonant discrimination based on human audition," *J. Acoust. Soc. Am.*, 65, 799-809.

Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979)."Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, 66, 1647-1652.

Shepard, R. (1972). "Psychological representation of speech sounds," in *Human Communication: A Unified View*, edited by E.E. David, Jr. and P.B. Denes. McGraw Hill, New York, 67-113.

Soli, S.D., and Arabie, P. (1979). "Auditory versus phonetic accounts of observed confusions between consonant phonemes," *J. Acoust. Soc. Am.*, 66, 46-59.

Soli, S.D., Arabie, P., and Carroll J.D. (1986). "Discrete representation of perceptual structure underlying consonant confusions," *J. Acoust. Soc. Am.*, 79, 826-837.

Sorin, C. (1987). "Psychophysical representation of stop consonant and temporal masking in speech," in *The psychophysics of speech perception*, edited by Schouten M. Martinus Nijhoff, Dordrecht, 241-249.

Stevens, K.N. (in preparation). *Acoustic Phonetics.*

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (1991). "An investigation of locus equations as a source of relational invariance for stop place categorization," *J. Acoust. Soc. Am.*, 90, 1309-1325.

Tanner, W.P. (1961). "Application of the theory of signal detectability to amplitude discrimination," *J. Acoust. Soc. Am.*, 33, 1233-1244.

Villchur, E., and Killion, M. (1975). "Probe-tube microphone assembly," *J. Acoust. Soc. Am.*, 57, 238-240.

Wang, M.D., and Bilger, R.C. (1973). "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Am.*, 54, 1248-1266.

Wish, M., and Carroll, J.D. (1974). "Application of individual differences scaling to studies of human perception and judgement," in *Handbook of Perception*, edited by E.C. Carterette and M.P. Friedman. Academic Press, New York, Vol. 2, 449-491.

Zurek, P.M., and Delhorne, L.A. (1987)."Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," *J. Acoust. Soc. Am.*, 82, 1548-1559.

Zwicker, E. (1954). "Die Verdeckung von Schmalbandgeräuschen durch Sinustöne," *Acustica*, 4, 415-420.

Zwicker, E. (1963). "Über die lautheit von ungedrosselten und gedrosselten challen," *Acustica*, 13, 194-211.

Zwicker, E. (1965). "Temporal effects in simultaneous masking by white-noise bursts," *J. Acoust. Soc. Am.*, 37, 653-663.

Zwicker, E., and Fastl, H. (1990). *Psychoacoustics: Facts and Models.* Springer-Verlag, Berlin.