1978

Predicting Consonant Confusions in Noise on the Basis of Acoustical Analyses

Judy Robin Dubno
The Graduate Center, City University of New York

PREDICTING CONSONANT CONFUSIONS IN NOISE ON THE BASIS OF ACOUSTICAL ANALYSES

by

JUDY ROBIN DUBNO

A dissertation submitted to the Graduate Faculty in Speech and Hearing Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1978

This manuscript has been read and accepted for the Graduate Faculty in Speech and Hearing Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

[signature]

Sept 22, 1978
Date

Harry Levitt
Chairman of Examining Committee

[signature]

Sept 22, 1978
Date

Irving Hochberg
Executive Officer

[signature]

Gerald A. Studebaker

[signature]

Irving Hochberg

[signature]

James M. Pickett

Examining Committee

The City University of New York

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the assistance of a number of people. My sincere appreciation is extended:

To Professor Harry Levitt, committee chairman, who has given his time, his knowledge, and most importantly, his encouragement from my very first days as a student. The rewarding research experiences he provided have inspired me to continue this area of study.

To Professor Gerald Studebaker, who has the insight and diligence of a researcher, and the infinite patience of a teacher, qualities which are indispensible.

To Professor Irving Hochberg, whose skillful leadership was instrumental in enriching the experience of my Doctoral studies at the Graduate School.

To members of the research staff of the Communication Sciences Laboratory, M. Jane Collins and Mary Joe Osberger, and computer analysts, Harvey Stromberg and Ronald Slosberg, for their technical expertise and moral support.

To my friends in Rooms 904 and 919, who make the Graduate Center the special place it is.

And to my family, who have supported my efforts in the past, and encouraged me to face the challenge of the future.

Thanks to all of you.

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

There are certain conditions under which one's ability to understand speech is reduced by the presence of an interfering sound. In some situations, a listener is capable of responding to one acoustic signal while ignoring another. However, the selective mechanism of the ear is not perfect. In order to understand why and how this interference takes place, it is necessary to examine the physical characteristics of both the signal and the interfering noise. One desired goal is to develop a theory or model to describe how the combination of signal and noise results in a reduction in the ear's ability to respond selectively.

When the acoustic stimulus that the listener is responding to is speech, this analysis becomes significantly more complicated. The speech signal is a complex acoustic stimulus having rapidly-changing temporal and spectral properties. Imbedded in this signal is the information needed by the listener to identify this stimulus as speech. In addition, many subtle acoustic cues are present which allow

the listener to differentiate one speech sound from another, and finally to understand what is said.

Many procedures have been developed to measure the amount of interference, or masking, produced by a particular noise. In the most simple case, a talker produces a speech sample (usually a list of syllables, words, or sentences) and listeners record their responses. The percentage of test items (i.e., syllables, words, sentences) understood is the articulation or discrimination score. The difference in the scores between the quiet and noise conditions can be interpreted as the amount of "masking" produced by that noise. The results obtained are dependent on the intensity of both the speech and the noise, the speech materials used, the characteristics of the speaker's voice, the equipment characteristics, and the listener's hearing acuity. Taking all variables into account, this method tells us something about how much interference is caused by a certain competing noise, but nothing about why or how this occurs.

A more revealing method involves the analysis of the incorrect responses made by the listeners under the quiet and noise conditions, rather than a simple comparison

of articulation scores. Furthermore, by restricting the listener to a small set of response alternatives, one can examine the errors made in terms of the available choices. Thus, a set of foils may be chosen which would allow the examiner to evaluate the specific characteristics in the speech signal which are most easily discriminated, and those which are most affected by interference.

Tests of the above type are known as closed-response-set tests. Typically, the characteristics of only one phoneme at a time are varied across foils. For example, in the set /ip, if, it, iθ, is, ik/, the discriminability of voiceless stops and fricatives following the vowel /i/ may be evaluated. In another set the same consonants may appear in the initial position, in another vowel context, etc. The data from such tests are usually reported in terms of a confusion matrix. In this table of target-response pairs, the rows represent the target utterances, and the responses to these phonemes appear in the columns. An analysis of the matrix may be performed to determine which phonemes are consistently misidentified, and the most common incorrect responses for each context in which the phoneme appears.

Confusion matrix analyses are also useful when listening conditions have been made more difficult by adding background noise or by selectively filtering the speech stimuli. The differential effects of these changes may be evaluated by comparing the perceptual confusions made in selected conditions with those made under a reference listening condition, such as speech in quiet. This type of analysis has been used extensively in recent years (Miller and Nicely, 1955; Wang and Bilger, 1973; among others). In both investigations, evaluation of perceptual features was, in large part, made up of measurements of the relative contribution and, therefore, relative importance of various features for distinguishing among speech sounds.

Analyses of confusion matrices obtained in studies on perception of English consonants show the features of voicing and nasality to be perceptually important, in that they are easily discriminable, independent of context and experimental condition (i.e., noise and/or filtering). The feature of place of articulation, on the other hand, appears to be weak, perceptually, in that there are frequent confusions involving this feature in studies of both masking and

filtering.

The earliest investigations of speech perception were geared toward the study of the relationships between the observed acoustical characteristics of speech and the perception of a particular speech sound or class of sounds. The results of these studies suggest that a great deal of the information necessary for the identification and discrimination of different speech sounds is acoustical in nature. Furthermore, we may begin to speculate as to what specific acoustical information (i.e., spectrum of the noise burst) is necessary, but not sufficient, for the perception of a specific perceptual feature (i.e., place of articulation). One may assume that if a certain perceptual error is made, some or all of the relevant acoustic information was not available to the listener. The lack of acoustic information may be a result of masking or the sound transmission characteristics of the listening environment. It is important to emphasize that most often there are multiple, overlapping cues for the identification of specific speech sounds. That is, no one bit of acoustic information sufficiently differentiates one phoneme from another. Therefore, if one acoustic

cue for place of articulation is lost, but another cue remains (i.e., second formant transition), it is likely that the listener will be forced to make his identification on the basis of the remaining acoustic information.

It is possible that the set of response alternatives includes an utterance which is similar to the target, in terms of their acoustic characteristics. It is likely that if acoustic cues for the target are unavailable, the resulting incorrect response would be that utterance with acoustic information which is most similar to the target.

On the basis of these assumptions, it seems possible that from selected acoustical measurements one should be able to predict listener responses to a set of target utterances under specific test conditions. Before such predictions can be made, the following preliminary steps should be taken. First, a finite set of utterances would be chosen as speech stimuli, and arranged in a closed-response-set format with appropriate response foils. In addition, environmental conditions (intensity and spectra of competing noise, earphone or sound-field listening, etc.) are chosen. Next, a set of acoustical characteristics of the speech

stimuli would be selected for analysis, and careful measurement procedures undertaken. The measurements are analyzed in conjunction with any interfering aspects of the specific listening conditions (i.e., speech-to-noise ratio, noise spectrum, earphone or speaker frequency-gain characteristics, etc.). Thus, for each target utterance, the acoustic information available to the listener could be mapped out.

The target acoustic characteristics are then compared to the comparable measurements of the response alternatives, as they appear in the reference (quiet) condition. The goal of this analysis is to find the set of acoustic characteristics which will account for the commonly-confused utterances, in addition to those pairs of utterances which are rarely confused. The key parameter would be one which has values which are very similar in pairs of utterances which are commonly confused, and very different in pairs which are rarely confused. After these key parameters have been identified, a similar analysis is made on the confusions resulting from listening conditions containing interfering noise. Isolated for analysis are those pairs of utterances which are rarely confused in the reference condition but are commonly con-

fused in noise. The acoustic parameters which differentiated these pairs in quiet are assumed to be unavailable due to the spectral and temporal characteristics of the interfering noise. Thus, confusion matrices may be predicted from measurements of the acoustic characteristics of both the speech signal and the interfering noise.

The purpose of the present investigation is to test these predictions by evaluating patterns of perceptual confusions made under difficult listening conditions in terms of selected acoustical parameters.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter has been divided into three sections. The first section contains a discussion of the work related to the acoustic characteristics of consonants and their perceptual significance. The next section is devoted to the factors affecting consonant and vowel confusions made by normal-hearing listeners, with special emphasis on the effects of noise on consonantal confusions. In addition, the more recent studies of confusions made by hearing-impaired individuals will be included. Finally, a discussion of spectral analysis is given as an introduction the the procedures used in the present investigation.

ACOUSTIC CUES FOR PERCEPTION OF CONSONANTS

Three groups of consonants, fricatives, plosives, and nasals have been studied extensively. Many of these investigations were concerned with the acoustic characteristics of these phonemes as they are used by the listener for identification and discrimination.

Fricatives

Acoustic cues for discriminating among the voiceless fricatives /f, θ, s, ʃ/ and their voiced cognates /v, ð, z, ʒ/ were studied by Harris (1958). Each phoneme was made up of a period of noise (frication) plus a vocalic segment consisting of formants and their appropriate transitions. The relative contributions of these portions were determined for CV type syllables. Except for the discrimination of /θ/ from /f/, the friction portion was the necessary and sufficient cue for identification of these segments. In other words, the transitions of the formants in the adjacent vowels contribute little to the identification of the consonant. For /θ/ and /f/, however, the correct judgment was dependent upon the vocalic portion.

Hughes and Halle (1956) obtained power-frequency

spectra of six fricatives, /f, v, s, z, ʃ, ʒ/, these being the voiced-voiceless pairs for each of three places of articulation (labial, dental, palatal). The measurements made on the spectra of the phonemes were described in terms of the frequency and intensity information sufficient for their correct identification. The results emphasized the effects of vowel context on the spectra, in that changes are noted which are dependent on the characteristics of the adjacent vowel sound.

Based on these data, Heinz and Stevens (1961) developed electrical analog approximations of 3 voiceless fricatives, /s, ʃ, f/. They consisted of a circuit characterized by one pole (resonance) and one zero (antiresonance), the frequency of the zero lying roughly one octave below the pole frequency. The center frequency and bandwidth of the pole and zero tended to vary from one fricative context to another, due to the changes in the vocal tract configuration as a result of the influence of the adjacent vowel. In addition, the spectra of the natural /ʃ/ indicated the need for an additional circuit for high-pass filtering below 3kHz. Similarly, the spectrum of /f/ required some addi-

tional low-frequency noise. Heinz and Stevens then generated synthetic fricatives from these analog circuits and presented them to subjects for identification. The model for each fricative (i.e., the frequency location and bandwidth of each pole and zero) is actually a linear electric circuit with the appropriate transfer function. The synthetic spectrum is the output of this circuit when it is excited by white noise. Increasing the resonant frequency resulted in a consistent shift in the responses from /ʃ/ to /tʃ / to /s/ to /f,θ/. The responses were not, however, affected by a change in bandwidth, or the addition of low-frequency noise to the fricatives with a higher resonant frequencies. Other cues (i.e., format transition of adjacent vowels and relative amplitude of the fricative and vowel) were then investigated. The results agreed with Harris (1958) in that the discrimination of /f/ and /θ/ was facilitated by the presence of the cues in the vowel portion, specifically the transition of the second format. On the other hand, /s/ and /ʃ/ responses were not affected by these cues but were dependent, to a large extent, on the frication, which for these phonemes is relatively intense. Furthermore, the starting frequency

(origin) of the second formant transition seemed to indicate the place of articulation of the fricative, as had been shown for plosives (Delattre, Liberman and Cooper, 1955). These cues appeared to be related to vocal tract length in that a low-frequency origin for the formant transition is identified with bilabial or labio-dental articulations, middle frequency origins with alveolar articulations, and high frequency origins with palatal points of articulation (Strevens, 1960).

A number of investigations have studied the acoustic cues for distinguishing the voiceless fricatives from their voiced cognates. The results of Denes (1955) suggest that shortening the duration of the frication noise in VC syllables produces a shift in the identification of the fricative from voiceless to voiced. The duration of the preceding vowel was also considered important, a finding which was later corroborated by Raphael (1972). For CV syllables, Cole and Cooper (1975) found the frication-duration cue to significantly outweigh the effect of vowel duration for voiced-voiceless distinctions in fricatives. Thus, for syllable-initial fricatives, frication-noise duration is a sufficient cue for voicing. However, for fricatives in the syllable-

final position, the voicing distinction is dependent upon the ratio of frication and vowel durations (Denes, 1955).

Plosives

As with the fricatives, investigations of plosive consonants have been concerned with determining cues which are important for identifying place of articulation, as well as cues for the voiced-voiceless distinction. The frequency characteristics of the noise burst following the plosive release appears to be an important cue for differentiating among /p, t, k/ or /b, d, g/ (Halle, Hughes and Radley, 1957). These phonemes are produced with different points of vocal tract occlusion (labial, alveolar, or velar) resulting in different lengths of vocal tract anterior to the source of the noise. Longer tube lengths (velar constrictions) should yield lower frequency resonances than shorter tube lengths (alveolar constrictions). The noise burst following a bilabial constriction may be affected by the resonance resulting from the rapid opening of the lips. As might be expected, these characteristics have been shown to be affected by the nature of the adjacent vowel. In a study by Cooper, Delattre, Liberman, Borst and Gerstman (1952), higher fre-

quency noise bursts, when paired with vowels, were heard consistently as /t/. However, lower frequency noise bursts were heard as either /p/ or /k/, depending upon the location along the frequency scale of the burst in relation to the second formant of the vowel. In addition, a single noise burst of 1440 Hz was perceived as /p/ when paired with /i/ or /u/, but as /k/ when paired with /ɑ/. These findings were interpreted as indicating a relationship between perception and the underlying articulatory movements. That is, the perception of the plosive in the CV syllable could be affected by the production of the following vowel; vowels with higher second formants are produced near the front of the mouth, and vowels with lower second formants are produced further back in the mouth.

A particularly important cue for determining place of articulation of plosives appears to be the direction of the second formant transition of the adjacent vowel (Cooper, et al., 1952; Delattre, et al., 1955). This perceptual cue was thought to arise from the articulatory movement from the place of constriction for the plosive to the configuration of the vocal tract necessary for the production of the

vowel. This cue also appears to be affected by context, without however affecting the perception of place. For example, a rising second formant transition is seen in the syllable /di/, but a falling transition is seen for /du/. Although the direction of the transition is different, the perception of /d/ remains intact. This was assumed to be related again to some underlying articulatory phenomenon. That is, the origin of the second formant transition for /d/ is held constant to some extent, even though the direction of the transition is changing with the following vowel. Origin frequency (or locus) is believed to be related to the plosive's point of constriction, resulting in the stability of the /d/ perception (Delattre, et al., 1955).

The magnitude of the second formant transition (i.e., the difference between the origin and the frequency at which the transition levels off) has also been shown to provide place of articulation information for plosives (Liberman, Delattre, Cooper and Gerstman, 1954). In addition, information from the third formant appears to be important for voiced plosives (Harris, Hoffman, Liberman, Delattre and Cooper, 1958). Cues from third formant transitions add to

the available information from the first and second formant transitions for improved discrimination of /b, d, g/.

Numerous cues have been isolated which are thought to help distinguish between voiced and voiceless plosives. By making a systematic change (cutback) in the onset of the first formants of the vowels in synthetic CV syllables, Liberman, Delattre and Cooper (1958) were able to convert voiced plosives to their voiceless cognates, without affecting place identification. Removing the initial portion of the transition resulted in other significant changes in the first formant characteristics, such as the elevation of its starting frequency. The addition of noise during this cutback period also increased the perception of voicelessness, partly because the cutback period corresponds to the normal period of aspiration. In summary, synthetic syllables perceived as voiced plosive-vowel pairs have rising first formant transitions beginning relatively early and at a low starting frequency, with no aspiration. Those perceived as voiceless plosive-vowel pairs have small to negligible first formant transitions, the onset of the formant beginning later in time and at a higher frequency, and with

aspiration present.

Lisker and Abramson (1964) have argued that the above set of acoustic cues for distinguishing between voiced and voiceless plosives could be derived from a single, underlying articulatory variable, voice onset time (VOT). This variable is related to the relative timing of the stop-closure release and the onset of vocal fold vibration. Differences in voice onset time have been shown to be the basis for perceptual separation of the plosives into three categories (i.e., voiced, unaspirated voiceless, and aspirated voiceless). In a more recent study, Stevens and Klatt (1974) point out that the initial voiced plosive has a well-defined, rising first formant transition, but the initial voiceless does not. Thus, differences have been found not only in voice onset time, but in the first formant transition, as well. An additional cue may be the presence or absence of a significant and rapid spectrum change at the onset of voicing. By independently manipulating the redundant cues of first formant duration and VOT, Stevens and Klatt demonstrated the complex trading relationship between the two cues.

For plosives in word-final positions, the consonant

may be released or unreleased. If the final plosive is unreleased, voice onset time and aspiration information is not available and other cues must be relied upon for the voiced/voiceless distinction. For these final plosives, the duration of the preceding vowel appears to be an important distinguishing cue (Raphael, 1972). As with the fricatives, the vowel appearing before a voiceless plosive is usually shorter than the vowel preceding a voiced plosive (i.e., /dus/ vs. /duz/).

Nasals

The nasal consonants /m, n, ŋ/ have been shown to be differentiated from their plosive cognates by the presence of the nasal murmur. The latter property manifests itself as a large concentration of low-frequency energy in the 250-300 Hz region (Malecot, 1956) resulting from the increased size of the resonating cavity which now includes both the oral and nasal areas (Fujimura, 1962). Differences in the spectral characteristics of the nasal murmur between nasal sounds do not appear to be well defined. Rather, information for determining place of articulation seems to

be conveyed by the formant transition of the following vowel. The direction of these transitions appears to be similar to those for the corresponding voiced plosives /b, d, g/ (Liberman, et al., 1954).

Most of the studies discussed above have attempted to isolate important acoustic cues for consonant perception using synthetic speech stimuli generated from highly simplified spectrograms. In doing so, a preliminary decision has been made as to which acoustic cues should be included in the synthesis process. Thus, it is possible that certain cues which may be characteristic of a given sound (but redundant for the purpose of speech synthesis) have been omitted. In cases where speech is distorted or masked by noise, these seemingly redundant cues may be of particular importance. In a similar manner, redundant cues may be used by hearing-impaired listeners. Hence, it is necessary to study the importance of cues occurring in real speech.

There have been a relatively small number of cues examined thus far, including mainly transitional and burst cues. The temporal information obtained from waveform analysis (i.e., duration, silent periods, onset characteristics,

relative amplitude) has not been thoroughly investigated (Cole and Scott, 1974). The acoustic and articulatory properties of consonants have been shown to be influenced to a great extent by context. In a perceptual study using acoustic information extracted from real speech, Winitz, Scheib, and Reeds (1972) emphasized the importance of the integration of various cues contained in the consonant and the vowel. The use of isolated synthetic segments may not reflect the true context-dependent speech perception process. In addition, in running speech timing constraints will often not allow the articulators to complete their movements to the target position before the onset of the next sound. Thus, formant frequencies, for example, may never reach their characteristic steady-state values. An interesting aspect of this effect, however, is that the "undershoot" of the formant frequency does not seem to adversely affect the perception of the utterance (Lindblom, 1963).

In addition, more emphasis needs to be placed on the role of the peripheral mechanism. Stevens and House (1972) point out that the type of analysis performed on the input acoustic stimulus by the auditory mechanism is

not a simple frequency analysis, but entails additional transformations of the processed signal. The analysis process must account for the ability to perceive cues in the signal such as rising vs. falling spectra, rapid vs. gradual onsets, and relative frequency positions (i.e., formants). The auditory system's efficiency in dealing with speech or speech-like signals containing these cues is beginning to be explored (Nábělek and Hirsh, 1969).

FACTORS AFFECTING SPEECH INTELLIGIBILITY

The earliest quantitative studies of the relative intelligibility of speech sounds were undertaken to evaluate the relative performance of telephone transmission systems (Campbell, 1910). Standardized lists of syllables, words, and sentences were later introduced and the "articulation scores" for each of these materials were examined for telephone circuits and frequency filtering systems (Fletcher and Steinberg, 1929). One of the earliest observations made with these tests was that the intelligibility decreased as the level of the speech approaches the threshold of hearing. On the basis of this observation, Fletcher and Steinberg (1929) recommended the use of these tests "in measuring

the deafness of an observer" (p. 851).

When measuring the intelligibility of speech by articulation methods, one may analyze the subject's responses in terms of an articulation function whereby the scores obtained are plotted as a function of signal level or speech-to-noise ratio. An alternative method (first described by Campbell, 1910) is to record the number of different responses for each speech stimulus. This table of target-response pairs, known as a "confusion matrix", may be used to determine which phonemes are consistently misidentified.

The articulation-test method was used extensively to assess the relative performance of communication systems, and of related factors such as masking and filtering (Egan, 1948). Non-acoustic factors were also studied, such as: (i) relative frequencies of occurrence of the speech sounds, (ii) the use of syllables vs. words or sentences, (iii) learning effects, and (iv) relative difficulty of test materials. Egan reported on the development of tests at the Psycho-Acoustic Laboratory. The monosyllabic lists were constructed with commonly-used words of equal average difficulty, and equal phonetic compositon representative of the

fundamental English speech sounds.

The importance of a number of the factors discussed by Egan was later confirmed in a study by Hirsh, Reynolds, and Joseph (1954) in which their differential effects were evaluated under both masking and filtering conditions. It has also been shown that previous knowledge of the message set can affect the determinant factors of intelligibility (Pollack, Rubenstein and Decker, 1959). They showed that in a known message set, intelligibility of the words presented in noise was determined solely on the acoustic similarity of the materials. However, in an unknown message set, the average frequency of usage of the words became a confounding factor.

Articulation tests have also been used to evaluate the relative importance to speech intelligibility of different frequencies within the speech spectrum in both quiet (French and Steinberg, 1947) and in noise (Pollack, 1948). By systematically filtering the speech signal under different noise conditions, Pollack was able to draw the following conclusions about the relative contributions of different portions of the speech spectra to speech intelligibility:

(1) Higher frequencies alone contribute little, but when coupled with low frequencies their contribution is significant. (2) The contribution of lower frequencies is small and tends to be detrimental at high intensity levels (due to upward spread of masking); at low intensity levels, however, low frequencies are beneficial in that they tend to add to the overall signal level. (3) The contribution of a single frequency band is <u>not</u> independent of the contribution being made at the same time by other bands. (4) When speech is presented in white noise, the relative contributions of the speech frequencies is not constant, but is a function of the relative levels of speech and noise for each frequency band. (5) When high frequencies alone are present in the signal (high-pass filtering), the improvement of intelligibility with increasing intensity is more rapid than in a low-pass filtering condition. In a related study, Hirsh, et al. (1954) found an interaction between utterance type and filtering. Nonsense syllables tended to be more adversely affected by low-pass and high-pass filtering than various types of monosyllabic and polysyllabic words.

A simple measure of the relative distribution of

high and low frequency information for speech intelligibility is the "cross-over" frequency. French and Steinberg (1947) found that the same articulation score may be obtained when only frequencies below 1900 Hz are passed as when only frequencies above 1900 Hz are passed. A slightly lower "cross-over" point (1600 Hz) was found by Hirsh, et al. (1954) using filters which did not cut off as sharply. An even lower cross-over (approximately 1000 Hz) is found when filtering is performed on speech presented in a background of noise (Pollack, 1948).

The effects of noise on speech intelligibility is a problem of longstanding interest. Miller (1947) identified three important aspects of the noise masker: (i) the intensity of the noise in relation to the intensity of the speech (i.e., speech-to-noise ratio), (ii) the spectrum of the noise, and (iii) the temporal characteristics of the noise. Thus, a wide-band, uninterrupted masker with a sloping spectrum similar to the long-term speech spectrum was found to be the most efficient masker. Miller emphasized the importance of the noise spectrum rather than its waveform. That is, masking is independent of the phase relations among the components of the noise.

The masking effects of narrow bands of noise at low to moderate noise levels are most easily interpreted in relation to the speech spectrum in the corresponding frequency band. Low frequency bands do not have much effect at low intensities due to the relatively strong speech components in that area. At higher intensities however, masking is quite effective, even spreading over the entire frequency range. High frequency bands, on the other hand, are effective in masking high frequency components of speech even at low intensities due to the weak speech components in that frequency region. However, there is very little downward spread of masking even for very high noise levels. As was the case for filtering, Hirsh, et al., (1954) found that noise level and signal-to-noise ratio have differential effects on the intelligibility of words and nonsense syllables.

An attempt was made to develop a computational procedure for predicting articulation scores, based on the speech-to-noise ratio in several frequency bands. A practical approach to this problem is to compute the Articulation Index (AI). The derivation of the AI is based on

the assumption that the contribution to speech intelligibility by a certain band of speech is independent of the contributions of other bands (French and Steinberg, 1947; Miller, 1947; Kryter, 1962a,b). Specifically, the spectrum is divided into 20 bands, each of which is assumed to make an equal contribution to intelligibility. The sum of the effective signal-to-noise ratios in each of these 20 bands is proportional to the AI (an AI of 1.00 corresponding to a sum of 600). The maximum contribution of each band is limited to 30, the minimum to 0 dB. These two values set the dynamic range over which speech peaks conveying information tend to vary. That is, if the level of the speech peaks falls below the level of the noise (i.e., a speech-to-noise ratio within the band of less than 0 dB), that particular band makes no contribution to the articulation index. If the speech peaks are more than 30 dB above the RMS level of the noise, the contribution is equal to 30. Thus, between these two limits, the effectiveness of the band is dependent upon the speech-peak-to-RMS-noise ratio in that band. Computationally, the level of the speech peaks is obtained by adding 12 dB to the RMS speech level

in the band.

Ideally, it would be useful to have a measurement device which would be independent of the test materials used, and allow for a prediction of the effect directly from it. The Articulation Index has been used for this purpose. Once the AI has been calculated, it can be used to predict nonsense syllable, word, or sentence articulation scores for a given listener or team of listeners. This prediction is obtained from empirically-derived curves relating the AI to percent intelligibility. A central assumption to these predictions is that percent intelligibility is a monotonic function of the AI.

Correction factors have been suggested to improve the accuracy of the Articulation Index procedure. These include corrections for the effects of high noise levels, peak clipping, interrupted or sharply-attenuated noise, variations in vocal effort, and reverberation (Kryter, 1926b). There are, however, many factors that have not been accounted for (i.e., severe peaks or valleys in the speech or noise spectra, severe filtering, nonlinear distortions or combinations of distortions, the use of speech

babble as a masker, or noise having significant energy below 600 Hz or above 4800 Hz). One basic assumption of the AI, the independence of the bands, is contradicted by the evidence showing the interdependence of different speech bands in contributing to intelligibility (Rosenthal, Lang, and Levitt, 1975). An alternative method could be the calculation of the cumulative contributions of the bands to intelligibility. A modification proposed by Kryter (1962b) in which an "effective" signal-to-noise ratio is calculated attempts to take into account upward spread of masking effects.

A set of contradictory findings have been reported in studies with speech and/or noise presented at high intensities. Although speech-to-noise ratio appears to be the primary variable, at high noise levels poorer intelligibility is measured than would be predicted solely on the basis of speech-to-noise ratio.

When filtered speech is presented in quiet, there is a significant reduction in intelligibility at high levels (French and Steinberg, 1947, p. 101). However, no such reduction is observed for high-intensity wide-band

speech presented in quiet (Fletcher, 1953, pp. 341-343). Finally, a reduction in wide-band speech intelligibility is seen when high-intensity speech is presented in a noise background (Pollack and Pickett, 1957a). In a subsequent paper, Pollack and Pickett (1957b) suggest that noise may be regarded as reducing the contribution of specific speech frequencies, as had been assumed for filtering. They point out that the Articulation Index does, in fact, consider these two effects in the same manner, despite the evidence that the two factors may have differential effects (Hirsh, et al., 1954). Supporting data have been presented (Pollack and Pickett, 1957b) showing that filtering of speech at high levels produces a decrement in speech intelligibility which is roughly equivalent to the reduction obtained when high-intensity speech is presented in noise. In addition, high speech and noise levels have been shown to cause a reduction in intelligibility with constant speech-to-noise ratios (Pollack and Pickett, 1958). This finding tends to support the assumption that over-all level, rather than speech-to-noise ratio, may be a determining factor in speech intelligibility.

PHONEME RECEPTION BY NORMAL-HEARING LISTENERS

Consonant Reception

Miller and Nicely (1955) studied perceptual confusions using 16 CV syllables with the vowel /ɑ/, produced by female talkers. The syllables were presented in a number of conditions of distortion, with adjustments made in speech-to-noise ratio, and low-pass and high-pass filtering. The 16 phonemes were categorized according to their acoustic and articulatory characteristics or cues. These so-called features were: voicing, nasality, duration, affrication, and place of articulation. The authors assumed that these features act independently of one another and attempted to evaluate the amount of confusion related to each feature under each distortion condition. Under conditions of low-pass filtering or wide-band noise masking, the confusions fell into consistent patterns. This is understandable, in that the broad-band noise would tend to mask the higher frequencies, resulting in similar distortion effects as the filtering. Under these conditions, the consonants fell into 5 clearly-defined groups: /ptk/, /bdg/, /f, θ, s, ʃ/, /v, ð, z, ʒ/, and /mn/. Most confusions were within each

group, few confusions occurring between the groups. The authors concluded from these results that voicing and nasality were most easily distinguished under conditions of noise and low-pass filtering, presumably because the acoustic cues important for the perception of these two features lie primarily in the low frequency region. Place of articulation, on the other hand, was affected by these distortions, presumably due to its dependence on weaker higher-frequency cues (such as the spectrum of the frication for voiceless plosives and fricatives). High-pass filtering caused all features to be similarly affected. Analysis of the confusion matrices showed a scattering of responses for these conditions, which was interpreted as showing a reduction in audibility for all features.

In an analysis of the Miller and Nicely data, Pickett and Rubenstein (1960) point out that when the low-frequency portion of the speech spectrum is available, perception of voicing was independent of place and manner of articulation. However, when only the high frequencies are heard, voicing interacts with place and manner, in that it is perceived better with alveolars than labials and better

with plosives than with fricatives. In their own study using broad-band noise and noise filtered at -12 dB/octave, they found similar effects of noise on voicing perception for syllables differing with respect to position (initial vs. medial or final) or manner (plosives vs. fricatives). The spectrum of the noise did, however, have a differential effect on voicing perception of syllables differing in place. For example, voiced-voiceless discrimination was better with alveolars than with labials. The low-frequency cues dominant in conditions of broad-band noise are operating independent of place. However, in low-frequency noise, the higher-frequency information more important for perception of alveolar consonants is available.

Busch and Eldredge (1967) compared scores for consonant identification in studies which used a "white" noise background with scores from a study in which "speech-shaped" noise was used. Their data showed negative correlations between these 2 groups, which was consistent with Pollack and Rubenstein's (1960) findings. The concentration of low-frequency power in the speech-shaped noise resulted in a deterioration of scores for nasals, liquids, and voiced stops, without significantly affecting the voiceless stops

and fricatives. On the other hand, the scores from the studies using broad-band noise showed the effects of the loss of middle and higher frequency cues. Busch and Eldredge also point out the differences in the results which are evident only when scores for initial consonants and final consonants are analyzed separately. They warn that grouping these data may tend to obscure important perceptual interactions.

The effect of the position of the consonant in the utterance has also been studied by Pederson and Studebaker (1972) using broad-band noise in several speech-to-noise ratios. For final consonants, the voiced plosives had the highest error rate, followed by the voiced fricatives, voiceless plosives and voiceless fricatives. For initial consonants, however, voiceless plosives and fricatives had more confusions than their voiced counterparts.

Abbs and Minifie (1969) found significantly fewer discrimination errors between pairs of syllable-final fricatives than between pairs of syllable-initial fricatives. In addition, fewer discrimination errors were found between pairs of fricatives which differed in voicing than between pairs in which the fricatives were either both voiced or

both voiceless.

It has been suggested that disagreements between studies of consonant reception could be due, at least in part, to differences in the spectra of the masking noise used (House, Williams, Hecker and Kryter, 1965). Another important factor is the time-envelope characteristic of the noise. Using as a masker a noise whose amplitude varies in time as the speech stimulus (referred to as speech-envelope noise), Horii, House, and Hughes (1971) were able to study the effects of a noise having an instantaneous speech-to-noise ratio which was essentially constant. For noise with a non-time-varying envelope, vowel intelligibility is found to be consistently higher than consonant intelligibility. Horii, et al. argued that the reason for this was that the standard procedure for obtaining speech-to-noise ratio results in a measure of vowel-to-noise ratio. Thus, the average level of the consonant may fall below the noise in some conditions. When the speech-envelope noise is used, however, the articulation functions for the consonants and the vowels are the same, when adjusted to have the same instantaneous signal-to-noise ratio. Miller and Nicely (1955), among others,

have indicated that noise tends to effect certain classes of consonants more than others. Using the envelope noise, however, Horii, et al. measured the intelligibility of stops, fricatives, nasals, liquids/glides, and affricates to be about the same. In addition, with increasingly poorer speech-to-noise ratios using the envelope noise, the ordering of relative intelligibility between the groups remains the same, and the differences in scores between consonant groups is reduced. Thus, they found no interaction between the class of the consonant and speech-to-noise ratio or noise level when phoneme-to-noise ratios are equated. In addition, the authors suggest that envelope cues, such as duration and level variations, are being utilized to a fuller extent.

Wang and Bilger (1973) measured consonant discrimination under different listening conditions. Using statistical procedures based on information theory, they attempted to isolate the features which accounted for good discrimination in different conditions while taking into account the internal redundancy of the features. The results were interpreted in terms of the "proportion of total transmitted information" provided by each feature, independent of the

others. The stimuli were CV and VC syllables, with maximum durations of 511 msec, presented in isolation (no carrier phrase). Changes in speech-to-noise ratio and broadband noise level resulted in changes in the relative importance (as defined in statistical terms) of the features. For example, voicing and nasality were well-perceived in noise but were surpassed by other features in quiet conditions. Differences were also seen as a function of consonant position. Furthermore, good performance on one feature seemed to be dependent, in some cases, on the intelligibility of another feature (i.e., siblance and duration). Their analyses indicated that a small number of features (i.e., voicing) tended to consistently account for a large proportion of transmitted information (although less so for final consonants than for initial consonants).

The authors point out that a large proportion of transmitted information may be accounted for in terms of the so-called articulatory and phonological features. However, no set of features were found which were independent of context, condition or experimental task. They conclude that these features may be helpful in describing

consonant confusion data in terms of the importance of acoustic cues. They may not, however, represent underlying perceptual processes.

Vowel Reception

Miller (1956) analyzed the confusions made between vowels presented in the /h-d/context. When the utterances were low-pass filtered to 670 Hz, there were few confusions between vowels which differed on the basis of the frequency of the first formant. However, vowels with similar first formants but which differed with respect to second formant frequency had higher rates of confusion.

This effect was also seen by Pickett (1957) using noises of different spectral content. When the second formant was masked by either high frequency or broad-band noise, the most common response was a vowel with a similar first formant frequency (the remaining cue). These vowels were also characterized by either a very high or very low second formant frequency. This was thought to be related to the higher frequency of occurrence of English vowels having relatively high or low second formant frequencies. In addition, when only one formant is available, differences

in duration of the vowels tended to act as an additional cue. On the other hand, the intensity differences between vowels typical of normal speech played a less important role. More intense vowels were more intelligible only in the low-frequency noise conditions. When the higher formant is masked by high-frequency or broad-band noise, the remaining first formant cue did not appear to improve intelligibility.

Several variables have been isolated which have been shown to affect a listener's vowel identification ability. Peterson and Barney (1952) have emphasized the importance of the listener's previous language experience. Their data indicate that if a person fails to distinguish between two vowels in production, he is not likely to discriminate between them in a listening task. In addition, when a confusion was made, it tended to be with a vowel that was adjacent to it on the vowel quadrilateral. Furthermore, the vowels that were correctly identified most often were the vowels that are produced at the extreme points of the articulatory mechanism.

The adjacent consonants have been shown to have an effect on the "secondary" acoustic characteristics of the vowels, such as duration, intensity, and fundamental fre-

quency (Lindblom, 1963). The effects of these consonant-vowel interactions on vowel perception has not, as yet, been clearly determined.

PHONEME RECEPTION BY HEARING-IMPAIRED LISTENERS

The study of consonant and vowel reception has recently been widened to include the investigation of error patterns made by hearing-impaired listeners. Included among these analyses were the effects of background noise, filtering effects, and the relationship (if any) of specific patterns of confusions with certain auditory pathologies or audiometric configurations.

Consonant Reception

Oyer and Doudna (1959) analyzed responses to the CID W-22 word lists by phonetic error type. Similar confusion patterns were found for patients with conductive and sensori-neural hearing loss, although the errors for the latter group were less consistent. Plosive and fricative errors were most common, as were phonemes in word-final position.

Pickett, et al. (1970) studied phoneme reception by

listeners with severe-to-profound hearing impairment. The greater availability of low-frequency cues, along with duration information, accounted for better reception of voicing, nasals, liquids, and glides. Higher frequency information necessary for place reception was lacking, even for those subjects with "better" hearing.

Several papers have attempted to relate the patterns of confusions with audiogram configuration. Lawrence and Byers (1969) studied the ability of subjects with high-frequency hearing impairments (of different durations) to discriminate among voiceless fricatives. The result, in general, was a high percentage of correct identifications with idiosyncratic patterns of response for the 5 subjects tested. Most errors were limited to two groups, confusions between /s/ and /ʃ/ and between /f/ and /θ/. To account for the relatively good performance, the authors suggest a relationship between fricative identification and second formant transition of the following vowel. More confusions were made when the fricative was paired with a vowel having a higher second formant (/i/ and /e/) where subjects' hearing was impaired, than for fricatives paired with vowels

having lower second formants (/o/ and /u/).

Owens, Benedict and Schubert (1972) tested subjects with sensori-neural hearing impairment, along with a group of normal-hearing subjects listening to filtered speech. Assuming that the resulting errors were due to the dependence on missing spectral information, an attempt was made to relate the confusions to the steepness of the slope of the hearing loss or its bandwidth. Such relationships could be shown for only a few phonemes in a few positions. It appears that hearing loss does not act as a simple "filter" and comparisons of this kind should be made with caution.

Sher and Owens (1974) studied the effect of high-frequency hearing impairment (of various etiologies) on all consonant confusion patterns. As in Owens, et al. (1972), the results were compared to normal-hearing subjects' responses to materials filtered to simulate the slopes of various high-frequency impairments. When the two groups were matched for audiogram and filter slope similar results, in terms of mean scores and phoneme errors, were found. Subjects with mildly sloping audiograms (or normal-hearing subjects listening to filtered speech with similar slope)

tended to perform significantly better than subjects with sharply sloping audiograms. The materials used in this study may have been subject to word-familiarity effects. In addition, choice of presentation level could have prevented certain subjects from receiving all high-frequency information due to their impairment in that region.

Recently, Bilger and Wang (1976) attempted to group subjects according to their phonemic confusions, and then to analyze the groups in terms of audiometric configuration. The phonemic confusions were analyzed in the same manner as their earlier study using normal-hearing listeners (Wang and Bilger, 1973). The redundancy inherent in perceptual features makes it difficult to assign a performance measure to one feature, not knowing for certain that the score was based on that feature alone. The statistical analysis procedure used by the authors rank ordered the features in terms of their proportion of transmitted information. Thus, although the percentage of transmitted information may be the same for two subjects, the features which have contributed the most information may differ. Subjects with similar "feature profiles" formed three groups which were found

to be similar in terms of audiometric configuration: (i) normal, control subjects and subjects with mild impairments, (ii) subjects with high-frequency hearing loss, and (iii) subjects with flat configurations. Using data from another study (Reed, 1975), group membership was predicted from audiometric data alone. The results were most accurate for predicting members of the high-frequency loss group.

Methods other than standard speech reception or speech intelligibility tasks have been used to investigate the perceptual processing strategies utilized by the hearing-impaired listener. These methods include the use of similarity judgments of phoneme pairs (Walden and Montgomery, 1975) and reaction-time measurements for same/different discrimination of phonemes (Reed, 1975). Results of these studies indicate that the hearing-impaired listener may rely on only one dominant feature which is perceptually distinct, rather than weighing each feature equally (Walden and Montgomery, 1975). Reed (1975) suggests that these features are used in a less efficient manner by the hearing-impaired than by normal-hearing listeners. Further investigation into these processes is needed, for both the normal-hearing

and hearing-impaired populations. In addition, the possible ramifications of using methods such as these in analyses of speech reception should be explored.

Vowel Reception

Vowel errors made by hearing-impaired listeners were studied by Ownes, Talbot, and Schubert (1968) using a closed-response format. The results showed this to be a relatively easy task for this population (94% correct vs. 46% correct on standard CID W-22 discrimination tests). Errors did not fall into a distinct pattern in terms of pathology of hearing impairment. However, the most common substitutions were, as expected, with neighboring vowels on the vowel quadrilateral. In addition, the most common incorrect response was the vowel /ʊ/. This finding could be inferred from a plot of first vs. second formant frequency, on which the vowel /ʊ/ shows a great deal of overlap with other vowels (Peterson and Barney, 1952).

Using severe-to-profoundly impaired subjects, Pickett, et al. (1970) found that vowel confusions were related to the similarity of the first formant frequency. The subjects' lack of discrimination ability for the sec-

ond (higher) formant frequency forces the reliance on first formant information. Back vowels (i.e., those with lower frequency second formants) were heard correctly more often than front vowels (higher frequency second formants). It has also been shown that these subjects have larger difference limen for higher formant frequencies than for lower formant frequencies, for which their DL's are equivalent to normals (Pickett and Martony, 1970).

Using synthetic vowels, Danaher, Osberger, and Pickett (1973) found that discrimination of the second formant transition is poorer when heard in the presence of first formant energy. This was assumed to be a result of upward spread of masking of the lower frequency formant. Additional masking effects were seen when a small transition was added to the first formant. Discrimination of the second formant improved when the amplitude of the first formant was reduced. However, further attenuation of the first formant resulted in a signal which was not recognized as speech. In a follow-up study, Danaher and Pickett (1975) found first formant masking effects present even when the two formants were separated to two ears. In ad-

dition, the first formant seems to produce a type of backward masking. The result of this was a reduction in the ability to discriminate cues preceding the vowel. Replicating a result of the first study (Danaher, et al., 1973), they found that the reduction in discrimination ability as a result of these masking effects was not related to audiometric configuration. Persons with similar audiograms do not necessarily show similar discrimination thresholds. The authors emphasize the importance of auditory training to improve the listening skills of those subjects who are most severely affected by these masking phenomena.

Data from Danaher, et al. (1973) indicated that some deterioration in discrimination ability occurs for all subjects (including those with normal hearing) when listening at high signal levels. They concluded that there may be some "normal" upward spread of masking at high sound levels. In addition, one should keep in mind when evaluating the results of these two studies that these measurements were of discrimination ability of synthetic vowel components and were not measures of speech sound recognition.

SPECTRAL ANALYSIS

One goal of speech analysis is to provide an accurate description of the many complex factors involved in the speech production process. One way of accomplishing this is to develop a model which would account for all of the important physical characteristics of human speech. Then it is necessary to find mathematical relationships to represent these physical parameters as simply as possible without sacrificing accuracy. The starting point for this description, therefore, must be a description of the important aspects of speech production.

Mechanism of Speech Production

The acoustical speech waveform can be described as an acoustic pressure wave which is generated when air from the lungs is forced through the vocal tract. Voiced sounds (including vowels) are produced when the vocal folds are caused to vibrate, opening and closing in an almost periodic sequence and producing quasi-periodic pulses of air pressure which are transmitted through the vocal tract. The rate of vocal fold vibration is known as the fundamental frequency and is defined as the reciprocal of the time interval be-

tween the periods of the acoustic pressure wave.

The volume velocity of the air flow through the glottis (the area between the vocal folds) is determined by the subglottal air pressure, in addition to the length, thickness, and tension of the folds themselves. The vocal tract (from the glottis to the lips) can be described as a non-uniform acoustic tube whose shape, as controlled by the velum, tongue, jaw, and lips, varies with time. Another sound transmission tube, the nasal cavity, is closed off by the velum during production of non-nasal sounds. Voiceless sounds (including fricatives and plosives) are generated when the vocal folds are held open, air forced past them, and then a constriction formed by the articulators. This creates a turbulence and a source of noise to excite the vocal tract. The waveform from a voiceless utterance shows no evidence of pitch periods, as opposed to the waveform of a voiced utterance, which is nearly periodic.

The sound sources for voiced and voiceless sounds described above create wide-band excitation of the vocal tract. The vocal tract in turn acts as a time-varying filter, imparting its transmission properties (including its

resonant frequencies) onto the signal output at the lips.

There have been attempts at modelling the above process, in order to describe the acoustic output from the speech production mechanism. One such model is that proposed by Fant (1960) in which speech is represented as the output of a discrete, time-varying linear filter, one which approximates the transmission properties of the vocal tract. The vocal tract changes shape relatively slowly during continuous speech, resulting in slowly-changing sound transmission properties. It is assumed, therefore, that the vocal tract shape may be approximated by a succession of stationary shapes. It becomes possible to define a transfer function for the vocal tract which is represented by its poles (resonances) and zeros (antiresonances).

For voiced sounds, the input to the time-varying linear filter representation of the vocal tract would be a triangular pulse train (the glottal waveform) in which the spacing between the impulses corresponds to the period of glottal vibration. For unvoiced sounds, a random noise with a flat spectrum is the filter input. One shortcoming of this simple model is that it does not account for mixing

of inputs (such as voiced fricatives) or for the addition of a side-branch filter (for nasals). The resulting vocal tract model is an all-pole model consisting of a small number of 2-pole resonators, defined as formants. In addition, a lip radiation factor takes into account the radiation of sound from the mouth by the transformation of the volume velocity waveform at the lips to the acoustic pressure waveform at some distance from the lips.

In the above model, speech signals are represented in terms of several time-varying parameters. These parameters are related to the transfer function of the vocal tract and the characteristics of the glottal excitation. By modeling the speech waveform directly instead of the spectrum, one may avoid the problems of applying a quasi-periodic signal to frequency-domain methods, such as Fourier analysis. Only a short segment of the speech wave is needed to obtain accurate results, allowing for accurate analysis of rapidly-changing speech events (such as plosive bursts) and higher-pitched speakers (women and children). In addition, certain measurements (such as formant frequency) may now be determined directly from the speech wave by separating

out the poles of the vocal tract from the source function.

A problem encountered in this type of model, however, is that the source (glottal waveform) is not uncoupled from the vocal tract. Thus, the resultant speech signal is influenced by the properties of both the quasi-periodic source and by the transfer function of the vocal tract. For example, if the source spectrum has a reduction in spectral energy (antiresonance) near one of the natural frequencies (resonances) of the vocal tract, it would be difficult to determine the resulting formant frequency in the spectrum. The nasal cavity presents a similar problem. In the use of analysis procedures based on this model, one must be sure that the important information in the signal (i.e., formant frequency) is not hidden by the influence of the source.

There are other properties inherent in the speech signal which must be taken into account in any speech analysis procedure. One such property is its time-varying nature. The durations of the phonemes of speech may vary from as short as 20 msec (plosives) to as long as several hundred msec (steady-state vowels). In addition, certain phonemes may comprise a sequence of temporal events which may not be

easily resolved by certain analysis procedures. Furthermore, the spectral structure of the speech signal changes dramatically with time, ranging from the almost steady, periodic voiced sounds to the random vibrations typical of fricatives. An additional factor is that the periodicity of the normal voice may range from 80 Hz to 350 Hz. Finally, as stated previously, the analysis of the speech signal must account for the interaction between the harmonic structure of speech and its envelope structure.

Speech Waveform Analysis.

Speech waveform analysis is useful for studying the properties of the glottal source. The acoustic waveform, when displayed on an amplitude vs. time scale (i.e., an oscillograph), allows for an analysis of the general envelope characteristics of its time-varying properties. These include voicing characteristics, short-term as well as long-term amplitude changes, and the identification of boundaries between certain speech segments (i.e., between voiced and voiceless sounds). In normal speech, the period from the start of one glottal cycle to the start of the next may vary. However, the nearly periodic structure of a voiced

utterance may easily be defined by waveform analysis by measuring the distance between major peaks. These pitch periods of glottal vibrations are not seen in voiceless sounds, as these sounds are produced by noise excitation of the vocal tract. In addition to glottal pulse analysis, timing measurements such as durations may easily be obtained from the speech waveform. In addition, the onset characteristics (abrupt vs. gradual) of the signal may be determined.

Speech Spectrum Analysis

A particularly powerful method of analysis is that of short-term spectrum analysis. The acoustic data is transformed into a spectral representation by performing short-term Fourier analysis of the speech waveform. The spectrum over a short period of time is obtained by viewing the waveform through a time window, or weighting function. In using short-term spectrum analysis, it is assumed that the spectral characteristics of the signal are essentially constant within the time window. The choice of the time window introduces the problem known in general terms as the "uncertainty relationship" (Gabor, 1946), which refers to the relationship between frequency and time resolution. As

a general rule, the longer the time window, the better the frequency resolution of the resulting spectrum analysis. However, if a very long time window is used, temporal resolution will be poor, since it is not possible to measure time-varying changes within a time window. Another way of viewing the problem is that if the frequency filters used in a spectrum analysis are narrow enough to make the frequency resolution small, then the signal is not well resolved in time. For example, to resolve the harmonic structure of a male voice, the frequency resolution should be well below 100 Hz. However, the resulting time resolution will be too long to resolve the individual pitch periods, which are on the order of 8 msec. If the time resolution is set to 5 msec to resolve the pitch periods, then the frequency resolution will be too wide (approximately 200 Hz) and the harmonics will not be resolved.

Time-frequency trading is most clearly seen in the use of wide-band (300 Hz) and narrow-band (45 Hz) spectrograms. Using the wide-band filter, spectrograms are obtained with pitch period resolution but large amounts of averaging or smearing occurs in the frequency domain. On

the other hand, the use of the narrow-band filter results in spectrograms showing the harmonic structure of the speech but glottal pulses in the time domain are lost.

In choosing the analysis time window, these two factors must be weighed. As the duration of the interval decreases to minimize the averaging performed on the changing speech signal, the frequency resolution becomes limited. However, to insure good frequency resolution, a long time window is needed. For analyses which intend to ignore individual glottal pulses, it has been shown that a successful compromise would be a time window of 2 to 3 periods (8 msec each) in length, or approximately 24 msec.

To reduce the frequency smearing produced by the time window,* the use of a weighted time window is advised (Blackman and Tukey, 1958), rather than a rectangular time window. Markel and Gray (1976) suggest the use

---

*The problem of frequency smearing is inherent in Fourier analysis of any finite time interval. Using a rectangular window, the signal is analyzed for a specific period of time, while everything occuring before and after this period is ignored. The resulting Fourier transform is convolved with the spectrum of the time window in the frequency domain, characterized by spurious peaks known as sidelobes. Using a window such as the Hamming window reduces the effect of these sidelobes.

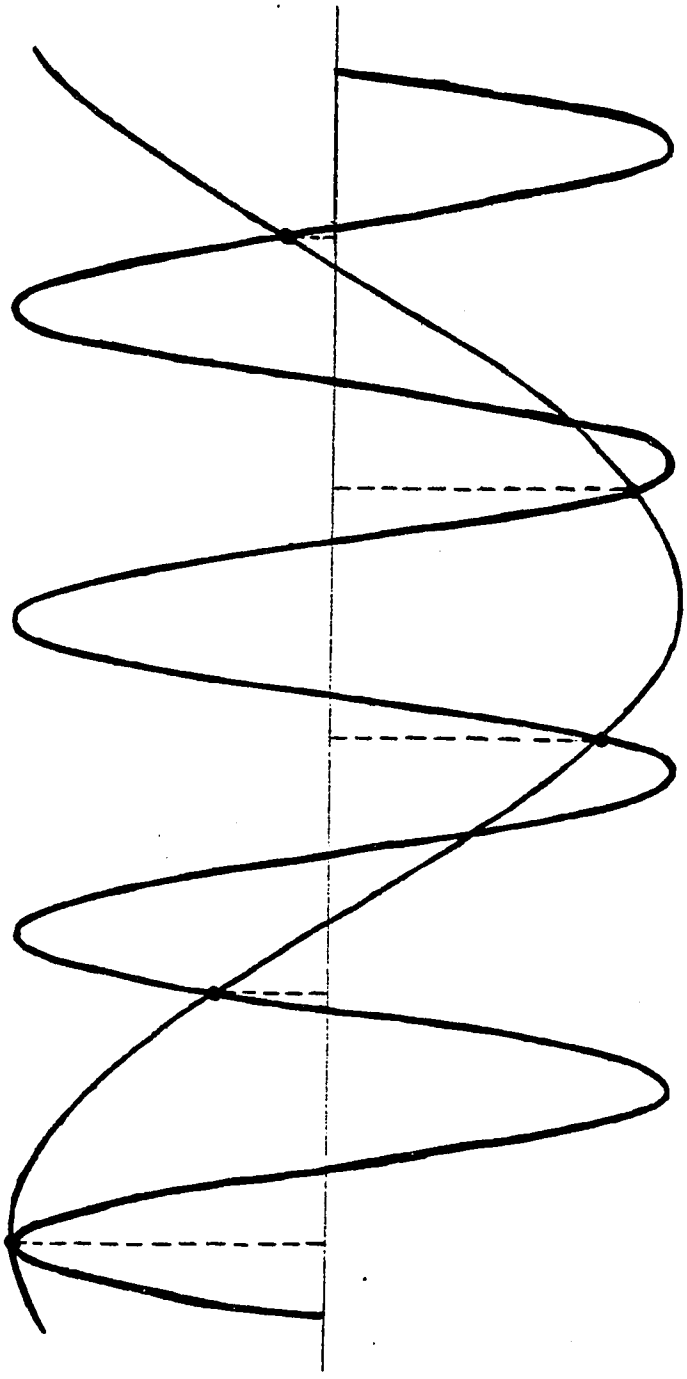
of windowing when analyzing intervals greater than 15 msec. An example of such a window is the Hamming window, which gives progressively less weight to the waveform near the edges of the window.

Digital Speech Analysis

The use of digital speech analysis systems has become increasingly popular in recent years. These procedures are capable of very complex and flexible analyses, resulting in accurate representations of the speech signal.

There are certain procedural constraints in digital processing which, when coupled with the properties of the speech signal, may result in inaccurate and misleading representations. Speech samples are entered into a computer by means of an A/D conversion of a continuous speech waveform. This involves the sampling of the waveform at a previously specified rate. It has been shown that the high-frequency components of a waveform can appear to be a low frequency if the sampling rate is too low (Bergland, 1969). An example of this effect, known as aliasing, may be seen in Figure 1. The possibility of aliasing can be removed by using a sampling rate which is high enough so

Fig. 1. An example of aliasing error resulting from a low sampling rate.



that the highest frequency in the signal would be sampled at least twice during each cycle. In other words, it is necessary to sample at a rate in excess of twice the bandwidth of the signal. Ideally, for speech having spectral components as high as 10 kHz, one whould sample at a rate of at least 20 kHz.

An additional digital processing constraint is related to the quantization of the signal. If the acoustic signals are quantized to less than 10 bits, the resulting background noise (quantization noise) will likely be audible. Therefore, to guard against this, 12 to 16 bits (1200 to 1600 linearly-spaced levels) should be used.

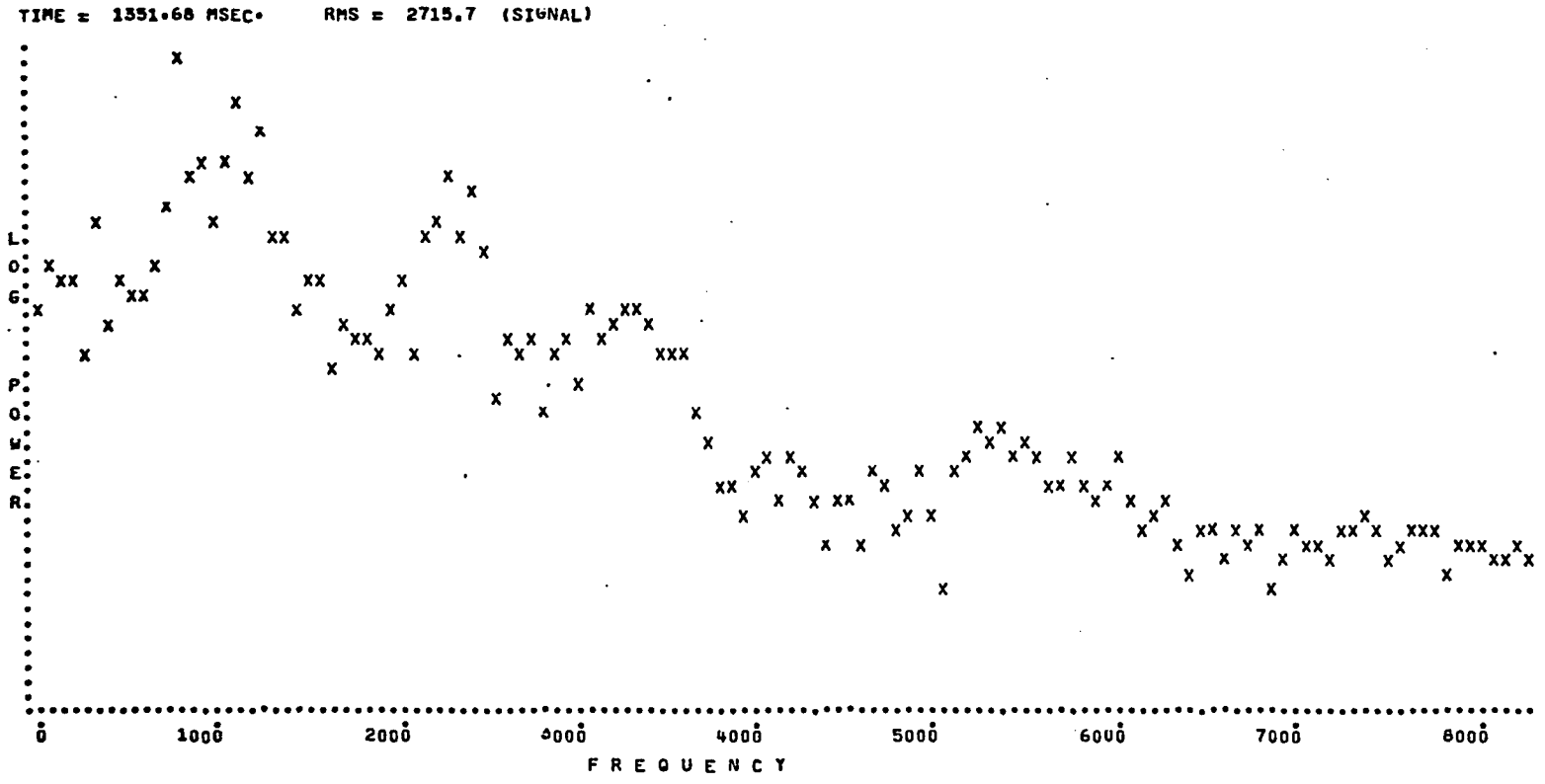
Digital speech analysis techniques, with these limitations in mind, may then be applied to the analysis of the speech signal. This is accomplished through the use of a model which describes the acoustic output from the speech production mechanism. If the assumption is made that this model, as previously described, accurately represents the speech process, the next step in speech analysis is to estimate the parameters of the model from the speech waveform.

Linear Predictive Coding

Recently, techniques utilizing linear predictive mathematics have been applied directly to speech analysis (Saito and Itakura, 1966; Atal and Schroeder, 1967; Markel, 1972). The linear predictability of the speech wave is the characteristic which forms the basis for this analysis procedure. The basic goal is to represent the speech waveform directly in terms of the parameters related to the transfer function of the vocal tract and the characteristics of the glottal source. Thus, the linear prediction model is very useful in helping to separate these two components for accurate measurements of formant frequency and fundamental frequency.

As previously described, short-term spectra only approximate the structure of the speech signal. This is due to the limitations of the digital processing techniques and the peculiar nature of the input signal, specifically its quasiperiodic nature and the presence of turbulence in the air flow (even during voiced sounds). The resulting power spectra have a great deal of variability, characterized by jagged power spectrum representations. Figure 2

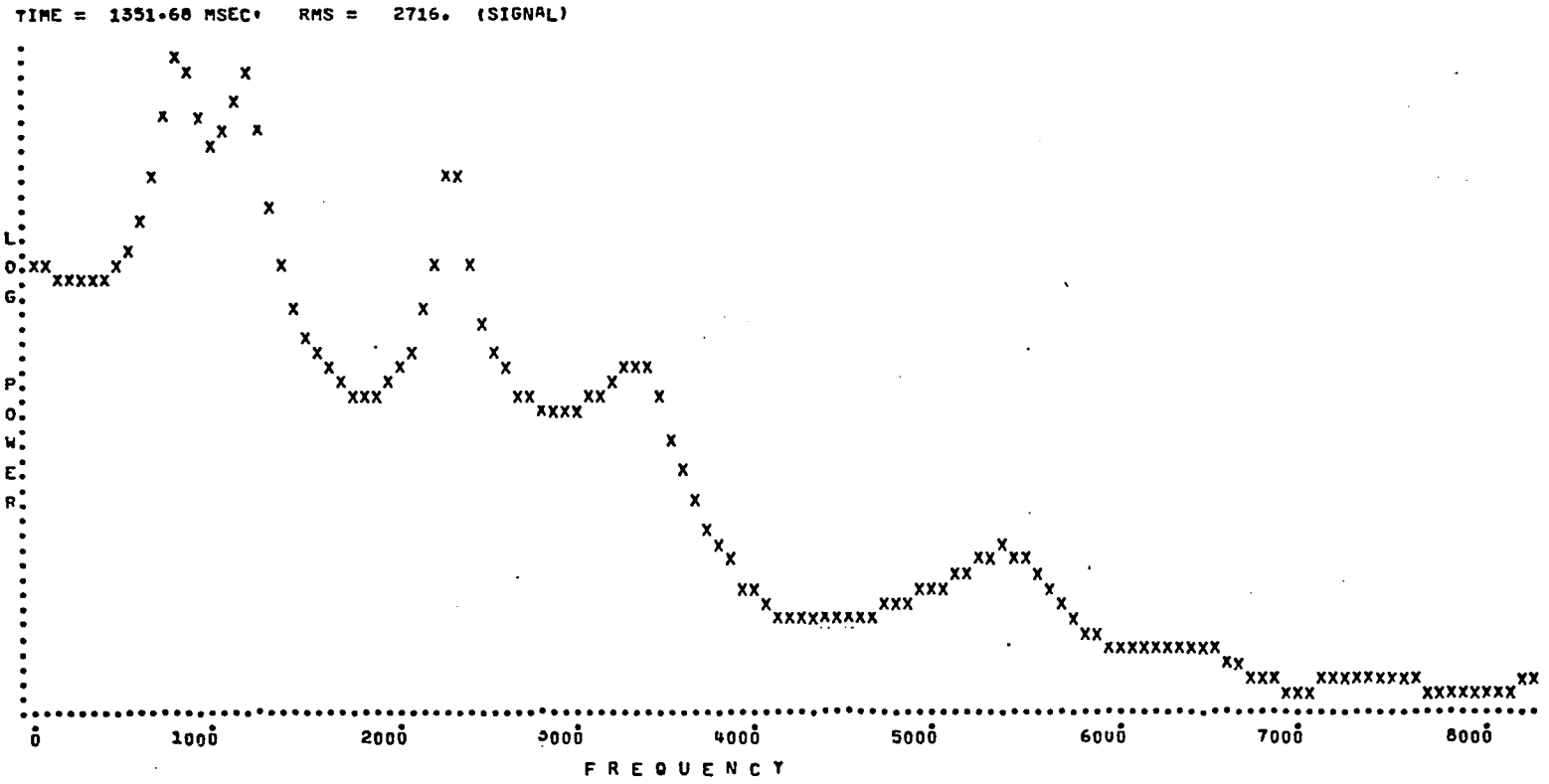
Fig. 2. Power spectrum output for the vowel /ɑ/ using a 30.72 msec time window.


TIME = 1351.68 MSEC.
RMS = 2715.7 (SIGNAL)
LOG POWER
0
1000
2000
3000
4000
5000
6000
7000
8000
FREQUENCY

shows a typical power spectrum output for a 30.72 msec time window. Among its other advantages. linear predictive coding appears to provide a good method of smoothing the speech spectra so as to simplify the problem of identifying formants. The smoothing effect may be seen in Figure 3 which shows the same data as contained in the time window of Figure 2 after having been analyzed by a linear prediction filter having 20 predictor coefficients.

In the linear prediction model, the speech signal is analyzed by predicting the present speech sample as a linear combination of the previous samples. The predicted sample is determined from these predictor coefficients. These predictor coefficients can be used to determine the parameters of the speech production model. In other words, the coefficients are the parameters of a digital filter (a linear filter whose filterning operations are implemented by digital computer). The coefficients, which account for the filtering action of the vocal tract, lip radiation, and glottal flow, may be derived from these parameters when applied to speech. The number of coefficients needed to successfully represent the speech segment is determined by

Fig. 3. Spectrum of vowel /ɑ/ after analysis by a linear predictive filter having 20 predictor coefficients (30.72 msec time window).


TIME = 1351.68 MSEC. RMS = 2716. (SIGNAL)
LOG POWER
0
1000
2000
3000
4000
5000
6000
7000
8000
FREQUENCY

the number of resonances (poles) and antiresonances (zeros) of the vocal tract in the range of frequencies being observed, in addition to the glottal pulse source and the radiation characteristics.

In general, only a few poles are needed to approximate the spectrum. Thus, the number of pole pairs, with a few exceptions, should equal the number of formants in the spectrum.* The notable exceptions to this rule are: (i) when the lowest formant in frequency is near the fundamental frequency, and (ii) when two formants lie close to one another on the frequency scale.

The resulting all-pole filter, or all-zero "inverse filter" (Markel, 1971), should represent the smooth spectral characteristics of the speech spectrum. Note that linear predictive coding is based on the use of an all-pole approximation; i.e., it assumes that the speech signal consists only of resonances. This is adequate for purposes of approximating a power frequency spectrum, but

*Two parameters are needed to specify each resonance or formant, e.g., either a pole pair, or two predictor coefficients.

raises problems for providing an accurate model (or description) of the speech mechanism, since speech sounds, such as nasals, do have strong antiresonances.

It appears that by selecting the analysis conditions correctly, the output of the linear predictor model (or inverse filter) will represent the estimated spectrum of the vocal tract response. Linear prediction can result in a linear, all-pole model of speech production by determining a small number of parameters directly from the speech waveform. Since the model spectrum is a representation of the smoothed data spectrum, formant frequency measurements can be made simply be peak picking. These formant estimates have been shown to be reasonably accurate when compared to other techniques, such as spectrographic analysis or cepstral analysis (Markel and Gray, 1976). The reliability of other measurements, such as fundamental frequency, have been evaluated by making estimates of synthetic speech with predetermined values. The results indicated that the estimates were accurate to within 1 to 2% of the actual values (Levitt, 1971). However, the accuracy of these methods as applied to <u>real</u> speech is unknown.

CONCLUSIONS

A review of the literature on speech perception by normal-hearing listeners has shown the importance of acoustic characteristics for identification and discrimination of the various speech sounds. One problem encountered with a majority of these studies, however, is the widespread use of synthetic speech stimuli containing a minimum of cues which may not accurately represent the multidimensionality of the stimuli. In addition, the efficiency of the auditory system has, for the most part, been overlooked. Comprehensive analyses of confusion matrices have, thus far, included only comparisons of certain groups of phonemes (fricatives with fricatives, plosives with plosives). There is not, as yet, a predictive model which could be used to predict the full confusion matrix under different listening conditions.

The various factors affecting overall speech intelligibility have also been described, with a great deal of attention placed on the effects of filtering and additive noise. Computational procedures such as the Articulation Index, and the use of speech-envelope noise have

emphasized the significance of speech-to-noise ratio measurements in assessing the speech-interfering characteristics of noise.

A number of studies have attempted to quantify the effects of noise on consonant and vowel intelligibility by estimating the importance of so-called articulatory or phonological features. A review of this research uncovers significant differences between studies as to the relative importance of these features. Lack of agreement has been shown for such basic measurements as the relative intelligibility of voiced vs. voiceless phonemes, and syllable-initial vs. syllable-final consonants. Significant effects of background noise spectra have been shown. The lack of independence of the features and the differential effects of noise on their importance makes it difficult to determine the significance of these analyses.

Many recent investigations with hearing-impaired listeners have, for the most part, been based on a simplified model of the effects of hearing impairment. This description, in which hearing loss is thought to act as a filter, fails to take into account the significant distor-

tion effects which have been shown to be present in the auditory system of these listeners.

In conclusion, the review of the literature reveals that there is a need for further investigation of the acoustical factors involved in phoneme reception. A detailed acoustical analysis of the speech signal and a precise measurement of speech-to-noise ratio may be obtained through the use of digital speech processing techniques described herein. From these measurements, predictions could be made as to the types of confusions one would expect in various noise conditions.

CHAPTER III

EXPERIMENTAL PROCEDURES

The primary purpose of this investigation was to attempt to predict from acoustical measurements, the pattern of perceptual confusions made under difficult listening conditions (e.g., against a noisy background). The procedures used to accomplish this consisted of the selection of a set of speech stimuli and listening conditions; the compilation of normative data, including confusion matrices, with these stimuli; and the measurement of selected acoustical parameters.

This chapter has been divided into two major sections. First, the subjects used in the listening experiment are described and the materials and apparatus used in both the listening experiment and subsequent acoustical analyses are delineated. The second section contains (i) a description of the experimental design of the listening experiment, (ii) a description of the procedures used to obtain the computer-generated spectra of the speech and noise stimuli, and (iii) a description of the techniques used in obtaining the acoustical measurements.

SUBJECTS

The subjects were 2 males and 4 females ranging in age from 23 to 28 years. All subjects had hearing thresholds, for each ear, within 10 dB of standard zero reference levels (ANSI, 1969) for octave pure tones between 250 and 8000 Hz. All subjects were familiar with the type of stimuli and format used in the listening experiment.

TEST MATERIALS

NONSENSE SYLLABLES

For the purpose of this study, it was necessary to have a set of speech stimuli which would allow for a detailed analysis of patterns of consonant confusions made by normal-hearing subjects under difficult listening conditions. The Nonsense Syllable Test, developed by Resnick, Dubno, Hoffnung, and Levitt (1975), was chosen as the basic test instrument for the following reasons: (i) the test was designed to include all of the major consonants of English. It was structured using a closed-response-set format such that nearly all of the most common perceptual confusions are available choices for the listener; (ii) the size of

each response set was not unwieldy, there being between 7 and 9 foils for each subtest; and (iii) the subtests contain a high proportion of sounds which are known to be difficult for the hearing-impaired listener (or hearing-aid user), a characteristic which may be important for later analysis with these two groups. In addition, high quality recordings of the test were available for immediate use.

The Nonsense Syllable Test (NST) has a modular format in that it is made up of a set of interchangeable subtests. Seven of these subtests were used in a major ongoing study (Levitt, Collins, Dubno, Resnick, and White, 1978). This is the most commonly used version of the NST and is referred to as the basic NST. For this study, data were also collected on an additional 4 subtests. These subtests are referred to as the Optional Subtests (OST).

The subtests of the NST (7 basic and 4 optional) consist of consonant-vowel (CV) and vowel-consonent (VC) nonsense syllables organized into sets of 7 to 9 syllables each. The 11 subtests differ in terms of three factors: (i) class of consonant (voiced or voiceless), (ii) position of consonant (initial or final), and (iii) vowel con-

text (/ɑ/, /i/, or /u/). Listener responses (i.e., the response foils) are limited to syllables within the same subtest. Each subtest contains one repeat item. The response foils were chosen so as to limit the response set as little as possible, but still be of manageable size. Thus, the response foils were selected so as to include the perceptual confusions which occur most frequently. Perceptual confusions involving a voicing error are among the least common for hearing-impaired listeners and hence the response foils were limited to consonants having the same voicing feature (i.e., voiced or voiceless). This limitation, which was a practical compromise, led to response sets of manageable size yet allowed for errors of all kinds (except those involving voicing).

Figure 4 provides a list of the response sets used in each subtest. Subtests 1, 2, and 3 contain final voiceless consonants in the three vowel contexts. Subtests 5, 8, and 9 contain initial voiceless consonants with the three vowels. Final voiced consonants with /ɑ/, /i/, and /u/ are represented in subtests 4, 10, and 11. Subtests 6 and 7 contain initial voiced consonants in the /ɑ/ context only.

Fig. 4. The test items making up the 11 subtests of nonsense syllables.

1	2	3	4	5	6	7	8	9	10	11
af	uθ	iʃ	ab	fa	la	na	pu	θi	iz	uŋ
aʃ	up	if	að	ta	ba	va	ʃu	pi	ib	uz
at	us	it	ad	pa	da	ma	tu	ʃi	id	un
ak	uk	ik	am	ha	ga	za	fu	ki	im	uv
as	ut	is	az	θa	ra	ga	su	hi	in	uð
ap	uf	iθ	ag	ʧa	ja	ba	hu	ʧi	iv	ub
aθ	uʃ	ip	an	sa	ʤa	ða	θu	si	ig	um
			aŋ	ʃa	wa	da	ʧu	ti	in	ud
			av	ka			ku	fi	ið	ug

Two subtests were needed in order to include the semivowels and the initial /r/, within response sets of practical size. The 11 subtests thus provided reasonably comprehensive coverage of the factors considered, voicing, consonant position, and vowel environment. The major limitations are that initial voiced consonants were split into two subtests and voiced CV's with /i/ and /u/ were not included.

In the recorded version of the test, the syllables are presented in the carrier phrase "You will mark _____, please". The test lists were recorded by both a male and female speaker.

BACKGROUND NOISE

A sample of background noise was recorded by Levitt, et al. (1978) for use with the NST. The recording was made in a large, carpeted cafeteria, edited to remove transients and intelligible conversation, and equalized for level. This competing noise was then dubbed onto the second track of all nonsense syllable test tapes. A detailed description of the recording, editing, and equalizing procedures, including the preparation of test tapes for both the nonsense syllables and the background noise,

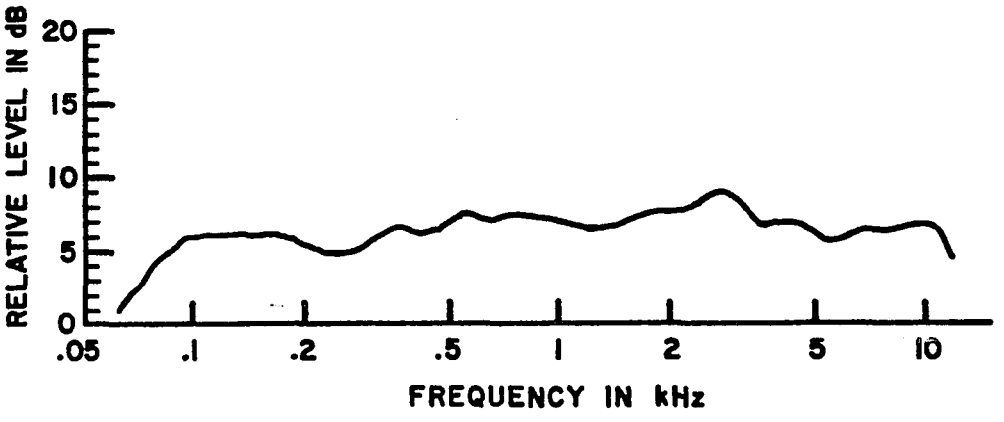
is given in the next section.

PREPARATION OF ORIGINAL RECORDINGS

Nonsense Syllables

The 11 sets of nonsense syllables were recorded by a male and female speaker, both of whom were judged by a skilled dialectician to have no marked regional accents. The original recordings were made by Levitt, et al. (1978) in a double-walled sound-attenuating booth (Industrial Acoustics Corporation). Measurements of octave-band noise level revealed levels of background noise to be less than 15 dB SPL between 250 and 8000 Hz, and 25 dB SPL in the 125 Hz band. A microphone (AKG Model D202E) was placed on a covered table situated in the center of the room and positioned approximately 5 inches below and 12 inches in front of the speaker's lips. The frequency response of the microphone is shown in Figure 5. A sound level meter (Bruel & Kjaer Model 2203) was placed behind the microphone and was used by the speakers to monitor extreme variations in level, during both the rehersal and recording sessions. The tape recorder (Nagra Model IVS) was run at 15 inches/second with

Fig. 5. Frequency response of the microphone used to record nonsense syllables and background noise.


RELATIVE LEVEL IN dB
20
15
10
5
0
.05
.1
.2
.5
1
2
5
10
FREQUENCY IN kHz

Scotch type 206 audio tape.

Background Noise

The original sample of background noise was produced in a large, crowded cafeteria room which had hard walls and carpeted floors. The recording was made using the identical microphone and tape recorder as in the nonsense syllable recordings. The result was a 20 minute sample which was to be edited at a later time.

MEASUREMENT OF SPEECH AND NOISE LEVELS

Nonsense Syllables

All tapes were checked by Levitt, et al. (1978) by making graphic level recordings (Bruel & Kjaer, Graphic Level Recorder Model 2305), which were used to measure speech peaks. The instrument was adjusted to a pen speed of 100 mm/sec (or approximately 50 dB/sec), and a paper speed of 10 mm/sec. The input signal and level recorder attenuators were set for a maximum pen excursion of 20 mm. With those settings, the tracings contained peaks which were within 10-20 dB above the zero line.

The syllables appear within the carrier phrase

"You will mark _____, please". It was decided that measurements were to be made which would permit equalization of the peak levels of the word "mark" for each syllable. The tape was then copied using selective attenuation so that the level of each "mark" would be the same throughout the tape. Note that the level of the test syllable varied over a markedly wide range, as would be the case in actual speech. A 400 Hz calibration tone was then recorded at the level which corresponded to the average equalized peaks of the word "mark". The recording was then rechecked aurally and a final graphic level recording produced to assure that the "mark" levels were equal (within $\pm$ 1.5 dB). The recordings made in this way, and provided they satisfied both of the above checks, became the <u>master</u> recordings of the nonsense syllables.

<u>Background Noise</u>

Editing of the noise tape was necessary prior to measurement of noise level. First, all transients and portions of intelligible conversation were spliced out. After a graphic level recording was made, all sections of tape containing peaks more than 3 dB greater than the mean

level were removed. The tape was checked aurally to make certain that the splices were not audible and did not disturb the continuous nature of the noise. A final 2 minute tape was produced with variations in level which did not exceed $\pm$ 1.5 dB. A tape loop was prepared so that a longer master noise tape could be produced. A 400 Hz calibration tone was added to the noise tape, its level equal to the average RMS level of the noise. As in the nonsense syllable tape, the final master tape (with calibration tone) was then checked aurally, and a graphic level recording produced as a final check on average levels.

PREPARATION OF TAPES

Listening Experiment Tapes

The original master recordings of Levitt, et al. (1978) were produced so that various protocols (orderings) could be prepared. There are 8 male and 8 female protocols of the 7 subtests making up the basic NST, and 6 male and 6 female orderings of the 4 optional subtests. The final protocols were a result of two independent randomizations. First, the subtest order was randomized. (For

the basic NST, 7 subtests were randomized). In addition, the order of presentation of syllables within each subtest was randomized. The dubbings necessary to produce these protocols were prepared on one of four matched tape recorders (two Ampex Model 500, Ampex Model 440B, Nagra Model IVS). In addition, subtest identifiers ("page one, page two",* etc.) were recorded and duplicated, as well as sufficient durations of tape containing the calibration tones. A speech sample (used for setting most comfortable loudness, if needed) was prepared for insertion after the calibration tone and before the beginning of the protocol. This segment was not used in the collection of the data for the present study.

The subtests were prepared in the following manner. The phrases were spliced and arranged in their predetermined order, separated by 5 seconds of timing tape, with the appropriate identifier at the beginning of each subtest. The subtests were then arranged in their assigned

---

*The subtests were identified in this way since the listener used a response booklet where each page corresponded to one subtest.

order and spliced together to form the protocol. The calibration tone and speech sample segments were then added to the start of the tape. These recordings of the 16 protocols were then dubbed again with the addition of the background noise to the second track of the tape. The 16 dubbed recordings (with the added noise on the lower track) are referred to as submasters. Because each of these 16 submasters contains a different ordering, each syllable is heard in a different noise context. The 16 submasters were then dubbed to produce 16 test tapes, each of which was carefully screened aurally and by graphic level recording before use. Only the test tapes were used in the subsequent experimental work. The submasters are used only for the purpose of preparing new test tapes.

Spectral Analysis Tape

A number of acoustical parameters selected for measurement required precise spectral analysis. Therefore, computer-generated spectra were prepared. In order to digitize the tape for computer analysis, a tape was prepared in which only the test syllables were recorded, the carrier phrases having been deleted. For convenience,

the syllables were spaced one second apart.

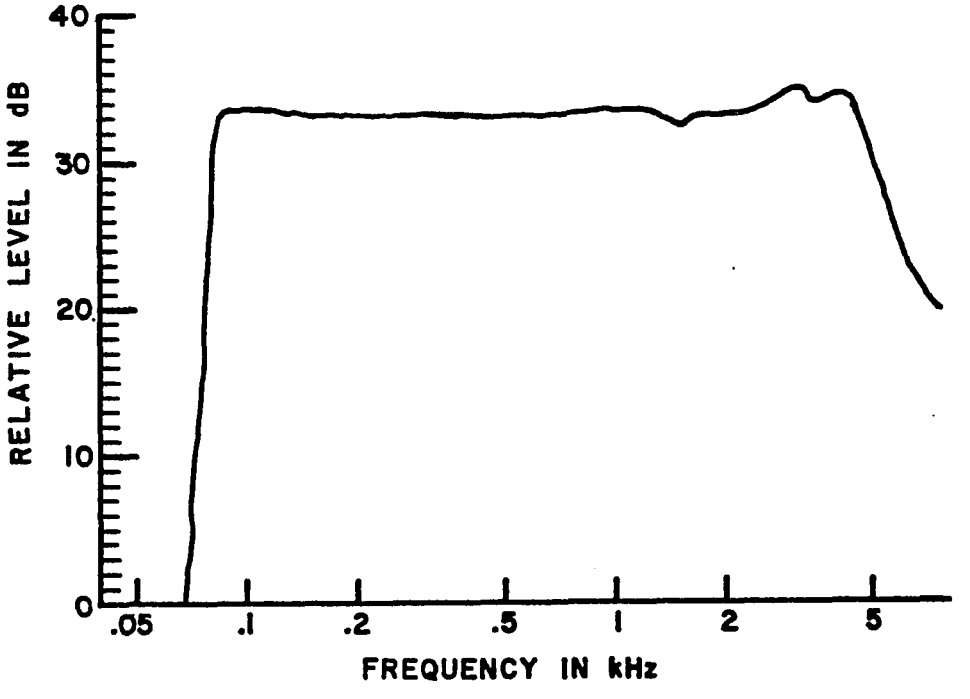
TEST APPARATUS

LISTENING EXPERIMENT

The nonsense syllable test tape was played at 7.5 inches/second on a two-channel Ampex Model 500 tape recorder. The output was fed through an audiometer (Grason-Stadler Model 1701) which was used for controlling, separately, the levels of both the speech and noise signals. The audiometer was calibrated to ANSI standards (ANSI S3.6-1969). The 2 signals were mixed and the output transduced by one of a matched pair of TDH-49 earphones mounted in MX-41/AR cushions. Frequency-response curves for the matched earphones were obtained using a 6 cc coupler (NBS Type 9A) and a constant current source. One curve, shown in Figure 6, shows a relatively flat ($\pm$ 3 dB) frequency response between 90 and 5000 Hz. All listening was done in a double-walled sound-attenuating suite (IAC).

Calibration of the test apparatus was performed prior to every test session. The calibration tones recorded at the beginning of each test tape served as the

Fig. 6. Frequency response of the TDH-49 earphone used in the listening experiment.


RELATIVE LEVEL IN dB
40
30
20
10
0
.05
.1
.2
.5
1
2
5
FREQUENCY IN kHz

test signals for both speech and noise stimuli. Both channels of the tape recorder were set at a specific VU reading (-5 VU), and a similar adjustment was made on the audiometer VU meter. The speech-to-noise ratio was adjusted using the audiometer attenuator dials and fine adjustments made using the audiometer calibration control knobs and monitored on the VU meter. The output from the earphone was then measured using a Bruel & Kjaer Sound Level Meter (Model 2203) using the standard 6 cc coupler. The same earphone was used as the load in checking the overall calibration of the electrical system. The accuracy of the electrical calibration was within $\pm$ 0.5 dB. The tape recorder, whose frequency response conformed to the NAB standard, was periodically checked, its heads cleaned and demagnetized.

ACOUSTICAL MEASUREMENTS

The pieces of equipment used during acoustical analyses were: (i) a sound spectrograph, (ii) an analog-digital converter system to convert the two-channel analog tape to digital form, (iii) a large, general-purpose computer which produced the speech and noise spectra from the digitized signals, and (iv) an interactive laboratory computer

and peripheral measurement devices which were used to make consonant and vowel duration measurements.

Sound Spectrograph

A sound spectrogram was made of each of the 72 individual syllables included in the acoustical measurement analysis. The tape recording consisting of the syllables only (the spectral analysis tape) served as the stimulus source. An Ampex Model 500 tape recorder, set to the same level as in the listening experiment (-5 VU) was used to deliver the speech signals to a Kay Sona-graph (Model 6061-B), set for total bandwidth of 80 to 8000 Hz. A wide-band spectrogram, with a 500 Hz calibration signal, was then generated for each syllable.

Analog-to-Digital Conversion

The two-track tape recording of the syllables, prepared as previously described, was low-pass filtered using a filter having an attenuation rate of 48 dB/octave and an 8 kHz cut-off frequency. The signals were then sampled at a frequency of 16.667 kHz per channel, using a 16-bit analog-to-digital converter. Prior to sampling, the level of the speech signal was adjusted such that the

peaks of the signal were comfortably within the dynamic range of the digitizer. The sampled speech signals were then transferred to digital tape for storage.

A large, general-purpose digital computer (IBM 360/168) was used to perform the spectrum analysis on the digitized signals. It is important that the parameters used in this analysis (which used the method of linear predictive coding) be carefully chosen. These parameters include sampling rate, number of filter coefficients, analysis interval, and window type. The effects of these various parameters were discussed previously in the Review of Literature chapter. The remainder of this section is devoted to a description of the parameters used in the digital spectrum analysis.

Sampling Rate

Almost all of the important spectrum information in speech lies below 10 kHz and a commonly used sampling rate for speech signals is 20 kHz. However, due to space limitations on the two-channel (speech and noise) system, the signals for analysis in this study were sampled at a rate of 16,667 samples per second (per channel), after

having been low-pass filtered to 8 kHz.

Number of Filter Coefficients in Linear Predictive Coding

In general, 2 filter coefficients (one pole-pair) are used to represent each resonance of interest. Thus, to ensure that all of the important resonances over the entire frequency range are included, the present analyses were performed using 10 pole-pairs. This represents a linear predictive filter having 20 coefficients.

Analysis Interval

In choosing the analysis time window, recall that the time-frequency trading relationship must be taken into account. A compromise selection was made and the analysis performed using a 30.72 msec window; this corresponds to 512 samples spaced 0.06 msec apart in time, the reciprocal of 0.06 msec being the sampling frequency (16.667 kHz) of the resulting frequency analysis. The frequency resolution is equal to the sampling rate (16.667 kHz) divided by the number of samples (512) in the time window, i.e., 32.55 Hz.

Windowing

As described in the Review of Literature, the use

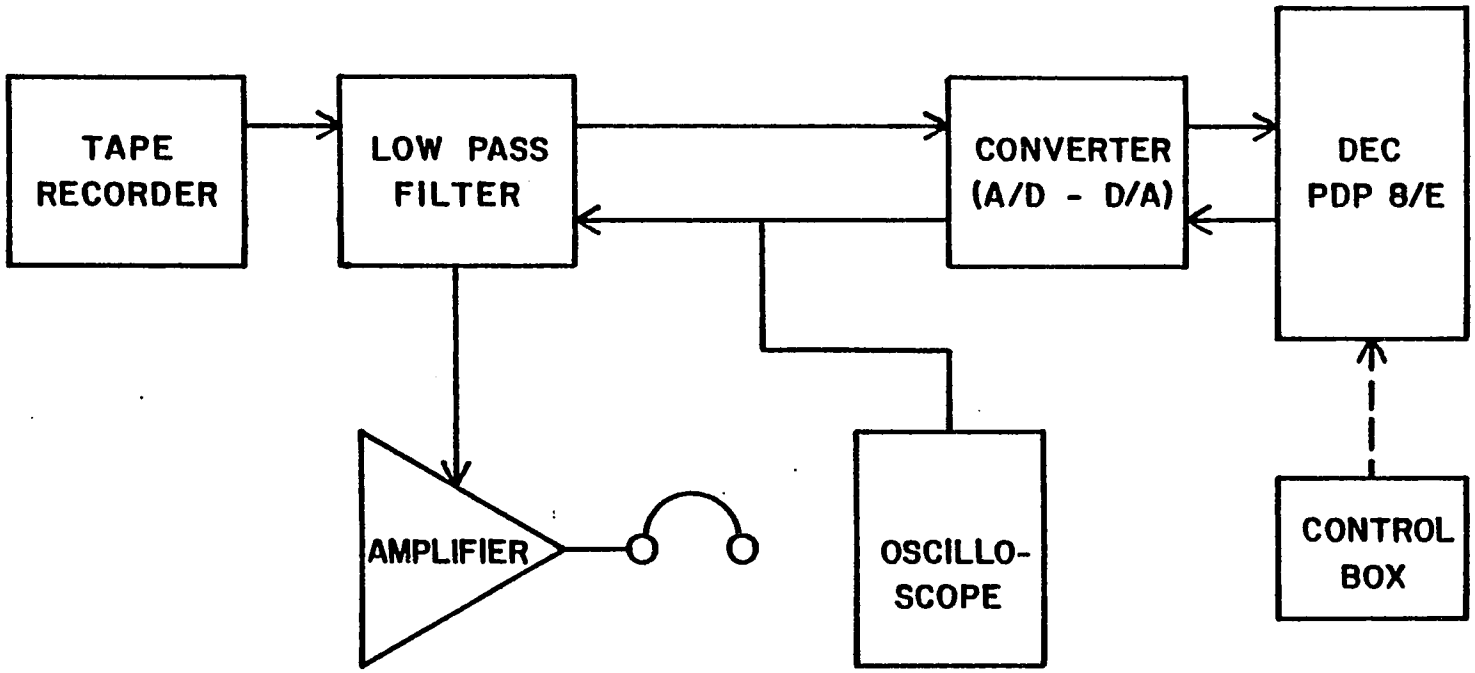
of a weighted time window is advantageous for analysis of a window of this size. The analysis in this study employed a Hamming window (Blackman and Tukey, 1958).

Interactive Computer System

Measurements of the duration of consonants and vowels were obtained by the use of an interactive computer system. The system was composed of a DEC PDP 8/E computer (core memory of 16K 12 bit words) along with the necessary measurement devices. The following analog equipment was used in conjunction with the computer: tape recorder (Ampex Model 500), earphones (TDH-49), filter (Allison A124B, variable filter), oscilloscope (Tektronix Type RM503), keyboard (Digital DEC Scope), and control box (custom-built). The digital device was a disk unit (DEC RKØ-E cartridge, 1.6 million 12 bit words). A block diagram of the interactive computer system is shown in Figure 7.

The same type used for the previously described digital analysis was used with this system. One syllable at a time was read into the computer and conversion to digital form was carried out by means of an analog-to-digital converter. The speech signal was first low-pass fil-

Fig. 7. Block diagram of the interactive computer system.


TAPE RECORDER
LOW PASS FILTER
CONVERTER (A/D - D/A)
DEC PDP 8/E
AMPLIFIER
OSCILLO-SCOPE
CONTROL BOX

tered to approximately 9 kHz and then sampled at a frequency of 20 kHz using a 12-bit quantizer.

The digitized syllable was then stored on the disk, played back and monitored through earphones while its waveform was viewed on the oscilloscope. A continuous digital-to-analog conversion of the signal allowed for a segment of the speech to be played back repeatedly. The length of the played-back observation interval was adjusted by means of a manual adjustment on the control box which manipulated the start and stop times of the interval independently. Thus, the duration of the speech segment could be measured by calculating the length of the observation interval containing the segment in question. The control box contained three knobs for controlling the start time and three knobs for controlling the stop time, permitting duration adjustments in steps of 0.05 msec, 15 msec, or 3 seconds. While observation-interval adjustments were made with the control box, the played-back portion of the syllable could be heard and viewed on the oscilloscope simultaneously. As the appropriate utterance (consonant or vowel) was adjusted, the start and stop times corresponding to each of

the knob settings were displayed on the monitor. The rules for deciding on relative durations are described in the Procedures section to follow.

PROCEDURES

LISTENING EXPERIMENT

The 11 sets of nonsense syllables served as the speech stimuli and the cafeteria noise, recorded on the second track of the test tapes, was used when background noise was required. Two parameters were varied: level of the speech signal, and speech-to-noise ratio (both the speech and the noise levels measured according to the procedures described in the previous section). Signal levels were chosen to provide a wide range of scores. Preliminary data revealed that complete articulation functions for these materials could be generated by using 5 signal levels: 20, 28, 36, 44 and 52 dB SPL. In order to compare consonant confusion patterns obtained in quiet with those obtained in noise, the same 5 signal levels were used, first without background noise (quiet condition) and again with the background noise set to a 5 dB speech-to-noise ratio.

The data were collected in two phases. First, 10 conditions (5 signal levels x 2 noise conditions) were used with both male- and female-speaker nonsense syllable lists, for a total of 20 listening conditions. The 20 conditions were randomized, as were the male- and female-speaker nonsense syllable protocols. This design was used to obtain data on the 7 basic subsets of nonsense syllables. Analysis of the preliminary data revealed significant random guessing at the 20 dB SPL signal level and nearly perfect scores (i.e., very few confusions) at the 52 dB SPL signal level. Therefore, the second phase of data collection, using the final 4 subsets, was modified to include only the 28, 36, and 44 dB signal levels. All other parameters remained the same, for a total of 12 listening conditions (3 signal levels x 2 noise conditions x 2 speakers). After obtaining the articulation-gain functions, the data obtained with the male speaker were then analyzed in greater detail in terms of the patterns of perceptual confusions and their acoustical correlates.

The subject was seated in the test suite and was given a test booklet for the first condition. The book-

let contained one page for each of the subtests, arranged in the order that they occur for that particular protocol. A sample page of a test booklet (for subtest 3) appears in Figure 8. For each of the target syllables, /ip, is, ik, iθ, it, if, iʃ/, responses are limited to that set of syllables. Subjects were instructed to guess when necessary.

After all of the subtests were completed, the tape was changed, the test equipment adjusted for the next condition, and a new test booklet provided which corresponded to the next protocol to be used. Subjects were tested individually, usually listening to 5 or 6 tapes per session.

Upon completion of all listening conditions for all subjects, responses were keypunched and prepared for data analysis.

ACOUSTICAL MEASUREMENTS

Introduction

Most of the acoustical measurements were made from the computer-generated speech and noise spectra. The computer output consisted of a graphic representation of the spectra of the speech and noise signals for each analysis

Fig. 8. A sample page of a nonsense syllable test booklet (subtest 3).

PAGE _____ NAME _____

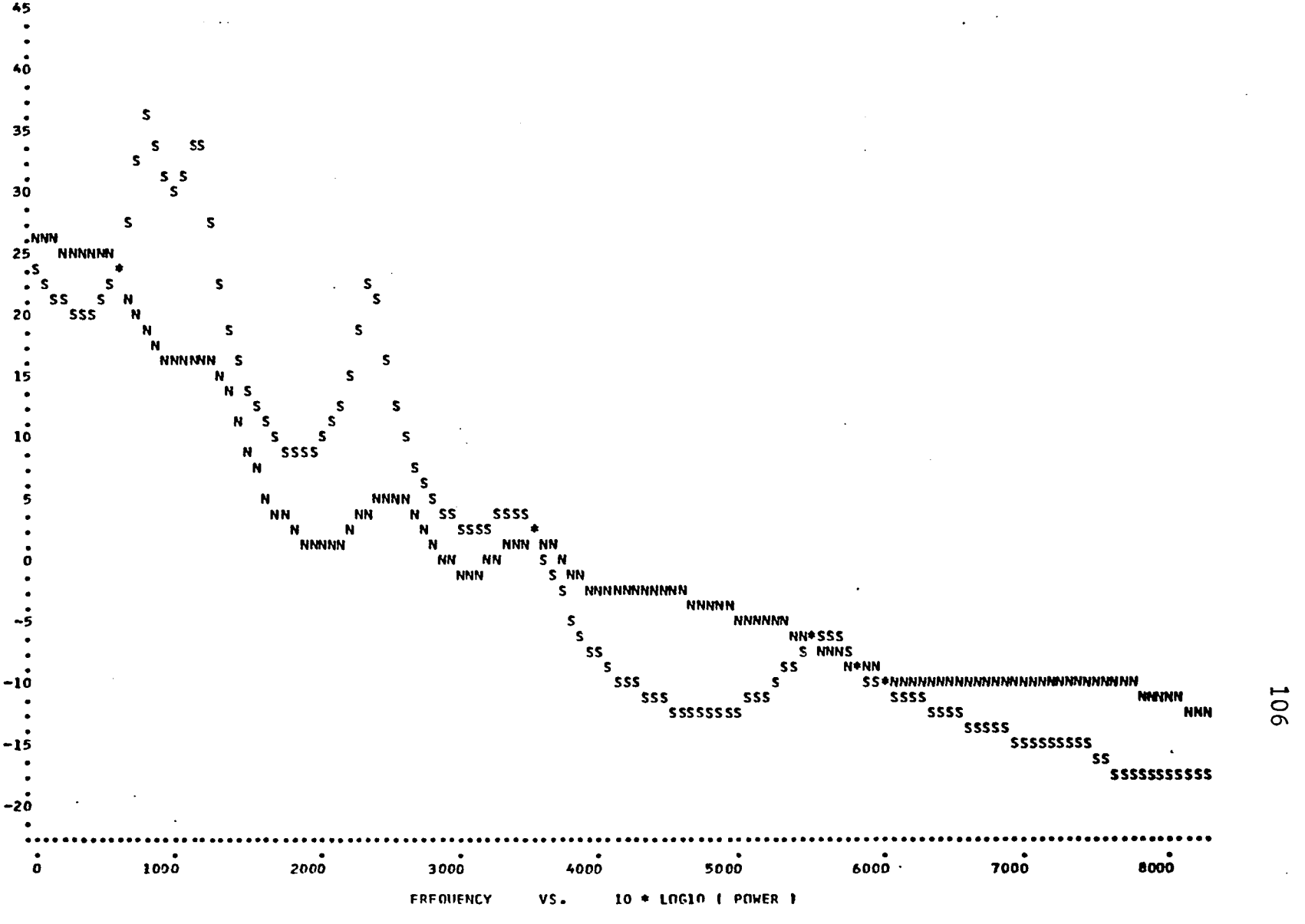
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH
EEP	EES	EEK	EETH	EET	EEF	EESH

interval (30.72 msec), printed on the same page. A typical output page (for a portion of the vowel in the syllable /aθ/) is shown in Figure 9. The spectra of one syllable and its accompanying noise requires about 10 to 15 pages of computer output, depending upon the syllable's duration. Along with the spectral representation, the program generates RMS amplitude and frequency measurements for the peaks in the spectra for each time interval. These numbers, in addition to the graphical representation of the spectra themselves, served as the source for the majority of the acoustical measurements made on the syllables and noise. The interactive computer system was used for the measurement of duration. A description of the procedures used in obtaining the acoustical measurements appears in the following sections. Any calculations or statistical analyses performed on these measurements are discussed in the Results chapter.

Vowel Formant Frequencies

The computer-generated spectra provided information to derive the frequencies of the peaks in the spectrum of the speech for each analysis interval. In order to deter-

Fig. 9. A typical page of computer output for speech (S) and noise (N) spectra. A 30.72 msec portion of the vowel in the syllable /ɑθ/ and its accompanying noise have been analyzed by linear predictive coding.


TIME = 1290.24 MSEC. (SIGNAL RMS = 1198.)
45
40
35
30
25
20
15
10
5
0
-5
-10
-15
-20
0
1000
2000
3000
4000
5000
6000
7000
8000
FREQUENCY VS. 10 * LOG10 (POWER)

mine the formant frequencies of the vowels in the CV and VC syllables, the following procedures were adopted. First, the time interval corresponding to the vowel portion of the syllable was determined. The sound spectrogram of the syllable was used as a visual aid. Next, the formant peaks in the spectrum for that time interval were found. The numerical values (in Hz) corresponding to those peaks were then read from the computer output. The frequencies of the first four peaks (F1, F2, F3, F4) were noted, and a check of these estimates was made by visual inspection of the spectrogram. In addition, the amplitude of the highest of the first four formants (F1 or F2, in all cases) was noted.

Second Formant Transition

A number of measurements were made on the second formant of the vowel. The time intervals corresponding to the transition period were found. This entailed careful inspection of the spectral peaks, the numerical values of those peaks, and the spectrogram of the syllable. For example, for a VC syllable, the transition period extended from the time at which the first change in the steady-state formant frequency was observed until the time corresponding

to the onset of the consonant. The second formant frequency for each of these time intervals was then noted. From these data, the magnitude and direction (i.e., rising or falling) of the formant transition was determined. The magnitude was specified as the change in frequency, in Hz, from the start of the transition to the end of the transition. For a small number of syllables (usually nasals) no change in the second formant frequency could be found, either from the computer analysis or by spectrographic inspection.

The reliability of these measurements was checked, whenever possible, by making the identical measurements from the repeat item in the subtest. This item is an exact duplicate (dubbing) of one of the syllables in the set, not a repeated utterance by the speaker. Thus, the acoustical measurements should be the same.

It is important to keep in mind the limitations imposed by the frequency resolution of the short-term digital spectrum analysis. The resolution for these analysis parameters was calculated to be 32.55 Hz (see previous section). That is, the formant-frequency measurements are accurate to within approximately 33 Hz. Larger changes in frequency

(formant transitions) are interpreted as occurring in multiples of 33 Hz. An additional problem was encountered with a small number of syllables (specifically with the vowel /u/). When formant frequencies lie close to one another (as in /u/), the formant peaks may merge and be interpreted as a single peak in the spectrum. When this occurred, the linear predictive coding procedure was repeated for those syllables, using a filter having a higher number of coefficients (24 coefficients, in this case). This method was successful in separating out the formant peaks in question.

Consonant Spectral Peaks

To determine the spectral peaks in the noise (fricative) portion of the consonants, a procedure similar to the one described for the vowel formants was used. The analysis intervals corresponding to the consonant portion of the syllable were determined. Again, the spectrogram of the syllable was used as a visual aid. The first and second highest peaks during this time period were noted on the computer spectra, and a check made by visual inspection of the spectrogram. The frequencies corresponding to these peaks were then read from the computer output.

Overall Consonant-Noise Bandwidth

The overall bandwidth of the frication portion of a consonant was measured. The overall bandwidth was defined as the difference between the upper and lower cut-off frequencies, which in turn were defined as the frequencies at which the falling spectrum was 3 dB below the average level of the main body of the power spectrum.

First, the time interval was selected which, by inspection, was found to contain the "broadest" consonant energy. Then a visual estimate of the main body of the power spectrum was made and the bandwidth of this energy was measured. This procedure, although involving a subjective judgment as to what constituted the main body of the power spectrum, seemed to work quite well. In all cases, a post hoc analysis revealed that the estimated overall bandwidth did, in fact, contain the previously-determined spectral peaks. In addition, the bandwidth measurements for each of the 72 syllables were performed twice, and the results of the two independent measurements showed only small differences, the largest being $\pm$ 25 Hz.

Total Energy of Consonant and Vowel

Measures were made of the total energy in the consonant portion and vowel portion of each syllable. These measures were obtained by a computer averaging technique performed on the estimated time interval corresponding to the individual consonant and vowel. To accomplish this, the precise start and stop times for each consonant and vowel had to be determined. This procedure was limited by the time resolution of the analysis. The analysis of the continuous speech waveform was made using consecutive 30.72 msec time windows, with no overlap or gaps. Therefore, any changes in level within this time period would be smeared by the analysis. With this in mind, the following procedure for determining the onset and offset of consonants and vowels was instituted.

The onset of the consonant in a CV syllable and of the vowel in a VC syllable was relatively simple to obtain. The time corresponding to the first window containing an abrupt change from the baseline level was taken as the onset of the syllable. For VC syllables, this first window usually showed the formant structure of the vowel; for CV

syllables, this first window usually showed a sharp change in overall spectrum level, without necessarily showing any formant structure. On rare occasions, a tape splicing click was detected in an early window, but this was clearly differentiated from the syllable onset by spectrographic inspection. Determining the stop time of a consonant in a CV syllable or a vowel in a VC syllable was more difficult. The C-V or V-C boundary point on the spectrum was usually characterized by a large difference in the amplitude between one window and the next. For example, for the syllable /ɑθ/, the last window showing evidence of /ɑ/ had a great deal more energy that the next window, signifying the onset of /θ/. Thus, the V-C boundary in this case was judged to be between those two windows.

The offset of the consonants in VC syllables was defined as the time corresponding to the last window with an amplitude greater than the background noise of the system. The offset of the vowel of a CV syllable was the time of the last window showing evidence of formant structure. For the stop consonants, the "stop closure" duration was included as part of the consonant duration, for these energy

measurements only. For example, for the syllable /ip/, the time between the offset of /i/ and the onset of /p/ was included with the consonant. The "stop closure" durations were considered separately in specifying the durational characteristics of the syllables (as defined below).

The result of these measurements was a tabulation of the onset and offset times for each vowel and each consonant of every syllable studied. The energy contained within these designated time periods for both the speech and the noise was then computed by summing the log of the energy in all of the windows within each time period. The results are specified in dB using as references the consonant and vowel having the lowest total energy. Thus, the consonants are measured in dB re: /hi/, and the vowels are measured in dB re: /iθ/. Consonant-to-noise ratios and vowel-to-noise ratios (in dB) were also derived.

Duration

Due to the relatively gross time resolution of the spectrum analysis, fine timing measurements, such as durations, could not be made from the computer-generated spectra. Thus, an interactive computer system, as described

in the previous section, was used. In this case, a control box was used to manually adjust the start and stop times of each speech segment. The time resolution of these adjustments, using the fine adjustment knob, is equal to the sampling interval, which in this case was 0.05 msec. (The sampling rate for this system was set at 20 kHz, resulting in intervals between samples of 1/20,000 or 0.05 msec.)

Measurements were made of the duration of the consonant, vowel, and stop closure (when applicable) for each of the 72 syllables under study. The procedures for these measurements were as follows. Immediately after the syllable was read into the computer, it was repeatedly displayed on the oscilloscope and simultaneously heard through the earphones. The onsets of consonants in CV syllables and vowels in VC syllables were relatively simple to determine, because they are preceded by near silence. Each of these points was defined as the time corresponding to the abrupt change in the waveform.

The consonant-vowel and vowel-consonant boundary was more difficult to determine. With the entire syllable appearing on the oscilloscope (/zɑ/ for example), the stop

time was held constant and fine adjustments were made in the start time until the abrupt change in the waveform was reached. This point (determined aurally and visually) corresponded to the onset of /z/. To determine the duration of /z/, all evidence of /ɑ/ was removed. This was done by adjusting the stop time, while keeping the start time constant. The stop time fine adjustment knob was carefully rotated, crossing the C-V boundary several times, until the precise point was established which represented the end of the consonant /z/ and the beginning of the vowel /ɑ/. With these knob settings, the entire portion of /z/ could be viewed on the oscilloscope and simultaneously heard through the earphones. The start and stop times corresponding to those knob settings were then recorded and the duration calculated by subtracting one from the other. The entire syllable was then viewed again and similar adjustments made to determine the duration of /ɑ/. The stop time of the vowel /ɑ/ was the time corresponding to the crossing of an arbitrary baseline on the oscilloscope. This rule worked well for determining the stop time of most consonants in VC syllables and all vowels in CV syl-

lables. Other, more abrupt offsets, such as the /k/ in /ik/ were recorded as the time corresponding to the abrupt reduction in the waveform.

The start and stop times for the stop closure durations were determined in a similar way. These measures were relatively simple as the boundaries in both cases are to and from near silence.

The method described above appears to provide estimates of duration which are at least as accurate and reliable as one could expect from spectrographic inspection procedures. It has been shown that the use of spectrograms or digital speech waveforms can yield reliable measurements of vowel duration (Allen, 1978). However, the validity of spectrographic estimates, specifically in terms of criteria for determining segmental boundaries, is questionable. Peterson and Lehiste (1960) state that it is possible to make duration measurements from spectrograms that are reliable to within one or two centiseconds. However, they emphasize that "instrumentational accuracy is in general considerably greater than the accuracy with which the segmental boundaries can be determined" (ibid.,

p. 694). Overlapping cues between consonants and vowels make such judgments difficult. Thus, the spectrogram may not necessarily represent the acoustic signal in the form in which it is received at the perceptual level. A particular segment of the spectrogram may not correspond precisely to a particular phonetic segment. The measurement of duration with auditory and visual cues may provide a better approximation of the signal in its encoded form.

In spite of the limitations of the spectrograms, they were used as a check on the duration measurements. Vertical lines were drawn at the estimated onset, offset, and boundary points. Next, all consonant segments were measured by ruler and the syllables rank ordered from shortest to longest consonant. This ranking was then compared to an identical consonant ranking based on the computer measurements. This procedure was repeated for vowels and for stop closure durations. The agreement between the two orderings was excellent. Segments with very similar durations (within 10 msec) were occasionally placed in different orders by the two methods.

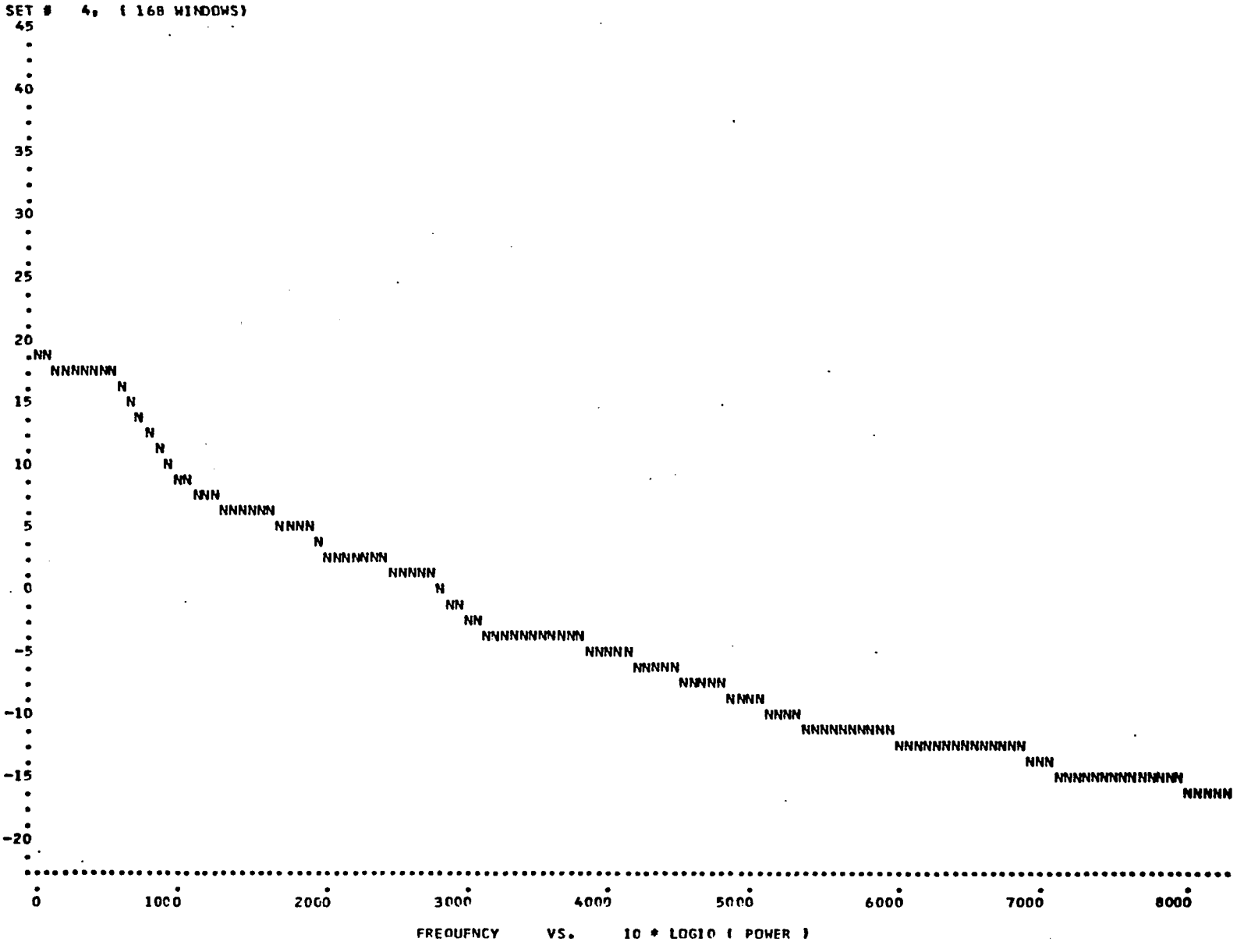
Noise Spectrum

Representations of the spectra of the background noise were obtained. Precise start and stop times for the noise accompanying each syllable were determined using the same method that was used for the total energy measurements. These times correspond to the total duration of each syllalbe. The spectra of the noise for each analysis interval was determined using the same number of filter coefficients as in previous analyses (20). A computer averaging technique was then performed on the estimated time interval corresponding to each noise segment. The result of this analysis was one spectral representation of the noise for each syllable. The spectra of the noise for each syllable within a subtest was then averaged resulting in one spectral representation for each subtest. The spectrum of the noise accompanying the syllables in subtest 4 is shown in Figure 10.

The spectra were then traced and compared segment-by-segment and were found to vary by not more than 5 dB across the entire frequency range. Variation between spectra averaged by subtest was less than 2 dB. Based on these

Fig. 10. Long-term averaged spectrum for background noise accompanying subset 4. Data contained in 168 windows of 30.72 msec length were averaged after linear predictive coding analysis. Note that plotting is on a linear frequency scale. A 5 dB reduction in amplitude was included to simulate the speech-to-noise ratio used in the listening experiment.

NOISE -5


SET # 4, (168 WINDOWS)
45
40
35
30
25
20
15
10
5
0
-5
-10
-15
-20
0
1000
2000
3000
4000
5000
6000
7000
8000
FREQUENCY VS. 10 * LOG10 (POWER)

analyses, it was felt that the averaged-subtest spectra were an accurate representation of the amplitude-frequency characteristics of the background noise.

The computer-averaged spectra of the noise for each subtest were used to determine the values for the acoustical characteristics of the syllables as they occur in noise. The measurements of speech-in-noise included vowel peak frequency; origin frequency, magnitude, and direction of the second formant transition; consonant spectral peaks; and the additional measurement known as crossover frequency.

The procedures outlined for these measurements were the same for vowel peak, origin frequency, and consonant peaks. If the peak-to-noise ratio (using the averaged noise spectra for each subtest as the reference) was 10 dB or worse, the peak with the highest peak-to-noise ratio was designated as the measurement for the noise condition. A change in origin frequency of the second formant transition resulted in a subsequent change in the magnitude of the transition. Direction of the transition did not change in noise, except for the cases in which the

entire transition is masked.

It was extremely difficult to determine the effect of noise on overall consonant bandwidth using the criterion established for the other acoustical measurements. An alternative procedure for the noise condition was established which resulted in an estimate of crossover frequency. The crossover frequency was defined as the frequency above which the spectrum of the consonant lies above the noise spectrum. The result is, in effect, a measurement of the region of the consonant spectrum that is available to the listener. The lower the crossover frequency, the wider the effective bandwidth in noise. A high crossover frequency implies that a large portion of the spectrum of the consonant was masked by the noise.

CHAPTER IV

RESULTS

LISTENING EXPERIMENT

Analysis of Intelligibility

The results of the listening experiment were analyzed as follows. The proportion of syllables correctly heard in each subtest for each condition was calculated for each of the six listeners. True percents were calculated from the measured data according to the following formula:

$$P_T = \frac{NS_M - 1}{N - 1} ,$$

where N is the number of possible responses and S_M is the measured score. This results in a score which includes a correction for random guessing. In addition, each nonsense syllable subtest contains one repeat item. Therefore, that syllable and its repeat had half-weight applied to it. For example, a set with eight possible response alternatives actually had nine target syllables, including the repeat item. Full-weight was given to seven syllables

(0.125 x 7), and half-weight was given to the two identical syllables (0.0625 x 2) for a total of 1.0. For further analyses, the proportions were transformed to arc sine units in order to stabilize the binomial sampling variability that is inherent in estimating a proportion from a finite set of items (Brownlee, 1965).

Despite differences in intelligibility scores between subjects, the decision was made to combine the data over this factor. An analysis of the errors made by each subject revealed similar patterns of confusions. Further, there were few interactions between subjects and other experimental variables. Thus, it was possible to average over subjects without serious loss of information.

After applying the arc sine transformation, the intelligibility data were analyzed using a three-way analysis of variance. The factors were: Subtest of Nonsense Syllables (1 through 11), Noise Condition (Quiet or S/N = 5 dB), and Speech Level (20, 28, 36, 44, and 52 dB SPL).

The results of the analysis of varaince are shown in Table 1. The table shows that each of the three factors

TABLE 1. Analysis of Variance of Intelligibility Scores (Nonsense Syllable Subtest by Speech Level by Noise Condition)

Source of Variation	Sum of squares	DF	Mean Square	F	Significance
Subtest (T)	5.62	6	0.94	32.14	0.001
Speech Level (L)	17.85	4	4.46	153.10	0.001
Noise Condition (N)	1.15	1	1.15	39.36	0.001
T x L	0.52	24	0.02	.74	0.769
T x N	0.29	6	0.05	1.67	0.173
L x N	1.04	4	0.26	8.95	0.001
T x L x N	0.70	24	0.03		
Total	27.16	69			

is significant. There are significant differences in intelligibility scores between subtests and, as expected, between noise conditions and between speech levels. A significant noise condition-speech level interaction is also present. There are no significant interactions between subtest and noise condition, or between subtest and speech level.

The intelligibility scores for the 11 subtest in each noise condition at each speech level are summarized in Table 2. The average values were obtained from the analysis of variance after the scores were transformed back to proportions from the arc sine units. The data show that improvement in intelligibility with increasing speech level changes as a function of syllable subtest type, although this interaction is not statistically significant. Although this treatment will not allow for a test of this effect, there appears to be evidence of a three-way interaction between subtest, speech level and noise condition. In the quiet conditions, subtests containing voiced consonants in combination with the vowel /ɑ/ appear to have more steeply rising functions, as is the case for the subtest containing

TABLE 2. Average Scores in Percent Correct for Each of the 11 Nonsense Syllable Subtests, for Five Speech Levels (dB SPL) and Two Noise Conditions (Quiet and S/N = 5 dB)

Speech Level	S/N	Nonsense Syllable Subtest 1	2	3	4	5	6	7	8	9	10	11	$\bar{x}$
20	Q	20.8	9.7	23.6	52.1	35.4	52.4	14.3	--	--	--	--	29.76
28	Q	66.7	41.7	38.9	74.0	71.9	89.3	56.0	61.4	72.9	40.6	40.6	59.45
36	Q	75.0	62.5	65.3	91.7	91.7	100.0	84.5	78.1	76.0	63.5	51.0	76.30
44	Q	87.5	81.9	86.1	96.9	96.9	100.0	91.7	93.8	92.7	78.1	87.5	90.28
52	Q	94.4	94.4	88.9	100.0	100.0	100.0	98.8	--	--	--	--	96.64
20	+5	18.0	27.8	29.2	32.3	34.4	57.1	32.1	--	--	--	--	32.99
28	+5	51.4	23.6	59.7	68.8	71.9	72.6	54.8	54.2	34.4	20.8	24.0	48.75
36	+5	79.2	56.9	59.7	89.6	71.2	91.7	75.0	64.6	54.2	37.5	39.6	65.38
44	+5	76.4	55.6	77.8	86.5	74.0	96.4	86.9	76.0	74.0	55.2	53.2	73.82
52	+5	80.6	62.5	72.2	83.3	82.3	96.4	83.3	--	--	--	--	80.09

that vowel with voiceless consonants in the initial position. For the noise conditions, this effect is present only for the voiced, syllable-initial consonants combined with the vowel /ɑ/ (subtest 6). This set, containing the glides and liquids /w, j, r, l/, is associated with consistent high scores for all conditions. The remaining voiced subtests with /ɑ/ and the voiceless, syllable-final subtest with /i/ appear to be more adversely affected by higher noise levels, despite the constant (5 dB) speech-to-noise ratio. This effect is characterized by functions which reach their peak at the 44 dB SPL speech level. Thus, in the noise conditions, an increase in speech level did not always lead to an increase in test score on all subtests.

A careful inspection was made of the errors in these three sets for each level, in quiet and in noise. The analysis reveals that the decrement in intelligibility score at the highest speech level could be accounted for on the basis of an increasing error rate with increasing speech level for three confusions only, namely /if/iθ/, /ɑv/ɑb/, and /vɑ/bɑ/*. The overall pattern of confusions for all

* The first syllable of the pair is the target; the second syllable represents the incorrect response.

other levels and for all other subtests, however, remains the same. In the course of these analyses it was found that, with the exception of the three confusions cited above, the error pattern in each confusion matrix was essentially the same as a function of level. That is, although the total number of errors decreased with increasing speech level, the relative frequency of occurrence of each confusion remained the same. Based on this finding, the decision was made to sum over speech level for the subsequent confusion matrix analysis.

The interaction between noise condition and speech level, as shown in Figure 11, appears to be largely confined to the lowest speech level. This figure also shows that, on average, rate of improvement in intelligibility with level appears to be slightly lower in noise than in the quiet condition.

Finally, there is an apparent lack of interaction between subtest and noise condition. These data are plotted in Figure 12. The syllable subtests were rank ordered on the horizontal axis according to the magnitude of the average score (over both quiet and noise conditions). Again,

Fig. 11. Intelligibility score, in percent correct, for nonsense syllables plotted as a function of speech level, in dB SPL, for both quiet and noise conditions.

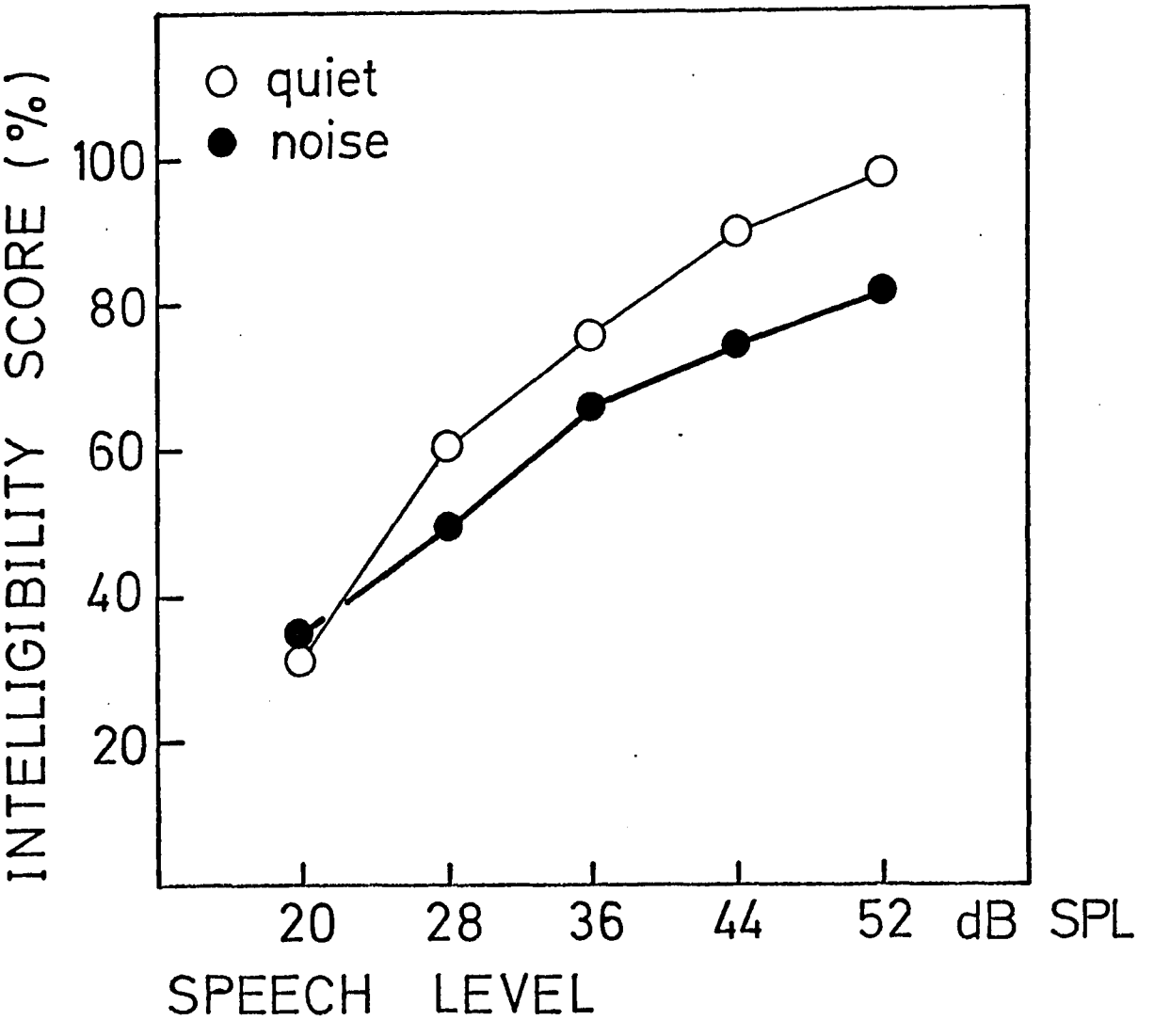
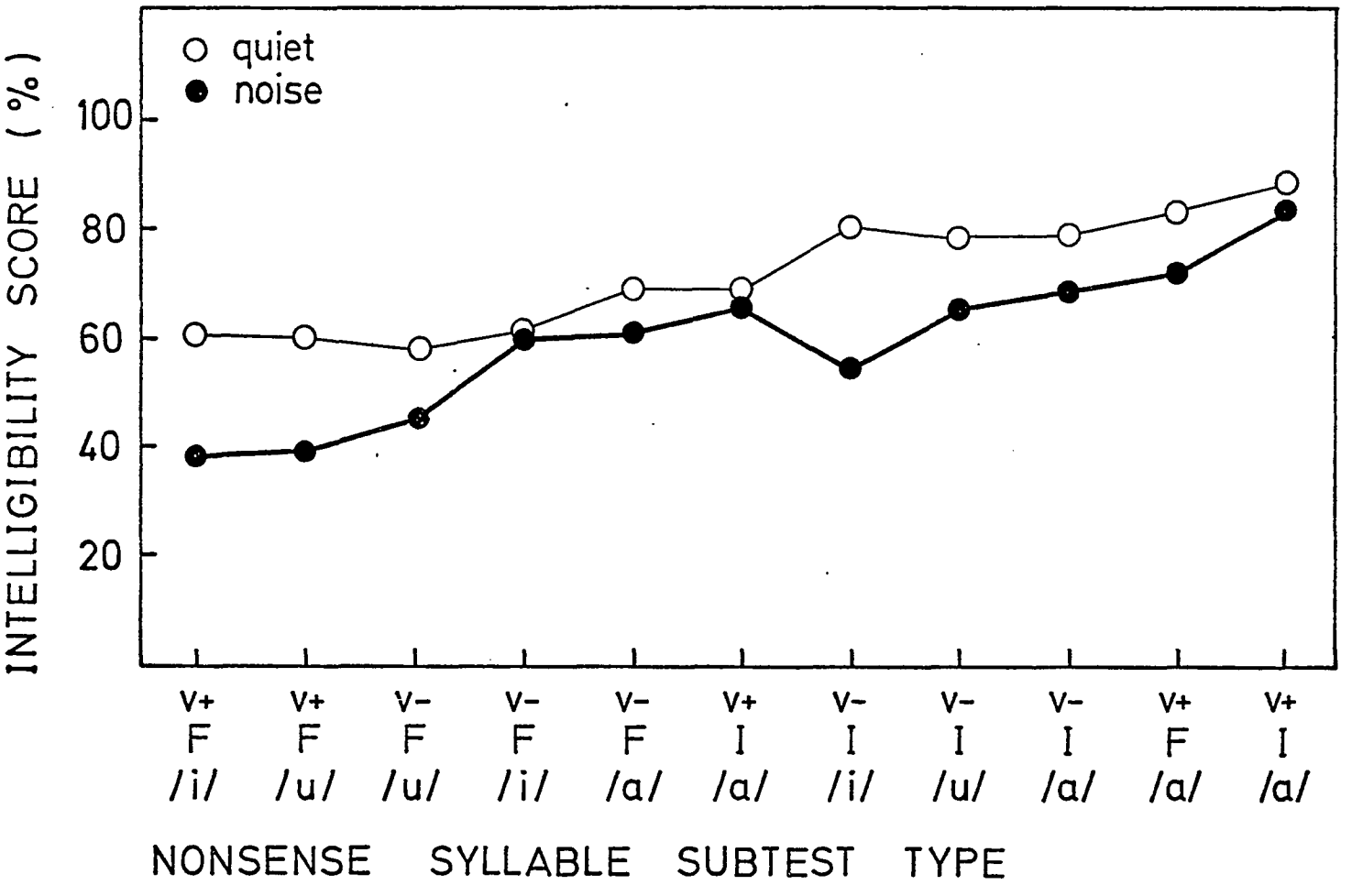

quiet
noise
INTELLIGIBILITY SCORE (%)
100
80
60
40
20
20
28
36
44
52
dB SPL
SPEECH LEVEL

Fig. 12. Intelligibility score, in percent correct, for each of the 11 subtests of nonsense syllables for both quiet and noise conditions. Consonants in each subtest are voiced (V+) or voiceless (V-), initial (I) or final (F), with vowels /ɑ/, /u/ or /i/.


quiet
noise
INTELLIGIBILITY SCORE (%)
100
80
60
40
20
V+ F /i/
V+ F /u/
V- F /u/
V- F /i/
V- F /a/
V+ I /a/
V- I /i/
V- I /u/
V- I /a/
V+ F /a/
V+ I /a/
NONSENSE SYLLABLE SUBTEST TYPE

on average, the highest scores were obtained on the subtests containing voiced consonants with /ɑ/, followed by voiceless, syllable-initial consonants with /i/, /ɑ/, and /u/. This figure also reveals the differential effects of noise on intelligibility. On average, intelligibility of voiceless initial and voiced final consonants with /i/ and all final consonants with /u/ appears to be most adversely affected by noise.

The importance of syllable subtest type revealed in the above analyses was explored further by separate analyses on the three factors differentiating the subtests, i.e., consonant voicing, consonant position, and vowel context. Transformed intelligibility scores were averaged according to subtest type, and three, three-way analyses of variance were performed with speech level and noise condition serving as the other two factors. The results show that there are significant differences in intelligibility scores between voiced and voiceless subtests, initial and final subtests, and between subtests containing the three different vowels. Again, speech level and noise condition main effects are significant in all three analyses. There are no significant

interactions between any of the syllable subtest types and noise condition. In addition, there are no significant position-speech level or vowel-speech level interactions, although a voicing-speech level interaction is found to be significant. These intelligibility scores, converted back to proportions, are plotted in Figures 13, 14, and 15, respectively. The data reveal that, on average, intelligibility is significantly better for syllable-initial consonants than for syllable-final consonants, in both quiet and noise conditions. In addition, the intelligibility of consonants combined with the vowel /ɑ/ is significantly greater than for consonants paired with either /i/ or /u/. Finally, voiced consonants appear to be significantly more intelligible than voiceless consonants only at the very lowest and very highest speech levels. The former finding could be due to the reduced overall energy in the voiceless consonants themselves. The reason for the decrement at the highest level could be a result of the greater effect of noise on the voiceless consonants.

Fig. 13. Intelligibility score, in percent correct, for nonsense syllables plotted as a function of speech level, in dB SPL, with consonant position as paramter.

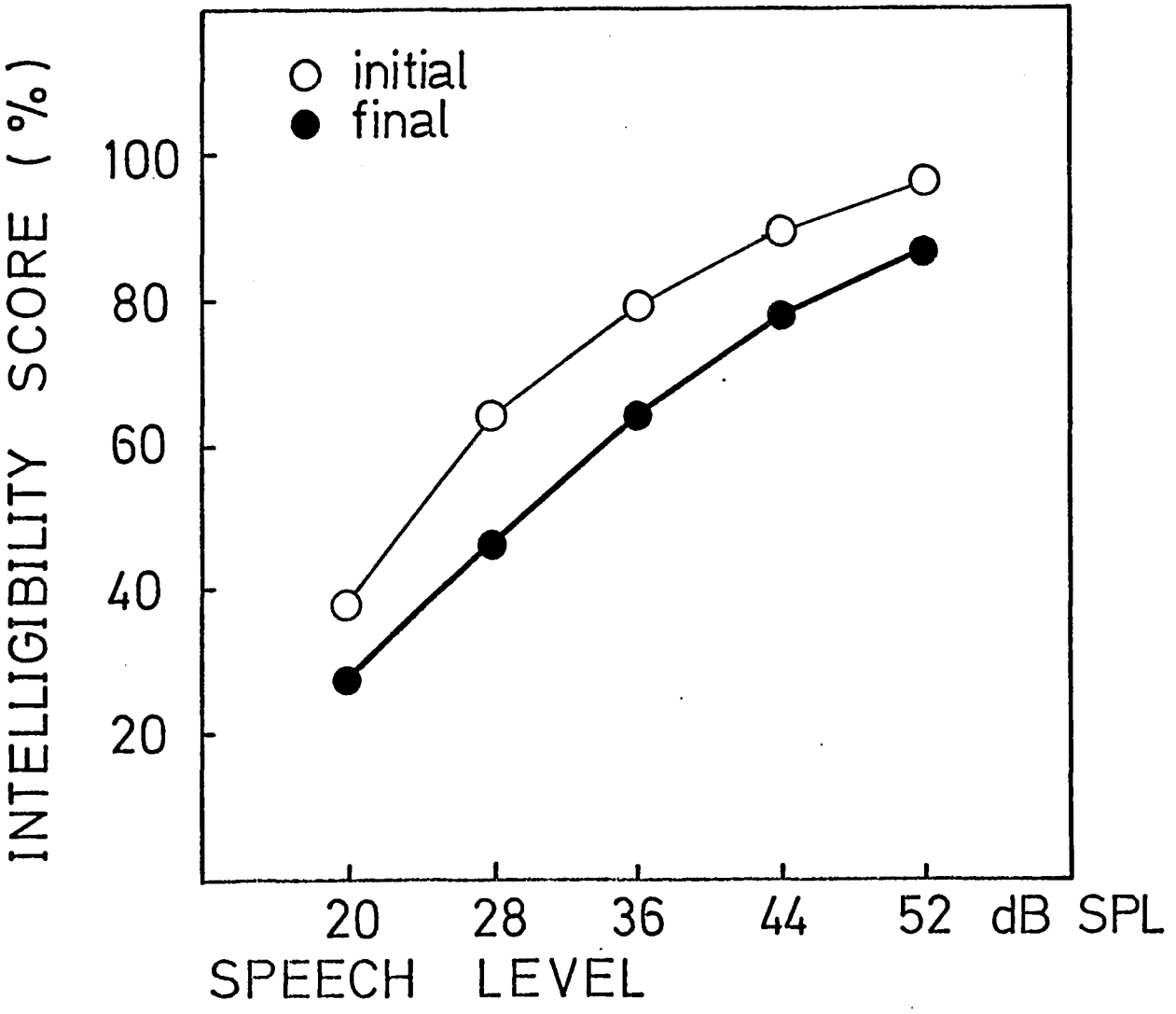

initial
final
INTELLIGIBILITY SCORE (%)
100
80
60
40
20
20
28
36
44
52
dB SPL
SPEECH LEVEL

Fig. 14. Intelligibility score, in percent correct, for nonsense syllables plotted as a function of speech level, in dB SPL, with vowel context as parameter.

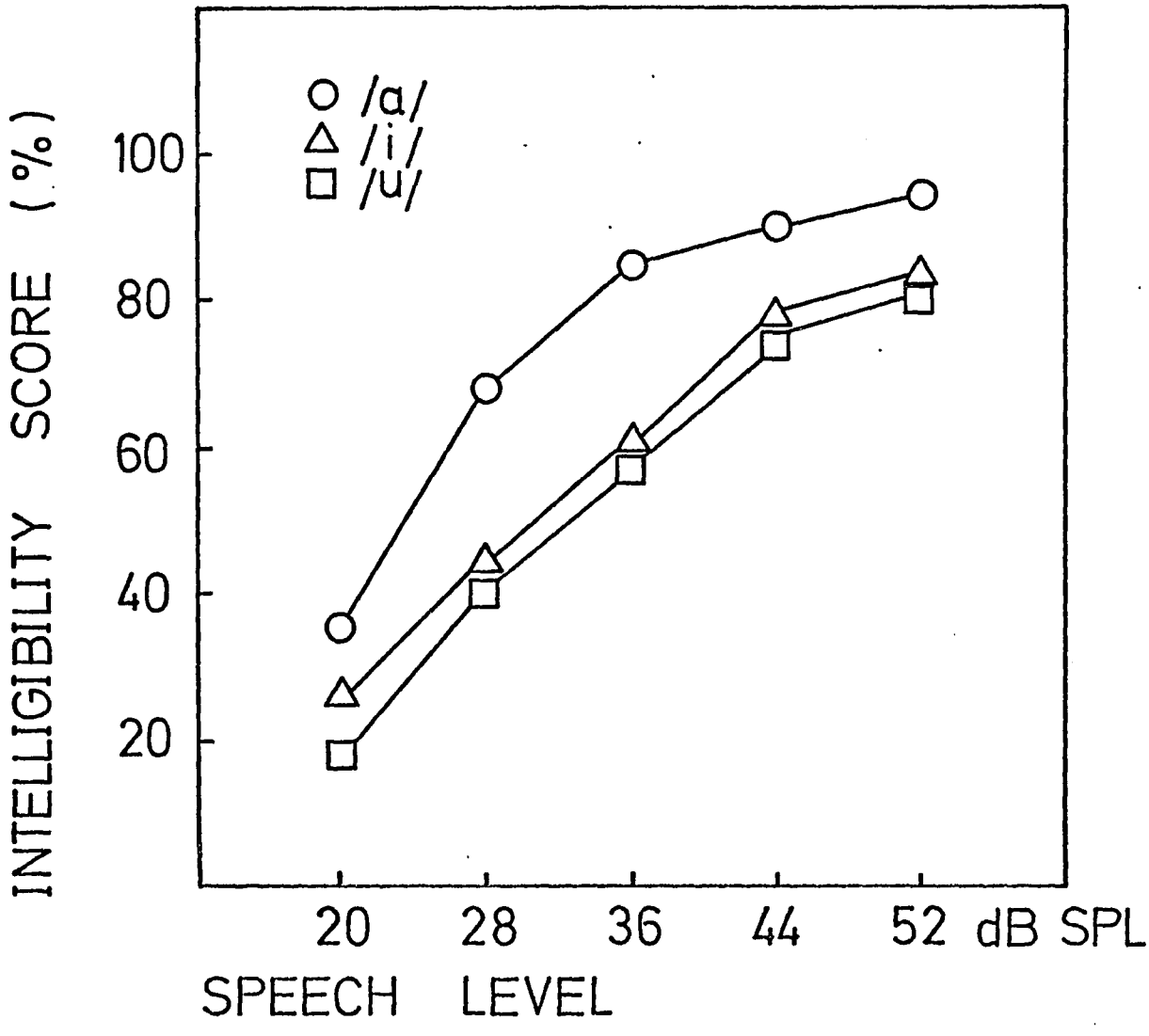
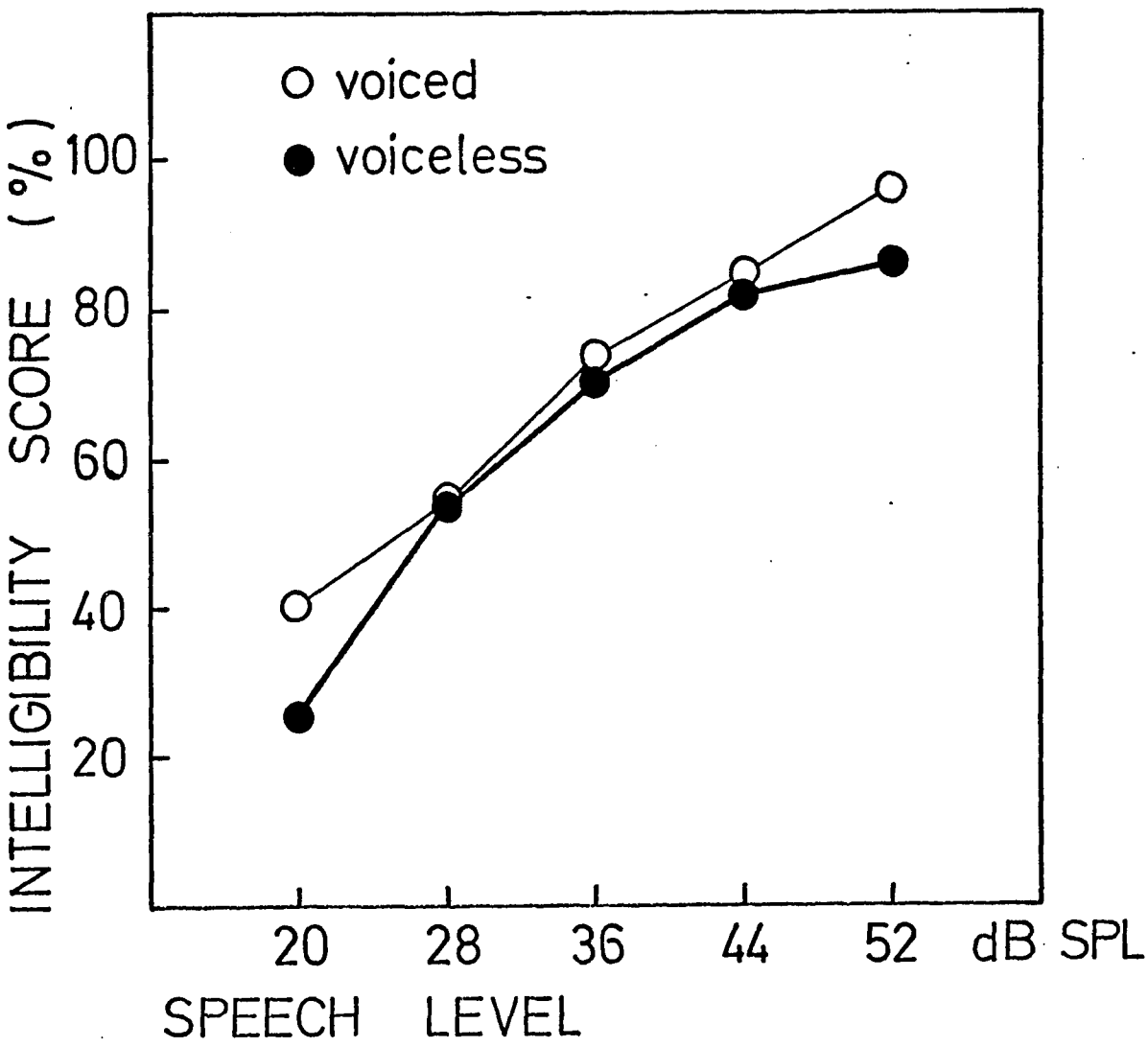

INTELLIGIBILITY SCORE (%)
100
80
60
40
20
○ /a/
△ /i/
□ /u/
20
28
36
44
52 dB SPL
SPEECH LEVEL

Fig. 15. Intelligibility score, in percent correct, for nonsense syllables plotted as a function of speech level, in dB SPL, with consonant voicing as parameter.


voiced
voiceless
INTELLIGIBILITY SCORE (%)
100
80
60
40
20
20
28
36
44
52
dB SPL
SPEECH LEVEL

Analysis of Correct Identifications

Confusion matrices for each of the nonsense syllable subtests were generated from the results of the listening experiment. For reasons stated previously, matrices to be analyzed were those summed over subject and speech level. Thus, there are two matrices for each syllable subtest, one each for the quiet and noise conditions. These matrices are shown in Tables 3 through 13.

For each matrix, the target syllables are those appearing in the rows, and the response syllables are those in the columns. The syllables comprising each matrix are arranged by manner of articulation and, within each manner class, by place of articulation. Thus, for each matrix, the plosives and fricatives appear first, followed by the nasals, glides, liquids and/or affricates. Within each manner class, the labial place of articulation appears first, followed by the labio-dentals, alveolars, and velars.

Inspection of the diagonal for each matrix reveals the percentage of correct identifications for each syllable in both quiet and noise. The data for the quiet condition are summarized in Table 14 according to consonant manner of

TABLE 3. Confusion Matrices for Nonsense Syllable Subtest 1, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ɑp	ɑt	ɑk	ɑf	ɑθ	ɑs	ɑʃ
ɑp	90.0	0.0	0.0	0.0	3.3	3.3	3.3
ɑt	6.7	68.3	11.7	3.3	3.3	0.0	6.7
ɑk	6.7	6.7	80.0	3.3	0.0	0.0	3.3
ɑf	20.0	0.0	8.3	56.7	13.3	1.7	0.0
ɑθ	3.3	8.3	8.3	16.7	50.0	6.7	6.7
ɑs	3.3	10.0	6.7	0.0	0.0	80.0	0.0
ɑʃ	3.3	3.3	0.0	0.0	5.0	3.3	85.0

NOISE

	ɑp	ɑt	ɑk	ɑf	ɑθ	ɑs	ɑʃ
ɑp	63.3	10.0	13.3	6.7	6.7	0.0	0.0
ɑt	11.7	73.3	6.7	1.7	0.0	3.3	3.3
ɑk	0.0	16.7	70.0	3.3	3.3	6.7	0.0
ɑf	36.7	6.7	8.3	41.7	3.3	0.0	3.3
ɑθ	23.3	3.3	8.3	6.7	53.3	3.3	1.7
ɑs	0.0	8.3	8.3	0.0	5.0	75.0	3.3
ɑʃ	0.0	3.3	3.3	3.3	0.0	0.0	90.0

TABLE 4. Confusion Matrices for Nonsense Syllable Subtest 2, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	up	ut	uk	uf	uθ	us	uʃ
up	46.7	6.7	13.3	23.3	6.7	0.0	3.3
ut	16.7	60.0	16.7	6.7	0.0	0.0	0.0
uk	18.3	1.7	68.3	5.0	0.0	6.7	0.0
uf	6.7	16.7	5.0	51.7	18.3	0.0	1.7
uθ	11.7	16.7	10.0	10.0	48.3	0.0	3.3
us	0.0	6.7	0.0	3.3	3.3	83.3	3.3
uʃ	0.0	3.3	0.0	3.3	3.3	3.3	86.7

NOISE

	up	ut	uk	uf	uθ	us	uʃ
up	23.3	13.3	33.3	20.0	10.0	0.0	0.0
ut	0.0	56.7	20.0	3.3	10.0	0.0	10.0
uk	18.3	21.7	31.7	21.7	5.0	1.7	0.0
uf	18.3	6.7	11.7	41.7	16.7	3.3	1.7
uθ	20.0	20.0	20.0	10.0	30.0	0.0	0.0
us	3.3	0.0	0.0	0.0	0.0	95.0	1.7
uʃ	0.0	3.3	3.3	0.0	3.3	0.0	90.0

TABLE 5. Confusion Matrices for Nonsense Syllable Subtest 3, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ip	it	ik	if	iθ	is	iʃ
ip	56.7	25.0	3.3	10.0	5.0	0.0	0.0
it	6.7	75.0	3.3	5.0	5.0	3.3	1.7
ik	0.0	6.7	76.7	3.3	6.7	0.0	6.7
if	6.7	10.0	13.3	33.3	36.7	0.0	0.0
iθ	6.7	10.0	13.3	33.3	36.7	0.0	0.0
is	0.0	0.0	8.3	0.0	0.0	91.7	0.0
iʃ	0.0	0.0	3.3	0.0	3.3	0.0	93.3

NOISE

	ip	it	ik	if	iθ	is	iʃ
ip	48.3	18.3	13.3	6.7	10.0	0.0	3.3
it	0.0	88.3	5.0	0.0	5.0	1.7	0.0
ik	3.3	3.3	91.7	0.0	1.7	0.0	0.0
if	3.3	10.0	23.3	20.0	43.3	0.0	0.0
iθ	6.7	16.7	23.3	26.7	23.3	0.0	3.3
is	0.0	5.0	0.0	1.7	0.0	93.3	0.0
iʃ	0.0	0.0	0.0	3.3	3.3	0.0	93.3

TABLE 6. Confusion Matrices for Nonsense Syllable Subtest 4, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ɑb	ɑd	ɑg	ɑv	ɑð	ɑz	ɑm	ɑn	ɑŋ
ɑb	90.0	0.0	6.7	3.3	0.0	0.0	0.0	0.0	0.0
ɑd	3.3	90.0	0.0	0.0	3.3	3.3	0.0	0.0	0.0
ɑg	0.0	6.7	86.7	3.3	3.3	0.0	0.0	0.0	0.0
ɑv	21.7	0.0	0.0	73.3	5.0	0.0	0.0	0.0	0.0
ɑð	3.3	6.7	3.3	5.0	68.3	10.0	3.3	0.0	0.0
ɑz	0.0	8.3	3.3	0.0	0.0	80.0	1.7	3.3	3.3
ɑm	0.0	0.0	5.0	3.3	0.0	0.0	86.7	0.0	5.0
ɑn	0.0	3.3	0.0	0.0	0.0	3.3	0.0	93.3	0.0
ɑŋ	0.0	0.0	0.0	0.0	1.7	0.0	0.0	3.3	95.0

NOISE

	ɑb	ɑd	ɑg	ɑv	ɑð	ɑz	ɑm	ɑn	ɑŋ
ɑb	80.0	3.3	6.7	6.7	0.0	0.0	0.0	3.3	0.0
ɑd	0.0	86.7	6.7	0.0	3.3	3.3	0.0	0.0	0.0
ɑg	6.7	6.7	73.3	0.0	10.0	0.0	0.0	3.3	0.0
ɑv	31.7	0.0	6.7	46.7	10.0	0.0	1.7	1.7	1.7
ɑð	0.0	6.7	5.0	3.3	63.3	10.0	0.0	11.7	0.0
ɑz	0.0	1.7	3.3	6.7	3.3	80.0	0.0	0.0	5.0
ɑm	6.7	0.0	3.3	11.7	0.0	0.0	73.3	0.0	5.0
ɑn	0.0	6.7	6.7	0.0	6.7	0.0	0.0	80.0	0.0
ɑŋ	3.3	0.0	0.0	3.3	0.0	0.0	0.0	0.0	93.3

TABLE 7. Confusion Matrices for Nonsense Syllable Subtest 5, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	pa	ta	ka	fa	θa	sa	ʃa	ha	tʃa
pa	61.7	6.7	0.0	15.0	3.3	0.0	0.0	13.3	0.0
ta	5.0	75.0	16.7	0.0	0.0	3.3	0.0	0.0	0.0
ka	5.0	8.3	83.3	3.3	0.0	0.0	0.0	0.0	0.0
fa	0.0	0.0	0.0	81.7	13.3	0.0	0.0	5.0	0.0
θa	3.3	3.3	3.3	0.0	86.7	0.0	3.3	0.0	0.0
sa	0.0	0.0	3.3	6.7	0.0	90.0	0.0	0.0	0.0
ʃa	0.0	0.0	0.0	0.0	3.3	0.0	90.0	0.0	6.7
ha	10.0	16.7	3.3	3.3	0.0	0.0	0.0	66.7	0.0
tʃa	0.0	0.0	0.0	0.0	0.0	0.0	1.7	0.0	98.3

NOISE

	pa	ta	ka	fa	θa	sa	ʃa	ha	tʃa
pa	53.3	8.3	15.0	20.0	0.0	0.0	0.0	3.3	0.0
ta	3.3	75.0	6.7	0.0	6.7	5.0	0.0	3.3	0.0
ka	13.3	11.7	53.3	0.0	0.0	0.0	0.0	21.7	0.0
fa	1.7	3.3	6.7	70.0	11.7	0.0	0.0	6.7	0.0
θa	0.0	3.3	3.3	6.7	80.0	0.0	0.0	6.7	0.0
sa	0.0	0.0	3.3	0.0	3.3	93.3	0.0	0.0	0.0
ʃa	0.0	0.0	0.0	0.0	0.0	0.0	80.0	0.0	20.0
ha	21.7	6.7	16.7	0.0	0.0	0.0	3.3	51.7	0.0
tʃa	0.0	3.3	0.0	0.0	0.0	0.0	6.7	0.0	90.0

TABLE 8. Confusion Matrices for Nonsense Syllable Subtest 6, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ba	da	ga	ja	ra	wa	la	dʒa
ba	81.7	0.0	0.0	0.0	3.3	5.0	6.7	3.3
da	3.3	90.0	0.0	3.3	0.0	0.0	0.0	3.3
ga	0.0	6.7	90.0	0.0	0.0	0.0	0.0	3.3
ja	0.0	3.3	0.0	93.3	0.0	0.0	0.0	3.3
ra	0.0	0.0	0.0	0.0	96.7	3.3	0.0	0.0
wa	0.0	0.0	0.0	0.0	0.0	96.7	3.3	0.0
la	1.7	0.0	0.0	0.0	10.0	6.7	81.7	0.0
dʒa	0.0	3.3	5.0	3.3	0.0	0.0	0.0	88.3

NOISE

	ba	da	ga	ja	ra	wa	la	dʒa
ba	70.0	0.0	1.7	1.7	5.0	1.7	20.0	0.0
da	0.0	90.0	10.0	0.0	0.0	0.0	0.0	0.0
ga	0.0	8.3	85.0	3.3	3.3	0.0	0.0	0.0
ja	0.0	3.3	0.0	96.7	0.0	0.0	0.0	0.0
ra	0.0	0.0	0.0	0.0	98.3	0.0	1.7	0.0
wa	1.7	0.0	0.0	0.0	1.7	93.3	3.3	0.0
la	1.7	0.0	3.3	0.0	8.3	23.3	63.3	0.0
dʒa	0.0	0.0	6.7	10.0	0.0	0.0	0.0	83.3

TABLE 9. Confusion Matrices for Nonsense Syllable Subtest 7, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ba	da	ga	va	ða	za	ma	na
ba	56.7	0.0	3.3	16.7	15.0	8.3	0.0	0.0
da	0.0	76.7	5.0	1.7	3.3	10.0	0.0	3.3
ga	0.0	16.7	80.0	0.0	0.0	3.3	0.0	0.0
va	16.7	0.0	0.0	70.0	11.7	0.0	0.0	1.7
ða	3.3	20.0	6.7	0.0	61.7	3.3	0.0	5.0
za	0.0	10.0	3.3	0.0	3.3	76.7	3.3	3.3
ma	3.3	0.0	0.0	6.7	3.3	0.0	80.0	6.7
na	0.0	11.7	6.7	0.0	0.0	0.0	0.0	81.7

NOISE

	ba	da	ga	va	ða	za	ma	na
ba	25.0	0.0	5.0	23.3	40.0	5.0	0.0	1.7
da	5.0	75.0	13.3	3.3	0.0	3.3	0.0	0.0
ga	0.0	10.0	86.7	0.0	0.0	0.0	3.3	0.0
va	26.7	1.7	0.0	50.0	20.0	0.0	1.7	0.0
ða	0.0	13.3	6.7	1.7	70.0	6.7	1.7	0.0
za	0.0	6.7	0.0	3.3	0.0	90.0	0.0	0.0
ma	13.3	0.0	3.3	3.3	0.0	0.0	76.7	3.3
na	0.0	6.7	1.7	0.0	0.0	0.0	0.0	91.7

TABLE 10. Confusion Matrices for Nonsense Syllable Subtest 8, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	pu	tu	ku	fu	θu	su	ʃu	hu	tʃu
pu	33.3	5.6	25.0	5.6	5.6	0.0	0.0	19.4	5.6
tu	0.0	94.4	0.0	5.6	0.0	0.0	0.0	0.0	0.0
ku	0.0	0.0	94.4	0.0	0.0	0.0	0.0	0.0	0.0
fu	0.0	5.6	0.0	55.6	0.0	25.0	0.0	13.9	0.0
θu	0.0	5.6	0.0	0.0	94.4	0.0	0.0	0.0	0.0
su	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
ʃu	0.0	0.0	0.0	0.0	0.0	5.6	94.4	0.0	0.0
hu	5.6	0.0	16.7	11.1	0.0	0.0	0.0	66.7	0.0
tʃu	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0	88.9

NOISE

	pu	tu	ku	fu	θu	su	ʃu	hu	tʃu
pu	44.4	8.3	16.7	8.3	0.0	5.6	0.0	16.7	0.0
tu	0.0	72.2	5.6	5.6	0.0	0.0	0.0	16.7	0.0
ku	5.6	0.0	77.8	0.0	0.0	0.0	0.0	16.7	0.0
fu	16.7	5.5	8.3	33.3	0.0	0.0	5.6	30.6	0.0
θu	0.0	5.6	5.6	11.1	66.7	0.0	0.0	5.6	5.6
su	0.0	0.0	0.0	0.0	0.0	94.4	0.0	5.6	0.0
ʃu	5.6	0.0	0.0	0.0	0.0	0.0	94.4	0.0	0.0
hu	16.7	5.6	0.0	22.2	11.1	0.0	0.0	44.4	0.0
tʃu	0.0	0.0	5.6	0.0	0.0	2.8	2.8	0.0	88.9

TABLE 11. Confusion Matrices for Nonsense Syllable Subtest 9, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	pi	ti	ki	fi	θi	si	ʃi	hi	tʃi
pi	72.2	5.6	5.6	13.9	2.8	0.0	0.0	0.0	0.0
ti	0.0	88.9	5.6	5.6	0.0	0.0	0.0	0.0	0.0
ki	0.0	0.0	94.4	0.0	0.0	0.0	0.0	5.6	0.0
fi	0.0	5.6	0.0	83.3	5.6	5.6	0.0	0.0	0.0
θi	5.6	0.0	2.7	41.7	44.4	0.0	5.6	0.0	0.0
si	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
ʃi	0.0	5.6	0.0	0.0	0.0	0.0	94.4	0.0	0.0
hi	5.6	8.3	5.6	2.8	0.0	0.0	5.6	72.2	0.0
tʃi	0.0	0.0	0.0	0.0	0.0	0.0	5.6	0.0	94.4

NOISE

	pi	ti	ki	fi	θi	si	ʃi	hi	tʃi
pi	36.1	5.6	11.1	22.2	5.6	0.0	0.0	13.9	5.6
ti	0.0	66.7	5.6	11.1	0.0	0.0	5.6	5.6	5.6
ki	5.6	5.6	77.8	5.6	0.0	0.0	0.0	5.6	0.0
fi	33.3	11.1	0.0	33.3	0.0	5.6	0.0	11.1	5.6
θi	2.8	5.6	2.8	52.8	27.8	0.0	2.8	5.6	0.0
si	0.0	2.8	0.0	2.8	5.6	88.9	0.0	0.0	0.0
ʃi	0.0	0.0	0.0	0.0	0.0	5.6	83.3	0.0	11.1
hi	33.3	8.3	5.6	16.7	8.3	0.0	0.0	27.8	0.0
tʃi	0.0	0.0	0.0	0.0	0.0	0.0	8.3	0.0	91.7

TABLE 12. Confusion Matrices for Nonsense Syllable Subtest 10, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ib	id	ig	iv	ið	iz	im	in	iŋ
ib	72.2	16.7	5.6	5.6	0.0	0.0	0.0	0.0	0.0
id	5.6	52.8	8.3	16.7	13.9	2.8	0.0	0.0	0.0
ig	0.0	11.1	83.3	0.0	5.6	0.0	0.0	0.0	0.0
iv	5.6	19.4	5.6	8.3	50.0	0.0	8.3	0.0	2.8
ið	5.6	11.1	0.0	30.6	52.8	0.0	0.0	0.0	0.0
iz	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
im	5.6	5.6	0.0	0.0	0.0	0.0	83.3	5.6	0.0
in	0.0	5.6	5.6	0.0	0.0	0.0	0.0	72.2	16.7
iŋ	0.0	0.0	0.0	11.1	0.0	0.0	11.1	16.7	61.1

NOISE

	ib	id	ig	iv	ið	iz	im	in	iŋ
ib	33.3	25.0	0.0	19.4	16.7	5.6	0.0	0.0	0.0
id	8.3	38.9	13.9	8.3	5.6	5.6	11.1	2.8	5.6
ig	0.0	8.3	63.9	5.6	16.7	0.0	0.0	5.6	0.0
iv	8.3	19.4	11.1	19.4	13.8	5.6	5.6	16.7	0.0
ið	11.1	13.9	0.0	16.7	50.0	2.8	5.6	0.0	0.0
iz	0.0	5.6	0.0	8.3	0.0	77.8	0.0	5.6	2.8
im	33.3	5.6	5.6	5.6	5.6	0.0	27.8	11.1	5.6
in	5.6	16.7	5.6	11.1	0.0	0.0	33.3	22.2	5.6
iŋ	5.6	5.6	5.6	0.0	0.0	0.0	5.6	16.7	61.1

TABLE 13. Confusion Matrices for Nonsense Syllable Subtest 11, Averaged Over All Speech Levels; Upper Matrix for Quiet Condition, Lower Matrix for Noise Condition

QUIET

	ub	ud	ug	uv	uð	uz	um	un	uŋ
ub	36.1	22.2	11.1	19.4	11.1	0.0	0.0	0.0	0.0
ud	0.0	66.7	0.0	11.1	11.1	0.0	5.6	0.0	5.6
ug	5.6	5.6	77.8	5.6	5.6	0.0	0.0	0.0	0.0
uv	27.8	5.6	8.3	33.3	19.4	0.0	5.6	0.0	0.0
uð	5.6	5.6	2.8	2.8	72.2	5.6	0.0	5.6	0.0
uz	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
um	0.0	5.6	0.0	0.0	0.0	0.0	72.2	16.7	5.6
un	0.0	5.6	0.0	5.6	5.6	0.0	8.3	63.9	11.1
uŋ	0.0	0.0	0.0	0.0	0.0	0.0	33.3	11.1	55.6

NOISE

	ub	ud	ug	uv	uð	uz	um	un	uŋ
ub	25.0	30.6	2.8	27.8	8.3	0.0	5.6	0.0	0.0
ud	5.6	50.0	5.6	0.0	5.6	0.0	22.2	11.1	0.0
ug	22.2	0.0	22.2	33.3	5.6	5.6	5.6	0.0	5.6
uv	16.7	11.1	11.1	50.0	5.6	0.0	5.6	0.0	0.0
uð	0.0	5.6	13.9	25.0	30.6	19.4	0.0	5.6	0.0
uz	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
um	13.9	8.3	16.7	0.0	0.0	0.0	52.8	5.6	2.8
un	5.6	0.0	0.0	5.6	8.3	11.1	22.2	30.6	16.7
uŋ	0.0	0.0	0.0	0.0	0.0	0.0	33.3	16.7	50.0

TABLE 14. Percent Correct Identification of Nonsense Syllables in the Quiet Condition, According to Consonant Manner of Articulation, as a Function of Consonant Voicing and Position

		Plosives	Fricatives	Nasals	Glides and	Affricates
Voiced	$\bar{x}$	75.4	66.4	76.8	92.1	88.3
	SD	6.0	25.8	13.0	7.1	--
	N	15	12	11	4	1
Voiceless	$\bar{x}$	73.3	74.7	--	--	93.9
	SD	17.1	20.5	--	--	4.7
	N	18	27			3
Initial	$\bar{x}$	78.2	79.7	80.9	92.1	92.5
	SD	17.0	16.2	1.2	7.1	4.8
	N	15	18	2	4	4
Final	$\bar{x}$	71.0	66.0	75.9	--	--
	SD	15.6	25.2	14.3	--	--
	N	18	21	9		
Total	$\bar{x}$	74.2	72.3	76.8	92.1	92.5
	SD	16.4	22.3	13.0	7.1	4.8
	N	33	39	11	4	4

articulation. The table shows that affricates are most easily identified, followed closely by the glides and liquids, and finally the nasals and plosives. Fricative consonants are the most difficult to identify, with especially small percentages for voiced fricatves and fricatives in the syllable-final position. Although overall percentages for plosives, fricatives, and nasals were lower, the identification of these phonemes appears to improve when in the syllable-initial position.

Data for percentages of correct identification in noise are summarized in Table 15. As in the quiet condition, the affricates, along with the glides and liquids, are most easily identified, with scores for the noise condition only slightly lower than in quiet. Identification of nasal consonants, most notably in the syllable-final position, appears to be adversely affected by noise. Plosive consonants are most difficult to identify in noise, especially those in the syllable-final position. In the noise condition, scores for voiceless fricatives are a great deal lower than scores in the quiet condition. Voiced plosives, on the other hand, are slightly more affected by noise. As in the quiet condi-

TABLE 15. Percent Correct Identification of Nonsense Syllables in the Noise Condition, According to Consonant Manner of Articulation, as a Function of Consonant Voicing and Position

		Plosives	Fricatives	Nasals	Glides and Liquids	Affricates
Voiced	$\bar{x}$	60.3	62.7	60.0	87.9	83.3
	SD	25.4	23.6	25.4	16.5	--
	N	15	12	11	4	1
Voiceless	$\bar{x}$	61.3	63.6	--	--	90.2
	SD	19.3	27.2	--	--	1.4
	N	18	27			3
Initial	$\bar{x}$	65.9	65.5	84.2	87.9	88.5
	SD	19.3	24.5	10.6	16.5	3.6
	N	15	18	2	4	4
Final	$\bar{x}$	56.7	60.2	54.6	--	--
	SD	23.5	27.6	24.8	--	--
	N	18	21	9		
Total	$\bar{x}$	60.9	63.3	60.0	87.9	88.5
	SD	21.9	25.9	25.4	16.5	3.6
	N	33	39	11	4	4

tion, plosives, fricatives, and nasals appear to be more easily identified when in the syllable-initial position.

The importance of the selection of response foils is shown by a comparison of the percent correct identifications of the voiced, syllable-initial plosives /bɑ, dɑ, gɑ/ in subtests 6 and 7 (Tables 8 and 9). With one exception, identification percentages are greater when the alternative responses consist of liquids, glides, and affricates, as in subtest 6, rather than the fricatives and nasals of subtest 7. The one exception is the syllable /gɑ/ of subtest 6 which has a slightly lower percentage (85.0%) in noise than its counterpart in subtest 7 (86.7%). Note, however, that the greatest confusion in both cases is with another plosive, /dɑ/.

Analysis of Confusions

Inspection of the remaining portion of the matrix reveals the patterns of confusions for the various syllable subtests in both quiet and noise. Target/response pairs with confusion percentages of 10% or greater were considered for analysis. Table 16 contains a summary of the confusions meeting this criterion, according to syllable subtest type.

TABLE 16. Number and Percent of Consonant Confusions of 10% or Greater Error in Both Quiet and Noise Conditions, as a Function of Consonant Voicing, Consonant Position and Vowel Context

	Quiet		Noise	
	Number of Confusions	Percentage	Number of Confusions	Percentage
Voiced	35	47.9	53	47.3
Voiceless	38	52.1	59	52.7
Initial	23	31.5	37	33.0
Final	50	68.5	75	67.0
/ a /	22	30.1	29	25.9
/ u /	29	39.7	45	40.2
/ i /	22	30.1	38	33.9
Total	73	100.0	112	100.0

In both quiet and noise conditions, the number of confusions is fairly evenly distributed for voiced and voiceless consonants and for the three vowel contexts, with slightly more confusions for voiceless consonants and for subtests containing the vowel /u/. However, syllables in which the consonant occurs in the final position appear to be highly susceptible to confusion in both quiet and noise.

Syllable pairs with confusion percentages of 10% or greater were then categorized according to the type of error made. Confusions in which the target and the response syllables were produced by different manners of articulation (e.g., confusions between fricatives and plosives) were designated as manner confusions. Confusions between two fricatives, two plosives, or two nasals were designated as place confusions. Figures 16 and 17 summarize the numbers and types of confusions in each category for the quiet and noise conditions. The overall number of place and manner confusions appears to be approximately the same in the quiet condition, as shown by Figure 16. Figure 17 shows that in the quiet condition the most common place confusions are between pairs of fricatives, followed by confusions

Fig. 16. Frequency of place and manner confusions for both quiet and noise conditions.

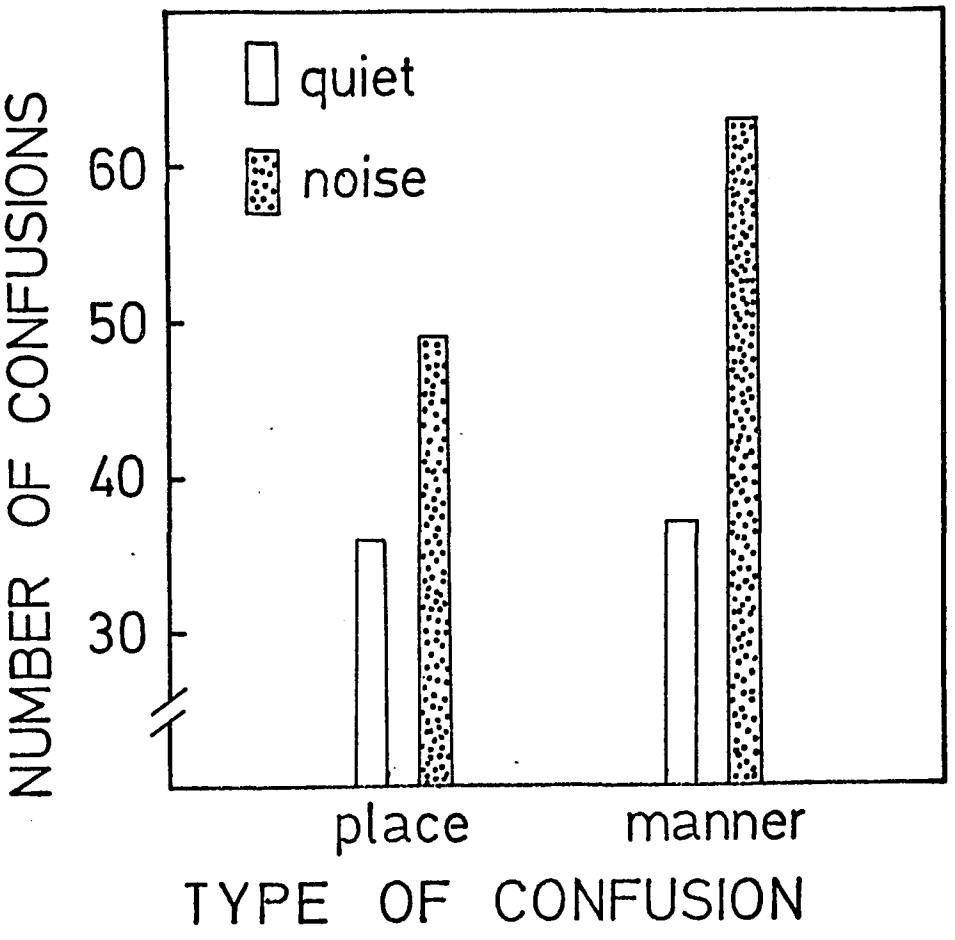
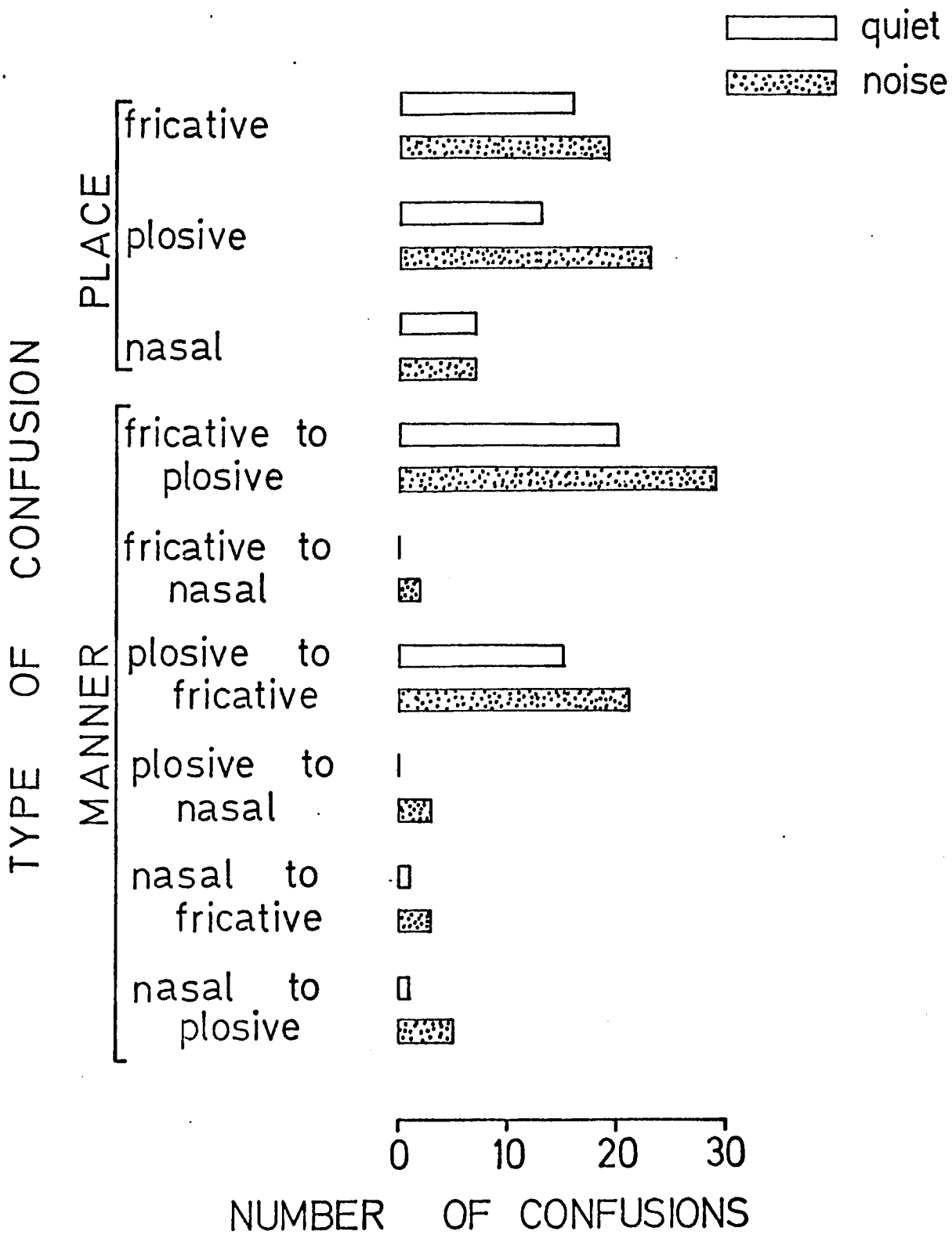

quiet
noise
NUMBER OF CONFUSIONS
60
50
40
30
place
manner
TYPE OF CONFUSION

Fig. 17. Frequency of different types of place and manner confusions for both quiet and noise conditions.


quiet
noise
TYPE OF CONFUSION
PLACE
fricative
plosive
nasal
MANNER
fricative to plosive
fricative to nasal
plosive to fricative
plosive to nasal
nasal to fricative
nasal to plosive
0
10
20
30
NUMBER OF CONFUSIONS

between plosives. Manner confusions in quiet are predominately between fricatives and plosives. In noise, Figure 16 reveals that the number of manner confusions increases to a much greater extent than is the case for place confusions in noise. In addition, Figure 17 shows that the most common place confusions in noise are between pairs of plosives, which is not the case in quiet. Ranking of manner confusion types remains the same in noise (i.e., fricative-to-plosive confusions most common), with the addition of several confusions between nasals and the other manner types.

It is important to note at this time that comparison of confusions between quiet and noise conditions, spoken of in these general terms, although interesting, may tend to be misleading. Comparison of the two conditions on the basis of numbers of confusions presents a highly simplified picture of this type of data. A more realistic analysis can be obtained from the interpretation of the confusions in terms of the actual percentages. To illustrate this point, the most commonly confused pairs of syllables, in terms of percent confusion, were selected from data for both quiet and noise. These major confusions are shown in Table 17.

TABLE 17. The Most Commonly Confused Pairs of Nonsense Syllables, in Terms of Percent Confusion, for Both Quiet and Noise Conditions

QUIET		NOISE	
Confusion	Percent	Confusion	Percent
iv/ið	50.0	θi/fi	52.8
θi/fi	41.7	if/iθ	43.3
if/i	36.7	bɑ/ðɑ	40.0
iθ/if	33.3	ɑf/ɑv	36.7
uŋ/um	33.3	up/uk	33.3
ið/iv	30.6	fi/pi	33.3
uv/ub	27.8	in/im	33.3
ip/it	25.0	hi/pi	33.3
pu/ku	25.0	im/ib	33.3
fu/su	25.0	uŋ/um	33.3
		ug/uv	33.3

For both conditions, more than half (60% and 55%) of the confusions are for voiced consonants, contrary to the trend suggested by the overall number of confusions ($\gtrsim$10%) as seen in Table 16. Major confusions seem to be more common in syllables containing the vowel /i/ and less common with the vowel /ɑ/ than would be inferred from the overall data of Table 16. One also finds an apparent contradiction between the data shown in Table 17 and the data in Figure 16. Figure 16 shows that number of place and manner confusions to be equivalent in quiet but very different in noise. Inspection of the major confusions in Table 17 reveals that in the quiet condition, 9 of the 10 confusions are of the place category (i.e., within the same manner class). In the noise condition, there are equal numbers of major confusions for place and manner.

Finally, it is true that, as expected, there are more confusions meeting the 10% criterion in the noise condition (112) than in the quiet condition (73). However, the increase in the number of confusions in noise is not simply the result of adding 39 more confusions to those isolated for the quiet condition. On the contrary, of the

73 confusions in quiet meeting the criterion, 37.0% of them are not among those isolated for the noise condition. In addition, it is quite understandable that certain confusions in noise would not be made in the quiet condition. However, it is more difficult to explain how 9 of the 112 confusions (8.0%) meeting the criterion in the noise condition showed higher percentages of confusions in the quiet condition.

The preceding analysis indicates that there is a complex interaction between error patterns for the quiet and noise conditions. Also, there are important interactions in error patterns within the commonly-used articulatory categories for the quiet and noise conditions. As noted previously, however, there appear to be few interactions between speech level and error patterns.

ACOUSTICAL MEASUREMENTS

Consonants and vowels which appeared (as either target or response) in confusions with percentages of 10% or greater were included in the acoustical analyses. This accounted for 72 of the 91 different syllables which make up the nonsense syllable subtests. Those syllables omitted

from the acoustical analyses on this basis were the syllables containing the fricative /ʃ/, the affricates /dʒ and tʃ /, the glides and liquids /j, r, l, w/, the plosives /bɑ, dɑ, gɑ/ of subtest 6, and the nasals /mɑ, nɑ/ of subtest 7. All other syllables were analyzed according to the procedures outlined in the Procedures chapter. The results for each measurement are presented in the following section.

Vowel Peak Frequency

The formant frequency of the vowel which was found to be the highest in amplitude (VPF) was recorded and is summarized in Appendix A. If the vowel peak-to-noise ratio was found to be 10 dB or worse, the peak with the highest peak-to-noise ratio was designated as the vowel peak frequency for the noise condition. These measurements also appear in Appendix A. The values for the quiet condition range from 227 Hz (F1) for the vowel in the syllable /un/ to 1172 Hz (F2) for the vowel in the syllable /ɑv/. The effect of the noise on this measurement is most clearly seen by a comparison of the values for subtest 3. The vowel peak frequency was associated with the first formant for all syllables in that subtest in the quiet condition. However, the spectral

characteristics of the noise resulted in a significantly poorer peak-to-noise ratio for the lower formant peaks than for the higher formant peaks. Thus, the vowel peak frequencies for these syllables in the noise condition are those associated with the second formant.

Second Formant Transition

Three measurements were made on the second formant transition. The origin frequency (ORIG) of the second formant transition, as well as the magnitude (MAG) and direction (DIR) of the transition was measured, according to the procedures described previously. The results of these three measurements are summarized for the quiet and noise conditions in Appendix B. The origin frequency values for the quiet condition range from 684 Hz for the origin of the transition in the syllable /uf/ to 2702 Hz for the origin of the transition in the syllable /iŋ/. The magnitude of the transitions in quiet range from 0 Hz for those syllables in which no transition could be isolated to 554 Hz for the transition in the syllable /si/. Direction of the transition was either rising (+), falling (-), or no change (0).

As noted in the Procedures section, if the noise spectrum exceeded the formant peak, this then affected the estimated start of the formant transition. A change in the origin of the frequency transition also produces a subsequent change in the magnitude of the transition. Consider the syllable /θi/ by way of an example. The extent of the transition in quiet is from the origin frequency 1790 Hz to 2279 Hz (magnitude = 489 Hz). However, the spectrum of the noise caused the previously-designated origin frequency to be masked. As the transition rises in frequency, the formant peak rises above the sharply-falling noise spectrum. Thus, the origin of the transition in noise appears to be higher in frequency. The extent of the transition in noise is from the new origin frequency 2018 Hz to 2279 Hz (magnitude = 261 Hz). If the noise is sufficient to mask the entire extent of the transition (as in /us/), a value of 0 Hz is assigned to the magnitude of the transition. Direction of the transition does not change in noise, except for the cases in which the entire transition is masked.

Consonant Spectral Peaks

Two highest peaks (CF1, CF2) in the frequency spectrum for each consonant were recorded. Measurements for the quiet and noise conditions are summarized in Appendix C. The decision to change the designated consonant peak frequency for the noise condition was made according to the criterion previously outlined. Values in quiet range from 391 Hz for the consonant in the syllable /pu/ to 8236 Hz for the consonant in the syllable /if/. The effect of the noise was most apparent in consonants which contained lower frequency spectral peaks, due to the spectral characteristics of the noise. When a change from the quiet measurement was indicated, due to the masking of a peak, the spectral peak which remained significantly above the noise was taken as the consonant spectral peak for the noise condition. Again, as a result of the spectral characteristics of the noise, the newly-designated peak was usually at a higher frequency than the measurement in the quiet condition.

Overall Consonant-Noise Bandwidth

Measurements of consonant-noise bandwidth (BW) in the quiet condition were carried out according to the

procedures previously described. The results of these estimates are summarized in Appendix D. Bandwidth of the consonant noise was calculated by subtracting the lower cutoff frequency from the upper cutoff frequency. The results range from a bandwidth of 400 Hz for the consonant in the syllable /ub/ to a bandwidth of 5267 Hz for the consonant in the syllable /af/.

Crossover Frequency

For the noise condition, the measurement of consonant bandwidth was replaced by that of crossover frequency (XF), as explained in the Procedures chapter. The results of these estimates, for the noise condition only, are summarized in Appendix E. The lowest crossover frequency was estimated at 667 Hz for the consonant /ŋ/ in the syllables /aŋ/ and /iŋ/. This implies that the spectrum of the consonant /ŋ/ above 667 Hz lies above the noise spectrum. The highest crossover frequency was estimated at 3533 Hz for the consonant in the syllable /iz/. In this case, the spectrum of /z/ below 3533 Hz was embedded in noise. Above 3533 Hz, the consonant spectrum lied above the noise spectrum.

Total Energy of Consonant and Vowel

Results of the consonant energy (CE) and vowel energy (VE) calculations appear in Appendix F. The consonant energy values reported are in dB re: the weakest consonant (/hi/). Relative to that level, the consonant having the highest total energy is the consonant in the syllable /sɑ/ (30.5 dB re: /hi/).

Results of the vowel energy measurements, recorded in dB relative to the weakest vowel, range from 0.0 dB for the vowel in the syllable /iθ/ to 9.5 dB for the vowel in the syllable /iʒ/. Note that the range of values for the vowels is much narrower than that for the consonants. In addition, the range of values of vowel energy for syllables within each subtest is narrower than for consonants.

Consonant-to-noise ratio (C/N) and vowel-to-noise ratio (V/N) are summarized in Appendix G. Measurements of consonant-to-noise ratio ranged from -21.8 dB for the consonant in the syllable /ɑk/ to 9.9 dB for the consonant in the syllable /si/. The lack of direct relationship between consonant energy and consonant-to-noise ratio is due to the variability in the average level of the background

noise which accompanies each syllable. This holds true for the vowel-to-noise ratio measurements which range from -2.0 dB for the vowel in the syllable /hu/ to 12.6 dB for the vowel in the syllable /ɑb/.

Durations

Durations of the consonants (CD), vowels (VD), and closures (CLD) are presented in Appendix H. The results of these measurements are as follows. The shortest consonant (13 msec) is the /b/ which appears in the syllable /bɑ/. The longest consonant (374 msec) is the /ŋ/ in the syllable /iŋ/. As is the case for consonant energy, the range of values within each subtest is large for consonant duration. Vowel durations, on the other hand, appear to be quite similar within each subtest. The shortest vowel (133 msec) is the vowel in the syllable /up/. The longest vowel (372 msec) is the vowel /ɑ/ which appears in the syllable /dɑ/.

Closure durations range from 76 msec in the syllable /ig/ to 145 msec for the syllable /up/. Differences in closure durations between syllables in the same subtest are small.

The same set of duration measurements represent both the quiet and noise conditions.

PREDICTIONS

Predicting Percent Correct Identifications

The data presented in the first two sections of this chapter served as the basis for further analyses. The purpose of these analyses was to determine the set of acoustical variables which would provide the best predictions of percent correct identification of nonsense syllables in quiet and in noise.

The first step in the analysis was the computation of Pearson product-moment correlation coefficients. Table 18 provides a list of the acoustical variables as measured in quiet that were most highly correlated with the percentage of correct identifications in quiet. A large positive correlation (e.g., between magnitude and percent correct in subtest 1) implies that the two variables tend to increase (or decrease) together. That is, as the magnitude of the second formant transition increases, the identification percentage increases. Coefficient values closer to zero imply the absence of a monotonic relationship between two variables. A negative correlation (e.g., between vowel duration and percent correct in subtest 4) implied an inverse

TABLE 18. Pearson Correlation Coefficients and Significance Levels for Selected Acoustic Variables Measured in Quiet Correlated with Percent Correct Identification in Quiet, By Syllable Subtest

Syllable Subtest	Acoustic Variable	r	Significance
1	MAG	0.82	*
2	VPF	0.80	
3	CF1	-0.80	
4	CF2	-0.70	*
	VD	-0.86	**
5	CF2	0.73	
	CD	0.68	
7	VD	0.74	
	ORIG	0.73	
8	CF2	0.71	
9	CE	0.55	
10	MAG	0.48	
11	MAG	0.55	

* < .05
** < .01

relationship between the two variables. That is, as the duration of the vowel becomes longer, the percentage of correct identification decreases. Note also that a majority of the correlation coefficients with acoustical variables in quiet did not reach either the 0.05 or 0.01 level of significance.

Table 19 presents the Pearson correlation coefficients between the acoustical variables as measured in the noise condition and the percentage of correct identifications for each subtest in noise. Note that consonant energy (CE) and consonant-to-noise ratio (C/N) appear to have large positive correlations with percent correct in noise. Vowel energy (VE) and vowel-to-noise ratio (V/N), however, are negatively correlated with percent correct in noise. As for consonant spectral peaks (CF1 and CF2), for some subtests, as the peak frequency becomes higher the percent correct identification also becomes greater (subtests 5 and 8 in quiet; subtests 5 and 10 in noise). In other subtests, an increase in the frequency of the peak is associated with a decrease in the identification percent (subtests 3 and 4 for quiet and noise). One possible explanation for this negative

TABLE 19. Pearson Correlation Coefficients and Significance Levels for Selected Acoustic Variables Measured in Noise Correlated with Percent Correct Identification in Noise, By Syllable Subtest

Syllable Subtest	Acoustic Variable	r	Significance
1	VN	-0.77	
2	CE	0.95	**
	CN	0.87	*
3	CF1	-0.84	*
4	CF2	-0.70	*
5	CF2	0.83	*
	XF	0.91	**
	ORIG	0.84	*
7	CN	0.79	*
8	ORIG	0.91	*
	CN	0.68	
9	VD	0.71	
	CE	0.87	*
	CN	0.91	**
10	CF1	0.44	
11	VE	-0.64	

* < .05
** < .01

correlation emerges from an examination of the frequency-gain characteristics of the transducer used in the listening experiment (see Figure 6). It is possible that the information provided by the consonant spectral peak variable was reduced when its value rose above the upper cutoff frequency of the earphone.

The next step in the analysis was to use the information provided by the correlational analyses to find a simple summary of the linear relationship between the acoustical variables and percent correct identification. The statistical procedure known as least-squares linear regression is commonly used for this purpose. In regression analyses, R^2, the square of the correlation coefficient, is used as a measure of the proportion of the variance in one variable explained by the other variable. Variance is "explained" in correlation analyses by linking the variation in one variable with the variation in another.

The simple regression procedure would not be appropriate for the type of analyses needed for this study. The multiple regression, on the other hand, is a procedure which finds the best linear prediction equation and its accuracy

of prediction from a number of independent variables (i.e., acoustical measurements) which are weighted and summed to obtain the best prediction of the dependent variable (i.e., percent correct identifications). From the multiple regression, a statistic is generated which relates the accuracy of the prediction equation and how much of the variation in the dependent variable is accounted for by the combined influence of the independent variables. It may evaluate the contribution of a specific variable and, based on this evaluation, may delete the independent variables which do not add significantly to the accuracy of the prediction. Evaluation of the F ratio for the regression equation provides a significance test for these measures, i.e., the probability that the measured R^2 of the equation is a result of sampling fluctuation or errors. The F ratio for the regression coefficients (the slope of the regression line, or B) gives a measure of statistical significance for each variable in the prediction equation treated individually.

The multiple linear regression procedure used in the present study is known as the Maximum R^2 Improvement

procedure (Barr, et al., 1976). The first step in this procedure is to determine the one-variable regression equation which provides the highest R^2, or the highest proportion of the explained variability. The next step is to find that combination of two variables which would result in an equation with maximum increment in R^2. This continues for three-variable prediction models, four-variable models, and so on until all variables are in the equation.

For this study, variables were allowed into the equation only until such time as the entrance of a variable resulted in an increase in the significance of the F ratio for the regression equation to be above a significant level (0.1). Entry into the model was also restricted if the variable resulted in statistically insignificant regression coefficients. Finally, a significant increment in R^2 had to be a result of the variable's entry in the equation for it to be included.

Table 20 shows a list of the acoustical variables (as measured in quiet) in the equations used to predict percent correct identification in quiet. The R^2 is shown for each model (either the 1, 2, or 3 variable model) and

TABLE 20. R^2 and Significance Levels for Acoustic Variables Measured in Quiet Entered into the Equation Predicting Percent Correct Identification in Quiet, by Syllable Subtest (significance refers to regression equation containing the entered variables)

Syllable Subtest	Acoustic Variable	R^2	Significance
1	MAG	0.676	*
2	VPF	0.646	
	BW,VE	0.897	*
	BW,VPF,VE	0.970	*
3	CF1	0.647	
	CF1,CE	0.980	**
4	VD	0.748	**
	VD,CF2	0.833	**
	VD,CF2,CF1	0.899	**
5	CF2	0.526	
7	(CF2)	(0.537)	
8	BW	0.695	*
	BW,CF1	0.797	
9	(CE)	(0.307)	
10	(MAG)	(0.235)	
11	DIR	0.384	

* < .05
** < .01

the level of statistical significance which was met for the prediction equation. Regression equations for subtests 7, 9, and 10, with R^2 reported in parentheses, did not reach the 0.1 level of significance. The "best" predictor of percent correct in quiet appears to be the prediction of correct scores for subtest 3 by the values for the consonant spectral peak (CF1) and the consonant energy (CE), ($R^2 = 0.980$).

Table 21 presents the acoustical variables (as measured in noise) used to predict the dependent variable of percent correct identification in noise. It appears that there are more "good" predictors of percent correct in noise than for the quiet condition. The two "best" predictions of correct scores in noise appear to be the prediction of correct scores for subtest 8 by the values for consonant-to-noise ratio (C/N), consonant energy (CE), and consonant duration (CD), ($R^2 = 0.999$), and the prediction for subtest 5 by the values for consonant spectral peak (CF1) and consonant energy (CE), ($R^2 = 0.993$).

The final step in the prediction of percent correct identification was to determine the one set of acoustical

TABLE 21. R^2 and Significance Levels for Acoustic Variables Measured in Noise Entered into the Equation Predicting Percent Correct Identification in Noise, by Syllable Subtest (significance refers to regression equation containing the entered variables)

Syllable Subtest	Acoustic Variable	R^2	Significance
1	VN	0.662	
2	CE	0.903	**
	CE,CN	0.967	**
3	CF1	0.703	*
	CF1,CN	0.936	*
	CF1,CN,MAG	0.985	*
4	CF2	0.483	*
	CF1,VD	0.625	
5	XF	0.828	**
	CF1,CE	0.993	***
7	CN	0.626	
	CN,CE	0.964	**
8	CN	0.872	*
	CN,ORIG	0.943	
	CN,CE,CD	0.999	**
9	CN	0.823	**
	CN,VE	0.917	**
10	(CF1)	(0.196)	
11	VE	0.414	
	CE,VE	0.612	

* < .05
** < .01
*** < .001

variables and their associated weightings which would provide the best equation to predict percent correct identification over all subtests. Acoustical variables in the prediction equations for quiet (Table 20) were rank ordered according to number of occurrences. A similar ranking was made for acoustical variables (from Table 21) in the equations predicting the noise data. Variables occurring in the most number of equations were selected. The two sets of selected variables are as follows:

Predicting Percent Correct Identification in Quiet:

First Consonant Spectral Peak
Second Consonant Spectral Peak
Consonant Energy
Magnitude of the Second Formant Transition
Overall Consonant-Noise Bandwidth

Predicting Percent Correct Identification in Noise:

Consonant-to-Noise Ratio
Consonant Energy
First Consonant Spectral Peak
Vowel Energy

A comparison of variables isolated for predicting percent correct identification in the quiet vs. the noise conditions reveals some similarities. Both predictions use the variables of consonant energy and consonant spectral

peak. Consonant-to-noise ratio is an additional variable included for the prediction of percent correct in noise.

Predicting Percent Confusions

The data from the acoustical analyses were also used to determine the set of acoustical variables which would provide the best possible prediction of the percent of confusions for selected pairs of syllables in quiet and in noise. The syllable pairs selected were the major confusion pairs (confusion percentages of 10% or greater) in addition to the target of each, paired with all of the other response alternatives. For example, for subtest 2 in noise, the /up/uk/ confusion occurs (percentage = 33.3%). This syllable pair is analyzed, along with the pairs /up/uf/, /up/ut/, /up/uθ/, /up/us/ having confusion percentages of 20.0%, 13.3%, 10.0%, and 0.0%, respectively.

All analyses described in the previous section were performed as stated, with the following differences. The The dependent variables were, in this instance, percent confusion in quiet or percent confusion in noise. The independent variables were the absolute differences between the acoustical measurements for the target and response

syllables in each pair. In the above example, the data set would include the five measures of percent confusion for the five pairs. The remaining data are the differences between the five pairs of syllables in terms of consonant energy, vowel energy, consonant spectral peak, etc.

The predictions of percent confusion were carried out the same way as for the predictions of percent correct identification. First, Pearson correlation coefficients were determined between the "acoustical-difference" variables and percent confusion in both quiet and noise. The largest coefficients for each condition are reported in Tables 22 and 23. In almost all cases, the variables are negatively correlated. That is, the smaller the difference between the acoustic variables, the larger the percent confusion. The largest negative correlations are with consonant energy (CE) for subtest 3 in quiet ($r = -0.62$) and consonant-to-noise ratio (C/N) for subtest 3 in noise ($r = -0.73$).

The variables in the equations predicting percent confusion in both quiet and noise appear in Tables 24 and 25. The "best" prediction in the quiet condition appears to be

TABLE 22. Pearson Correlation Coefficients and Significance Levels for Selected Acoustic-Difference Variables Measured in Quiet Correlated with Percent Confusion in Quiet, by Syllable Subtest

Syllable Subtest	Acoustic-Difference Variable	r	Significance
1	VD	-0.47	
2	VPF	-0.56	**
	CE	-0.54	**
	BW	-0.44	*
3	CE	-0.62	*
	CD	-0.55	*
4	ORIG	-0.43	
5	CE	-0.34	
7	ORIG	-0.39	*
	MAG	-0.32	
8	MAG	0.37	
9	VPF	-0.39	
10	CE	-0.39	**
	CD	-0.29	*
	VE	0.25	
11	CD	-0.42	**
	VD	-0.28	

* < .05
** < .01

TABLE 23. Pearson Correlation Coefficients and Significance Levels for Selected Acoustic-Difference Variables Measured in Noise Correlated with Percent Confusion in Noise, by Syllable Subtest

Syllable Subtest	Acoustic-Difference Variable	r	Significance
1	XF	0.21	
2	VPF	-0.52	**
	CD	-0.36	
	CE	-0.59	**
	CN	-0.53	**
3	CD	-0.46	
	CE	-0.70	**
	CN	-0.73	**
4	CF1	-0.32	
	ORIG	-0.45	*
	DIR	-0.31	
5	CF1	-0.54	**
	CF2	-0.52	*
	XF	-0.47	*
	CE	-0.47	*
	ORIG	-0.36	
7	ORIG	-0.53	**
	MAG	-0.53	**
8	VPF	0.52	**
	CD	-0.30	
	VD	0.31	
	VE	-0.35	*
	MAG	-0.29	
	DIR	0.33	
	VN	0.39	*
9	CE	-0.32	
	CN	-0.46	*

TABLE 23. (Continued)

Syllable Subtest	Acoustic-Difference Variable	r	Significance
10	XF	-0.23	
	VD	-0.30	*
11	CF2	-0.32	*
	XF	-0.25	*
	CD	-0.37	**
	VD	-0.31	*

* < .05
** < .01

TABLE 24. R^2 and Significance Levels for Acoustic-Difference Variables Measured in Quiet Entered into the Equation Predicting Percent Confusion in Quiet, by Syllable Subtest (significance refers to regression equation containing the entered variables)

Syllable Subtest	Acoustic-Difference Variable	R^2	Significance
1	VD	0.220	
2	VPF	0.319	**
	VPF,CD	0.361	**
3	CE	0.386	*
	CE,CF1	0.636	**
	CD,CE,CF1	0.715	**
4	ORIG	0.185	
5	CE	0.119	
7	ORIG	0.155	*
	ORIG,CF1	0.227	*
8	(MAG)	(0.141)	
9	(VPF)	(0.150)	
	(VPF,BW)	(0.292)	
	CF2,BW,CE	0.685	*
10	CE	0.153	**
	CE,VE	0.206	**
	CE,VE,ORIG	0.266	**
	CE,ORIG,VE,MAG	0.299	**
11	CE	0.176	*
	CD,CF2	0.327	**
	CD,VD,CF2	0.350	**

* < .05
** < .01

TABLE 25. R^2 and Significance Levels for Selected Acoustic-Difference Variables Measured in Noise Entered into the Equation Predicting Percent Confusion in Noise by Syllable Subtest (significance refers to regression equation containing the entered variables)

Syllable Subtest	Acoustic-difference Variable	R^2	Significance
1	(CE)	(0.090)	
2	CE	0.346	**
	CE,DIR	0.441	**
	CE,DIR,CD	0.511	**
	CE,CD,MAG,CN	0.581	**
3	CN	0.534	**
	CN,CF1	0.683	**
	CE,CF1,VPF	0.802	***
	CE,VPF,CN,CF1	0.841	***
4	ORIG	0.202	*
	ORIG,CF1	0.379	*
	ORIG,CF1,CF2	0.470	*
5	CF1	0.288	**
	CF1,CE	0.469	**
	CF1,CE,ORIG	0.556	***
7	ORIG	0.208	
	ORIG,CN	0.318	
8	VPF	0.270	**
	VPF,CD	0.331	**
	VPF,CD,CN	0.396	**
9	CN	0.207	*
	CN,VE	0.270	*
	VE,CN,CD	0.352	**
	CN,CD,VE,CF1	0.405	**
10	VD	0.114	*
	VD,MAG	0.157	*
	VD,MAG,XF	0.210	*
11	CD	0.136	**
	CD,CF2	0.215	**
	CD,CF2,VN	0.237	**

* .05
** .01
*** .001

the equation for subtest 3, containing the acoustic-difference variables of consonant duration (CD), consonant energy (CE), and consonant spectral peak (CF1), (R^2 = 0.715). For the noise condition, the same subtest (3) appears to have the best prediction (R^2 = 0.841). However, this prediction equation contains the four acoustic-difference variables of consonant-to-noise ratio (C/N), consonant spectral peak (CF1), consonant energy (CE), and vowel peak frequency (VPF).

The acoustic-difference variables occurring in the most equations were isolated for quiet and noise and were used to generate equations and coefficients which could best predict the overall data. These variables were as follows:

Predicting Percent Confusion in Quiet:

Consonant Energy
Consonant Duration
Origin of the Second Formant Transition
Vowel Duration
First Consonant Spectral Peak
Second Consonant Spectral Peak
Magnitude of the Second Formant Transition

Predicting Percent Confusion in Noise:

Consonant-to-Noise Ratio
Consonant Energy
Consonant Duration
First Consonant Spectral Peak
Origin of the Second Formant Transition

A comparison of the acoustic-difference variables used to predict percent confusion in the quiet vs. the noise conditions shows four common variables: consonant energy, consonant duration, consonant spectral peak, and origin of the second formant transition. As was the case for predicting percent correct identification, the consonant-to-noise ratio variable was included in the noise prediction of confusion. In addition, note that the duration variables were included in predicting percent confusions but were not used in the equations for predicting percent correct.

CHAPTER V

DISCUSSION

Three major areas for discussion emerge from the results of this study: (1) the pattern of correct identifications and confusions of the nonsense syllables in quiet and in noise, (2) measurement of the acoustic characteristics of the nonsense syllables and the noise, and (3) a statistical model for predicting correct identifications and confusions from the acoustic analyses. The following chapter contains a discussion of the data obtained in this study as they relate to these three areas.

PATTERN OF CORRECT IDENTIFICATIONS AND CONFUSIONS

The analyses of intelligibility and of correct identifications may be compared to the findings of other studies. Inspection of comparable confusion matrices reported by Wang and Bilger (1973) reveals a similar hierarchy of correct identifications with respect to

consonant class for both quiet and noise conditions. The rank ordering of the consonant classes was quite different, however, from that reported by Horii, et al. (1971) for consonants masked by white noise. Nasal consonants had the lowest overall intelligibility score in noise in the present study. In white noise, however, Horii, et al. report nasal consonants to have the second highest average intelligibility score, second only to the liquids and glides. Affricates were reported to have the lowest scores in white noise, but were most easily identificed in the present study. Differences such as these could be accounted for by noting the differences between the spectrum of white noise and the spectrum of the noise used in the present study, most notably the additional low frequency energy. In fact, the hierarchy of the consonant classes for the present study is more similar to the hierarchy for the "envelope-noise" data of Horii, et al., in which the differences in masking between the various consonants classes was much smaller.

The effect of the accompanying vowel on consonant intelligibility was found to be significant by Wang and

Bilger (1973) for both quiet and noise. However, their data showed consonant intelligibility to be highest when paired with the vowel /u/, and lowest with the vowel /ɑ/. This finding is opposite to the results of the present study in which highest scores were obtained for syllables containing the vowel /ɑ/ and lowest scores for syllables with the vowel /u/ (see Figure 14). The acoustic analysis of the syllables used in the study showed that the total energy of the /ɑ/ vowel, averaged over all consonants exceeded that of the /u/ vowel. This helped most for the speech in noise case where the improvement for the /ɑ/ vowel was the greatest.

The analyses performed on the confusion matrices for the quiet condition were compared to similar data for four normal-hearing control subjects reported by Bilger and Wang (1976). Comparisons for the noise conditions were made with the data reported in Wang and Bilger (1973). Although there were large differences in the absolute frequency of confusions between the Bilger/Wang studies and the present study, the pattern of confusions for each syllable type was similar in many respects. For eaxample,

in both sets of data more confusions were found between syllables containing voiceless consonants than voiced. Similarly, in both sets of studies more confusions were found between pairs of VC syllables than between pairs of CV syllables. With respect to the Bilger/Wang data, it should be noted that although the number of CV syllables showing confusions was smaller, when a confusion occurred the error rate was very high.

The differences in error patterns between the studies related primarily to the role of the accompanying vowel, as was noted above, and to differences in the relative frequencies of place and manner errors. Confusions involving manner of articulation were not as common in the Bilger/Wang studies as was the case in the present study. The few manner errors reported in the Bilger/Wang data were of the fricative-to-plosive type. In this study, manner errors of all kinds were observed (see Figure 17); however, by far the most frequent manner errors were between fricatives and plosives. Confusion matrices reported by Miller and Nicely (1955) also show manner confusions to be quite rare. The matrix generated

using environmental conditions closest to those used in the present study revealed only one large error of the manner type (/bɑ/vɑ/). The differences between the studies could be attributed in part to differences in the spectra of the masking noise that were used. The cafeteria noise used in this study had more intense low-frequency components which presumably had greater masking effects on the low-frequency nasal component, and other manner of articulation cues.

Percentages of confusion in quiet were computed for the Bilger/Wang data*. Comparisons were made between these data and the syllable pairs having the highest frequencies of confusion in the present study. These comparisons are summarized in Table 26. For the quiet condition, only two of the confusions (/θi/fi/, /iθ/if/) showed comparable high percentages. Three cases were comparable for the noise condition, specifically /θi/fi/, /up/uk/, and /hi/pi/.

Comparisons of this kind serve to point out that

---

* Percentages reported for Bilger and Wang (1976) and for Wang and Bilger (1973) were derived from matrices for their four syllable sets, CV-1, VC-1, CV-2, and VC-2, in each of the three vowel contexts. The data were made available to the author by Dr. Marilyn D. Wang.

TABLE 26. Percent Confusion for Pairs of Syllables Found to be the Most Commonly Confused Pairs in the Present Study (Comparsion data derived from Wang and Bilger, 1973; Bilger and Wang, 1976)

QUIET				NOISE			
Confusion	Dubno (1978)	Wang & Bilger (1973) CV-1/VC-1	CV-2/VC-2	Confusion	Dubno (1978)	Bilger & Wang (1976) CV-1/VC-1	CV-2/VC-2
iv/iʒ	50.0	1.7	3.3	θi/fi	52.8	53.2	--
θi/fi	41.7	83.6	--	if/iθ	43.3	3.1	16.1
if/iθ	36.7	1.7	5.0	bɑ/ʒɑ	40.0	8.5	--
iθ/if	33.3	15.0	21.7	ɑf/ɑv	36.7	11.5	18.8
uŋ/um	33.3	--	0.0	up/uk	33.3	29.0	--
iʒ/iv	30.6	1.7	3.3	fi/pi	33.3	11.0	11.1
uv/ub	27.8	0.0	0.0	hi/pi	33.3	--	51.2
ip/it	25.0	0.0	--	in/im	33.3	--	1.6
pu/ku	25.0	0.0	--	im/ib	33.3	--	0.9
fu/su	25.0	0.0	1.7	uŋ/um	33.3	--	6.8
				ug/uv	33.3	2.1	2.4

differences between studies of this type can be quite large. Differences are to be expected when one considers the differences in procedures and environmental conditions used to generate the two sets of data. The speech stimuli in the Bilger/Wang studies controlled for duration of the syllable (511 msec) and rise time (1 msec), conditions which were not controlled in the present study. Presentation level overlapped to some degree but were not identical. The levels in the Bilger/Wang studies were on the order of 30 dB above the highest level used in this study. Also, the Bilger/Wang studies matched the speech stimuli for level, thus eliminating that variable as a cue for identification. Results of the current investigation indicate that differences in overall energy between consonants may be a crucial variable for reducing confusion rates. Other procedural differences were the use of feedback, the elimination of the carrier phrase, and the use of a large (16 item) response set in the Bilger/Wang studies. There were also differences in the noise spectrum, the effects of which have already been noted.

Despite the differences between the studies, the

data showed some similarities and several generalizations may be made. Firstly, the confusion matrices show distinct patterns. These confusion patterns are fairly complex, as was shown in their description in the Results chapter. Secondly, the confusion patterns appear to be similar as a function of speech presentation level, although there are a few exceptions to this finding. Thirdly, the confusion patterns for different sets of nonsense syllables are differentially affected by the addition of background noise. Furthermore, vowel context appears to be an important factor, as certain confusions occur commonly with one vowel, but rarely with another. The results also indicate that the set of response alternatives has a significant effect on the patterns of confusion.

ACOUSTICAL MEASUREMENTS

Measurements were made of acoustical variables believed to be potentially important for predicting the patterns of correct identifications and confusions of of nonsense syllables in quiet and in noise. Predictions of the confusion matrix were then made based on these

measurements. The effectiveness of these predictions is heavily dependent upon the accuracy of the acoustic measurements. One must keep in mind that the measurements presented in this study are of a specific speech sample and should not be interpreted as being representative of the average male speaker. However, the value of these measurements is that they pertain to the speech and noise samples used in generating the confusion matrices.

Digital speech processing techniques were used in obtaining most of these measurements, since these procedures have been found to be reasonably accurate and highly repeatable. The measurements taken directly from the computer-generated spectra included vowel formant frequencies, measures of second formant transition, and consonant spectral peaks. On average, the results for formant frequency are comparable to those reported by Peterson and Barney (1952), and to those reported by Lehiste and Peterson (1961) for second formant transitions. Spectral peaks are similar to those reprted by Halle, et al. (1957) for plosive consonants, and to those reported by Hughes and Halle (1956) for fricative consonants. Some differences were found, of course,

as might be expected when measurements are taken from the recordings of only one speaker.

Estimates of overall consonant-noise bandwidth and crossover frequency were made using procedures developed specifically for this study. There was no other source of this type of measurements for direct comparison.

Measurements of total energy of the consonant and of the vowel, and calculation of consonant-to-noise and vowel-to-noise ratios were derived from the results of computer averaging performed on these segments. The relative energy levels of the consonants were averaged over the three vowel contexts and compared to similar relative levels derived from Fletcher (1953). The two sets of measurements were rank ordered from the most intense to the least intense consonants. The consonants tended to order themselves in a similar manner for the two studies. The largest exceptions were as follows. Measurements of /s, z, v/ from the present study resulted in substantially higher rankings of relative energy for these consonants than measurements of Fletcher. On the other hand, measurements of /t, k, h/ resulted in lower

rankings for the present study for these consonants than for Fletcher. Total energy for the three vowels used in this study were averaged across all consonant contexts and compared to the relative energy levels reported by Fletcher (1953). Both studies found the vowel /ɑ/ to be the most intense of the three, with /u/ and /i/ having similar lower energy levels.

As stated previously, the durational measurements made from digital speech waveforms were at least as accurate and as reliable as one could expect from spectrographic analysis. Differences in vowel duration as a function of consonant environment were similar to those reported by House and Fairbanks (1952) and Raphael (1972). Consonant durations were comparable to those reported by Fertig (1976).

The computer-generated spectral represenation of the background noise indicates that, on average, the noise is stable from subtest to subtest, in terms of overall level and spectral content. Therefore, differences in confusion patterns between subtests in the noise condition are probably due to factors other than between-subtest changes in noise level or spectrum.

In summary, the estimates of the acoustic characteristics of the speech stimuli and background noise are fairly consistent with other published data. In the future, however, further use of digital speech processing techniques, which are constantly being refined, should result in even greater levels of accuracy and reliability in specifying the acoustic characteristics of speech. In addition, as the use of linear predictive coding becomes more common in the spectral analysis of speech, it is likely that the ideal parameters for the use of this analysis technique will emerge.

PREDICTING CORRECT IDENTIFICATIONS AND CONFUSIONS

Sets of acoustic variables were isolated which provided the best prediction of percent correct identifications and percent confusions of nonsense syllables in quiet and in noise. These sets of variables were taken from the results of multiple linear regression analyses. The analyses provided regression equations in which the variables included account for the largest proportion of the variance in the prediction of either percent correct identification

or percent confusion. The variables used to predict scores for correct identification in quiet and in noise were the absolute values of the measured speech parameters, as measured in quiet and in noise. The variables used to predict the percentage of confusion between two syllables were the differences between the acoustic measurements between the two syllables. This was done both for speech in quiet and in noise.

The accuracy of the prediction equations resulting from this analysis is necessarily limited by the individual acoustic variables that were chosen for measurement. The selection of variables was made on the basis of previous research in the areas of phoneme reception and speech intelligibility in both quiet and noise, as outlined in the Review of Literature chapter. Certain measurements could, perhaps, be eliminated from future analyses of this type, or measured according to different criteria, as will become evident.

Inspection of the variables included in the regression equations (Tables 20, 21, 24, 25) reveals that for each subtest, the largest proportion of the variance (in quiet

and in noise) is accounted for by different sets of variables. However, three variables stand out (consonant energy, consonant spectral peaks, and consonant-to-noise ratio) as being particularly important, since they were isolated in a majority of the subtests. A large proportion of the variance is accounted for by acoustic variables, but different regression equations with different coefficients are used to predict the results for each subtest in each condition.

The implications of this finding is that the information provided by acoustic variables may be necessary for reducing confusions between certain sets of syllables but are of less value in other sets. Similarly, a check of the first variable entered into each regression equation is an indication of which single variable accounted for most of the variance (see Tables 20, 21, 24, 25). Again, these variables differed as a function of syllable subtest and listening condition.

In comparing the accuracy and reliability of the four sets of predictions (percent correct identifications and percent confusions in quiet and in noise), it appears

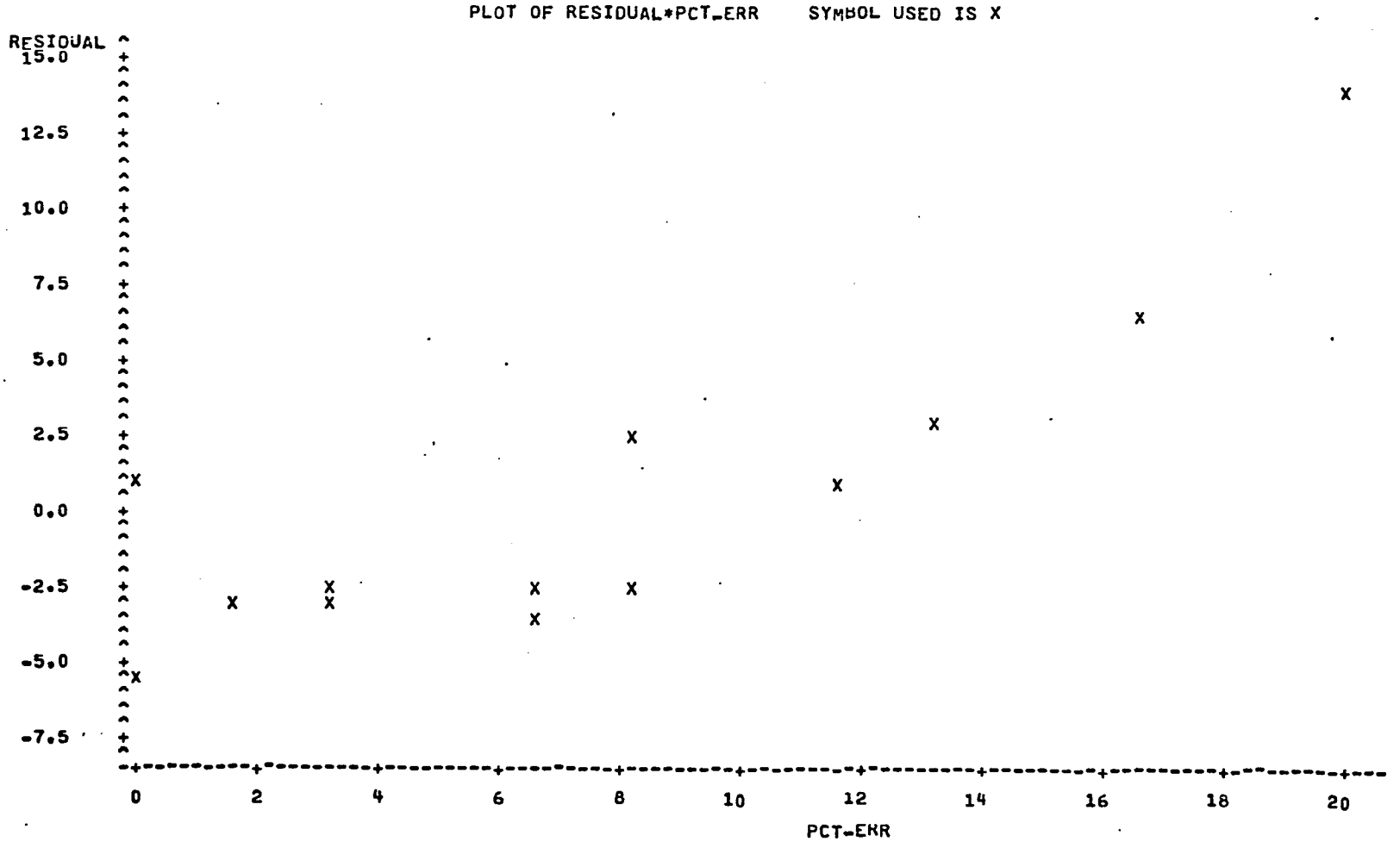
that, on average, a larger proportion of the variance is accounted for in predicting percent correct than in predicting confusions. That is, one may predict correct responses from a knowledge of acoustic information with much greater accuracy than one can predict percent confusion.

The prediction of correct identifications is basically much easier than that of predicting confusions. Firstly, only absolute measurements of the acoustic variables are needed, whereas for the confusion predictions a difference between the acoustic variables for the target and response stimuli were used in the equations. The latter were necessarily less precise. Further, the differences computed for certain variables were smaller than the values known to be just detectable by the auditory system. Perhaps in future analyses, differences less than the difference limen for certin acoustic variables (intensity, duration, spectral change, etc.) should be kept from contributing to the prediction, as such differences are probably not providing useful information to the auditory system. One area of research which needs to be explored

before such adjustments could be made in the procedure would be the accurate estimation of difference limen for spectral and temporal characteristics of real speech.

An analysis of the differences between observed and predicted confusions shows a mean standard deviation of 6.82 percentage points. Further, a trend was observed in that the larger the measured confusion, the larger the deviation between observed and predicted confusion rates. This trend can be seen in Figure 18, which is a plot of the residuals as a function of percent confusion for subtest 1 in the quiet condition. The prediction tended to underestimate the size of the largest confusion. This trend may be due in part to a statistical property of the data, in that cells with very low confusion rates have better test-retest repeatability than those with higher error rates. Another factor which should be considered is that whenever the frequency of confusion was high, one of the several acoustic variables considered might have actually shown a large difference, thus affecting the predicted score. However, it may be that when several acoustic cues occur simultaneously, their joint benefit

Fig. 18. Differences between predicted and observed confusions plotted as a function of percent confusion, for syllable subtest 1, in the quiet condition.


PLOT OF RESIDUAL*PCT_ERR SYMBOL USED IS X
RESIDUAL
15.0
12.5
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
0
2
4
6
8
10
12
14
16
18
20
PCT-ERR

to the listener is different from the sum of the individual contributions from each acoustic cue. This possibility is particularly intriguing in view of the observation by Rosenthal, et al. (1975) that the combination of low and high frequency acoustic cues are more effective in enhancing intelligibility than the summed contribution of the two sets of cues operating individually. In contrast, Walden and Montgomery (1975) have suggested that hearing-impaired listeners tend to rely on only one perceptually-important variable at a time. The mutual interactions between acoustic cues need to be considered.

In summary, the results of this study have identified some distinct patterns of consonant confusions among nonsense syllables. These patterns are similar but not identical to the results of other studies.

Differences between confusion patterns in the quiet and noise conditions indicate that the effects of noise may be profound and complex. Differences in relative frequencies of confusions over a range of speech levels, on the other hand, are not as large.

Confusion matrices can be predicted with a reasonable

degree of accuracy. However, variables found to be the best predictors change between experimental conditions and syllable subtest types. It is interesting to note that variables isolated as predictors of percent confusion for a particular subtest in quiet do not necessarily predict the confusions which will occur when the subtest is heard in noise.

Three variables (consonant energy, consonant spectral peaks, consonant-to-noise ratio) stand out as being particularly important, since they were isolated in a majority of the syllable subtests. This result suggests that these cues should, perhaps, be enhanced in speech communication and acoustic amplification systems. However, it should be remembered that the study reported here was a correlational study and that a causal study needs to be performed in which the acoustic cues are systematically removed and the predicted effects on the confusion matrix assessed. Further, in order to apply these data to the further development of acoustic amplification and auditory training systems for the hearing impaired, it is necessary to determine the differences between the hearing-impaired and normal-

hearing listeners in their utilization of available acoustic cues.

CHAPTER VI

SUMMARY

The purpose of this investigation was to predict, from acoustical measurements, the pattern of correct identifications and confusions of nonsense syllables in both quiet and noise conditions.

A set of speech stimuli was chosen which would allow for a detailed analysis of patterns of consonant confusions made by normal-hearing subjects under difficult listening conditions. A closed-response, nonsense syllable test, consisting of 11 interchangeable subtests, was chosen for use in the study. The subtests differ with respect to voicing of the consonant, consonant position, and vowel context. A sample of edited, cafeteria noise was used in conditions which required background noise.

In the listening experiment, two conditions were varied: level of the speech signal and speech-to-noise ratio. Speech levels ranged from 20 to 52 dB SPL. These speech levels were used in both the quiet and noise

(S/N = 5 dB) conditions. Subjects for the listening experiment were six normal-hearing adults who were familiar with the type of stimuli used in the experiment.

A set of acoustic characteristics of the speech stimuli were chosen for analysis, and careful measurement procedures were undertaken, including the use of computer-generated speech and noise spectra. The measurements included estimates of vowel formant frequency; second formant transition origin frequency, magnitude, and direction; consonant spectral peaks; overall consonant-noise bandwidth; total energy of consonants and vowels; and consonant, vowel, and closure durations.

A second set of measurements were generated which represented these characteristics as they appear in conjunction with the background noise. Overall average spectra of the noise was used as the speech-in-noise measurement device. A second set of measurements was collected for vowel peak frequency in noise, transition information in noise, consonant spectral peaks in noise, crossover frequency (an indirect measurement of bandwidth in noise), and consonant-to-noise and vowel-to-noise ratios.

The major findings of the listening experiment were as follows. Improvement in intelligibility for the 11 sets of nonsense syllables with increasing speech level was seen as a function of subtest type. Differences in articulation-gain functions were also noted for the two environmental conditions of quiet and noise. Subtests containing syllable-initial consonants and subtests containing consonants paired with the vowel /ɑ/ had consistently higher scores at all speech levels. Scores for voiceless subtests, however, were significantly lower than for voiced subtests at the lowest and highest speech levels only.

Confusion matrices were generated from the results of the listening experiment. As the error patterns in each matrix were found to be essentially the same as a function of speech level (with three minor exceptions), the matrices used in subsequent analyses were summed over that factor. Analyses of the percentages of correct identification for each matrix revealed affricates, glides and liquids to be most easily identified in both quiet and noise. Hierarchy of intelligibility for plosive,

fricative, and nasal consonants, however, differed as a function of both noise condition and consonant voicing. Scores for these phonemes in both noise conditions were higher when they appeared in the syllable-initial position. Level of percent correct identification for certain syllables was also shown to be affected by the response set in which it appears.

A summary of the major confusions revealed that the number of confusions was fairly evenly distributed between syllables containing voiced and voiceless consonants and among the three vowel contexts. A greater number of confusions, however, were found for consonants in the the syllable-final position. A relatively large number of confusions between consonants differing with respect to manner of articulation was found, primarily between fricative and plosive consonants.

Analysis of the pairs of syllables with the highest percentage of confusion showed somewhat different results, most notably between quiet and noise conditions, indicating the presence of complex interactions between error pattern and listening condition.

The results of the acoustic measurements made on the speech stimuli in both quiet and noise were correlated with percentage of correct identification under these conditions. A set of acoustic variables was isolated which was found to account for the largest proportion of the variance in predicting correct scores for each subtest, in each condition. Differences in acoustic variables between two syllables were used to predict the percentage of confusion between the two syllables. Another set of acoustic variables was isolated which accounted for most of the variance in that prediction, for each subtest, in each condition.

Inspection of the variables included in the prediction equations revealed that, for each subtest, the largest proportion of the variance (in quiet and in noise) was accounted for by different sets of variables. However, three variables stood out (consonant energy, consonant spectral peaks, and consonant-to-noise ratio) as being particularly important, since they were isolated in a majority of the subtests. This finding implies that the information provided by certain acoustic variables

may be necessary for reducing confusions between certain sets of syllables, but are of less value in other sets. In addition, on average, a greater level of accuracy was reached in predicting percent correct identifications than in predicting confusions. A trend in the residuals indicated that the predictions tended to underestimate the size of the largest confusion.

The mutual interactions between acoustic cues need to be considered in future analyses of this type. In addition, the performance of a causal study is indicated in which acoustic cues are systematically removed and the predicted effects on the confusions assessed. Finally, to apply these data to the further development of acoustic amplification and auditory training systems for the hearing impaired, it is necessary to determine the differences between the hearing-impaired and normal-hearing listeners in their utilization of available acoustic cues.

APPENDICES

APPENDIX A

VOWEL PEAK FREQUENCIES

Quiet Condition

Subtest	Syllable	VPF	Subtest	Syllable	VPF
1	ɑp	814	7	bɑ	1042
	ɑt	814		dɑ	1107
	ɑk	814		gɑ	423
	ɑf	1139		vɑ	1074
	ɑθ	814		ðɑ	423
	ɑs	814		zɑ	1107
2	up	293	8	pu	358
	ut	260		tu	326
	uk	293		ku	326
	uf	260		fu	326
	uθ	260		θu	326
	us	358		su	326
				hu	293
3	ip	195	9	pi	358
	it	293		ti	293
	ik	260		ki	326
	if	293		fi	326
	iθ	260		θi	326
	is	260		si	358
				hi	260
4	ɑb	1139	10	ib	228
	ɑd	1107		id	195
	ɑg	1042		ig	228
	ɑv	1172		iv	260
	ɑð	749		ið	293
	ɑz	1172		iz	293
	ɑm	1042		im	293
	ɑn	1009		in	260
	ɑŋ	1074		iŋ	293
5	pɑ	1074	11	ub	293
	tɑ	1107		ud	293
	kɑ	1107		ug	293
	fɑ	1074		uv	293
	θɑ	1139		uð	293
	sɑ	1107		uz	293
	hɑ	1107		um	227
				un	326
				uŋ	391

Note: All measures are reported in Hz.

Noise Condition

Subtest	Syllable	VPF	Subtest	Syllable	VPF
1	ɑp	814	7	bɑ	1042
	ɑt	1237		dɑ	1107
	ɑk	814		gɑ	423
	ɑf	1139		vɑ	1074
	ɑθ	814		ʒɑ	423
	ɑs	814		zɑ	1107
2	up	293	8	pu	358
	ut	260		tu	326
	uk	293		ku	326
	uf	260		fu	326
	uɵ	260		ɵu	326
	us	358		su	326
				hu	293
3	ip	260	9	pi	358
	it	2311		ti	293
	ik	2409		ki	326
	if	2344		fi	326
	iɵ	2409		θi	326
	is	2279		si	358
				hi	260
4	ɑb	1139	10	ib	228
	ɑd	1107		id	195
	ɑg	1042		ig	228
	ɑv	1172		iv	260
	ɑɵ	749		iʒ	293
	ɑz	1172		iz	293
	ɑm	1042		im	293
	ɑn	1009		in	260
	ɑŋ	1074		iŋ	293
5	pɑ	1074	11	ub	293
	tɑ	1107		ud	293
	kɑ	1107		ug	293
	fɑ	1074		uv	293
	ɵɑ	1139		uʒ	293
	sɑ	1107		uz	293
	hɑ	1107		um	227
				un	326
				uŋ	391

Note: All measures are reported in Hz.

APPENDIX B

ORIGIN FREQUENCY, MAGNITUDE, AND DIRECTION OF SECOND FORMANT TRANSITIONS

QUIET CONDITION

ST	SYL.	ORIG	MAG	DIR	ST	SYL.	ORIG	MAG	DIR
	ɑp	846	196	+		bɑ	1107	65	-
	ɑt	1302	98	+		dɑ	1562	390	-
1	ɑk	1334	195	+	7	gɑ	1497	325	-
	ɑf	1107	32	-		vɑ	1139	65	-
	ɑθ	1204	32	+		ðɑ	1270	131	-
	ɑs	1139	65	-		zɑ	1270	131	-
	up	977	0	0		pu	1335	358	-
	ut	977	131	+		tu	1530	163	-
2	uk	879	0	0		ku	1270	163	-
	uf	684	65	-	8	fu	977	33	-
	uθ	1270	489	+		θu	1302	228	-
	us	1009	195	+		su	1367	130	-
						hu	911	97	-
	ip	2344	97	-					
	it	2279	228	-		pi	2018	196	+
3	ik	2572	98	+		ti	2148	33	+
	if	2214	227	-		ki	2214	130	+
	iθ	2246	227	-	9	fi	2051	228	+
	is	2116	228	-		θi	1790	489	+
						si	1660	554	+
	ɑb	1074	33	-		hi	2474	163	-
	ɑd	1595	391	+					
	ɑg	1367	97	+		ib	2441	131	-
	ɑv	1042	97	-		id	2181	228	-
4	ɑð	1302	98	+		ig	2669	195	+
	ɑz	1367	163	+		iv	2409	32	-
	ɑm	1042	32	-	10	ið	2311	130	-
	ɑn	1562	162	+		iz	2083	228	-
	ɑŋ	1009	33	-		im	2409	65	-
						in	2539	65	-
	pɑ	1042	33	-		iŋ	2702	130	+
	tɑ	1204	32	-					
	kɑ	1009	130	+		ub	1432	0	0
5	fɑ	977	97	+		ud	1497	423	+
	θɑ	1139	33	-		ug	.	0	0
	sɑ	1172	65	-		uv	.	0	0
	hɑ	944	130	+	11	uð	1432	325	+
						uz	1367	358	+
						um	1107	33	+
						un	1432	260	+
						uŋ	1107	0	0

Note: ORIG and MAG are reported in Hz; +DIR = rising, -DIR = falling, 0 DIR = no change; Dot (.) = not measureable.

NOISE CONDITION

ST	SYL.	ORIG	MAG	DIR	ST	SYL.	ORIG	MAG	DIR
1	ɑp	1042	0	0	7	bɑ	1107	0	0
	ɑt	1237	33	+		dɑ	1562	390	-
	ɑk	1337	195	+		gɑ	1497	325	-
	ɑf	1107	32	-		vɑ	1074	0	0
	ɑθ	1204	32	+		ðɑ	1172	98	-
	ɑs	1139	65	-		zɑ	1172	98	-
2	up	.	0	0	8	pu	.	0	0
	ut	977	131	+		tu	1367	0	0
	uk	879	0	0		ku	1270	163	-
	uf	.	0	0		fu	977	33	-
	uɛ	1009	228	+		θu	1302	228	-
	us	.	0	0		su	1367	0	0
						hu	.	0	0
3	ip	2344	97	-	9	pi	2018	196	+
	it	2279	0	0		ti	2148	33	+
	ik	2572	98	+		ki	2214	130	+
	if	2214	227	-		fi	2051	228	+
	iθ	2246	196	-		θi	2018	261	+
	is	2116	228	-		si	2116	98	+
						hi	2311	0	0
4	ɑb	1074	33	-	10	ib	2571	0	0
	ɑd	1497	293	+		id	.	0	0
	ɑg	1367	97	+		ig	2669	195	+
	ɑv	1042	97	-		iv	2409	32	-
	ɑð	.	0	0		ið	2311	130	-
	ɑz	1270	66	+		iz	2279	32	-
	ɑm	1042	32	-		im	2409	65	-
	ɑn	1562	162	+		in	2539	65	-
	ɑŋ	1009	33	-		iŋ	2702	130	+
5	pɑ	1042	33	-	11	ub	.	0	0
	tɑ	1204	32	-		ud	.	0	0
	kɑ	1009	130	+		ug	.	0	0
	fɑ	1042	65	+		uv	.	0	0
	θɑ	1139	33	-		uð	.	0	0
	sɑ	1172	65	-		uz	.	0	0
	hɑ	944	130	+		um	.	0	0
						un	.	0	0
						uŋ	.	0	0

Note: ORIG and MAG are reported in Hz; +DIR = rising, -DIR = falling, 0 DIR = no change; Dot (.) = not measureable.

APPENDIX C

CONSONANT SPECTRAL PEAKS

QUIET CONDITION

ST	SYL.	CF1	CF2	ST	SYL.	CF1	CF2
1	ɑp	3353	.	7	bɑ	1107	.
	ɑt	3548	2279		dɑ	3255	4036
	ɑk	1465	3776		gɑ	1530	3646
	ɑf	3581	6706		vɑ	6934	7129
	ɑθ	7878	3743		ðɑ	6706	7910
	ɑs	3971	4427		zɑ	6836	7454
2	up	3255	1302	8	pu	391	1009
	ut	3581	6803		tu	6673	5794
	uk	1497	3581		ku	1400	.
	uf	7878	7910		fu	7910	6738
	uθ	3646	6803		θu	6673	6152
	us	3613	7650		su	6706	6087
					hu	1204	3555
3	ip	3418	6543	9	pi	1790	3483
	it	3743	5924		ti	6576	3678
	ik	1953	2279		ki	2669	3646
	if	8236	6868		fi	7194	6771
	iθ	6640	7682		θi	6966	3711
	is	3939	7552		si	6641	7747
					hi	2148	3516
4	ɑb	3418	1367	10	ib	1432	2506
	ɑd	3581	3482		id	2702	3743
	ɑg	2409	1465		ig	2409	3613
	ɑv	1335	3516		iv	2539	3353
	ɑð	6608	5599		iðˇ	6803	7943
	ɑz	6708	3874		iz	4167	6738
	ɑm	1172	4264		im	2441	1270
	ɑn	2702	1790		in	1628	4460
	ɑŋ	1074	2734		iŋ	2930	1172
5	pɑ	391	2376	11	ub	1237	2376
	tɑ	3678	4460		ud	2539	3353
	kɑ	1563	2181		ug	1302	2311
	fɑ	7843	6706		uv	6608	7747
	θɑ	7878	7422		uðˇ	6641	7943
	sɑ	4102	6803		uz	6510	7845
	hɑ	944	2474		um	2344	1204
					un	2441	1107
					uŋ	2376	1009

Note: All measures are reported in Hz; Dot (.) = not measureable.

NOISE CONDITION

ST	SYL.	CF1	CF2	ST	SYL.	CF1	CF2
1	ɑp	3353	.	7	bɑ	1107	.
	ɑt	3548	5762		dɑ	3255	4036
	ɑk	5892	3776		gɑ	1530	3646
	ɑf	3581	6706		vɑ	6706	7129
	ɑθ	7878	6706		ðɑ	6836	7910
	ɑs	3971	4427		zɑ	6836	7454
2	up	3255	6217	8	pu	391	1009
	ut	3581	6803		tu	6673	5794
	uk	1497	3581		ku	1400	.
	uf	7878	7910		fu	7910	6738
	uθ	3646	6803		θu	6673	6152
	us	3613	7650		su	6706	6087
					hu	1204	3555
3	ip	3418	6543	9	pi	1790	3483
	it	3743	5924		ti	6576	3678
	ik	1953	2279		ki	2669	3646
	if	8236	6868		fi	7194	6771
	iθ	6640	7682		θi	6966	3711
	is	3939	7552		si	6641	7747
					hi	2148	3516
4	ɑb	3418	6510	10	ib	1432	6673
	ɑd	3581	3482		id	2702	3743
	ɑg	3646	5664		ig	2409	3613
	ɑv	3516	6836		iv	2539	3353
	ɑð	6608	5599		ið	6803	7943
	ɑz	6702	3874		iz	4167	6738
	ɑm	1172	4264		im	2441	1270
	ɑn	2702	1790		in	1628	4460
	ɑŋ	1074	2734		iŋ	2930	2474
5	pɑ	2376	3483	11	ub	1237	6185
	tɑ	3678	4460		ud	2539	3353
	kɑ	1563	2181		ug	1302	2311
	fɑ	7843	6706		uv	6608	7747
	ɢɑ	7878	7422		uð	6641	7943
	sɑ	4102	6803		uz	6510	7845
	hɑ	944	3646		um	2344	1204
					un	2441	1107
					uŋ	2376	1074

Note: All measures are reported in Hz; Dot (.) = not measureable.

APPENDIX D

OVERALL CONSONANT-NOISE BANDWIDTHS

Subtest	Syllable	BW	Subtest	Syllable	BW
1	ɑp	866	7	bɑ	933
	ɑt	866		dɑ	1733
	ɑk	1056		gɑ	600
	ɑf	5267		vɑ	2934
	ɑθ	2734		ʒɑ	2734
	ɑs	3934		zɑ	4267
2	up	666	8	pu	467
	ut	1000		tu	2267
	uk	1200		ku	666
	uf	2800		fu	1934
	uθ	3067		θu	2600
	us	5067		su	2734
				hu	600
3	ip	867	9	pi	1133
	it	3867		ti	2333
	ik	867		ki	1867
	if	3934		fi	2800
	iθ	3467		θi	2667
	is	4600		si	3334
				hi	1134
4	ɑb	467	10	ib	467
	ɑd	1000		id	533
	ɑg	1533		ig	1334
	ɑv	933		iv	934
	ɑʒ	3000		iʒ	3200
	ɑz	4734		iz	4467
	ɑm	734		im	800
	ɑn	1666		in	534
	ɑŋ	2000		iŋ	1067
5	pɑ	800	11	ub	400
	tɑ	4400		ud	1267
	kɑ	1200		ug	600
	fɑ	2800		uv	1800
	θɑ	2734		uʒ	1866
	sɑ	4667		uz	2600
	hɑ	2000		um	1134
				un	1200
				uŋ	733

Note: All measures are reported in Hz.

APPENDIX E

CROSSOVER FREQUENCIES

Subtest	Syllable	XF
1	ɑp	2867
	ɑt	3000
	ɑk	3200
	ɑf	3133
	ɑθ	3333
	ɑs	3200
2	up	2733
	ut	2933
	uk	1400
	uf	3067
	uθ	3200
	us	3000
3	ip	2733
	it	3200
	ik	1933
	if	2933
	iθ	3467
	is	3333
4	ɑb	2800
	ɑd	2933
	ɑg	1400
	ɑv	2000
	ɑʒ	3200
	ɑz	3333
	ɑm	800
	ɑn	1300
	ɑŋ	667
5	pɑ	2000
	tɑ	3133
	kɑ	1400
	fɑ	2800
	θɑ	2867
	sɑ	3200
	hɑ	1733
7	bɑ	2200
	dɑ	2933
	gɑ	1267
	vɑ	3200
	ʒɑ	3467
	zɑ	3267
8	pu	2200
	tu	2267
	ku	1067
	fu	3267
	θu	3400
	su	3000
	hu	1133
9	pi	1600
	ti	3000
	ki	2133
	fi	2933
	θi	3333
	si	3200
	hi	2933
10	ib	1267
	id	2533
	ig	1600
	iv	3000
	iʒ	3400
	iz	3533
	im	1000
	in	1400
	iŋ	667
11	ub	2067
	ud	2600
	ug	1067
	uv	3067
	uʒ	3467
	uz	3067
	um	1000
	un	1067
	uŋ	867

Note: All measures are reported in Hz.

APPENDIX F

TOTAL ENERGY OF CONSONANTS AND VOWELS

ST	SYL.	CE	VE
1	ɑp	4.2	8.0
	ɑt	0.4	1.8
	ɑk	0.8	0.9
	ɑf	21.1	3.9
	ɑθ	6.0	4.4
	ɑs	20.8	4.9
2	up	5.0	4.6
	ut	10.0	4.4
	uk	3.0	5.4
	uf	7.8	2.7
	uθ	8.3	1.8
	us	25.7	4.7
3	ip	4.8	5.6
	it	15.3	1.8
	ik	13.0	3.2
	if	10.3	4.0
	iθ	9.8	0.0
	is	29.2	1.4
4	ɑb	13.4	6.9
	ɑd	13.4	5.3
	ɑg	14.8	6.5
	ɑv	17.7	6.2
	ɑɣ	14.4	6.3
	ɑz	21.6	6.6
	ɑm	19.3	4.3
	ɑn	23.4	6.7
	ɑŋ	22.4	7.7
5	pɑ	10.5	4.7
	tɑ	18.2	4.4
	kɑ	11.9	3.8
	fɑ	3.5	5.6
	θɑ	11.2	5.1
	sɑ	30.5	6.3
	hɑ	13.6	7.0

ST	SYL.	CE	VE
7	bɑ	9.3	7.8
	dɑ	9.3	4.1
	gɑ	12.0	7.0
	vɑ	9.3	5.9
	ðɑ	14.2	4.9
	zɑ	20.8	5.2
8	pu	20.3	1.9
	tu	13.6	0.0
	ku	11.0	1.6
	fu	2.0	0.9
	θu	9.3	2.4
	su	26.1	2.3
	hu	0.4	1.0
9	pi	7.5	3.6
	ti	14.8	2.7
	ki	12.8	5.0
	fi	4.3	3.3
	θi	9.1	3.0
	si	29.2	3.9
	hi	0.0	3.1
10	ib	15.0	5.2
	id	15.0	3.5
	ig	19.1	8.2
	iv	19.6	4.4
	iɣ	20.2	9.5
	iz	22.6	3.7
	im	25.0	7.6
	in	23.6	4.3
	iŋ	25.5	9.1
11	ub	19.8	6.1
	ud	16.8	5.3
	ug	17.3	7.5
	uv	25.2	5.1
	uð	15.1	4.8
	uz	25.4	4.3
	um	25.5	5.0
	un	26.4	6.5
	uŋ	27.0	6.8

Note: Measures for CE are reported in dB re: /hi/;
Measures for VE are reported in dB re: /iθ/.

APPENDIX G

CONSONANT-TO-NOISE AND VOWEL-TO-NOISE RATIOS

ST	Syl.	C/N	V/N
1	ap	-18.8	10.1
	at	-20.6	4.3
	ak	-21.8	4.2
	af	- 2.4	15.2
	aθ	-17.8	5.4
	as	- 3.2	5.9
2	up	-17.7	4.3
	ut	-16.4	3.0
	uk	-21.2	5.9
	uf	-14.0	7.8
	uθ	-13.0	6.5
	us	2.1	5.2
3	ip	-19.2	7.5
	it	- 8.4	3.9
	ik	- 9.0	2.1
	if	-12.8	5.8
	iθ	-12.8	0.5
	is	4.2	4.6
4	ab	- 9.6	12.6
	ad	- 8.0	6.5
	ag	-10.2	6.5
	av	- 6.3	7.4
	aʒ	- 6.5	7.6
	az	- 2.1	5.4
	am	- 3.6	5.9
	an	- 0.8	6.5
	aʃ	- 5.0	5.6
5	pa	- 2.0	4.2
	ta	- 4.5	4.2
	ka	-12.5	2.0
	fa	-19.1	3.2
	θa	-12.8	3.4
	sa	5.8	2.4
	ha	- 8.7	6.6
7	ba	-15.4	7.1
	da	-12.2	2.3
	ga	- 9.5	6.8
	va	-13.8	5.8
	ʒa	- 9.1	3.1
	za	- 1.5	2.5
8	pu	0.2	1.8
	tu	- 4.5	0.1
	ku	- 6.5	2.0
	fu	-19.0	0.6
	θu	-12.8	1.0
	su	5.4	0.5
	hu	-17.7	-2.0
9	pi	- 9.8	4.5
	ti	- 1.3	2.3
	ki	- 4.7	4.8
	fi	-13.0	0.4
	θi	-12.4	1.6
	si	9.9	4.1
	hi	-11.2	5.6
10	ib	- 5.7	6.3
	id	- 6.1	4.3
	ig	- 5.7	6.1
	iv	- 3.7	4.8
	iʒ	- 4.8	9.4
	iz	- 0.8	3.4
	im	0.9	6.7
	in	- 3.0	3.2
	iʃ	1.4	7.9
11	ub	1.0	8.0
	ud	- 6.3	8.0
	ug	- 6.2	8.3
	uv	- 1.2	7.5
	uʒ	- 7.2	6.3
	uz	2.0	5.2
	um	2.0	7.0
	un	5.6	3.7
	uʃ	2.0	8.1

Note: All measures are reported in dB.

APPENDIX H

DURATIONS

ST	Syl.	CD	VD	CLD
1	ɑp	78	146	115
	ɑt	38	153	141
	ɑk	89	163	129
	ɑf	218	180	.
	ɑθ	243	192	.
	ɑs	256	211	.
2	up	102	133	145
	ut	64	166	139
	uk	73	142	137
	uf	230	179	.
	uθ	230	192	.
	us	294	166	.
3	ip	64	161	137
	it	77	179	129
	ik	89	141	131
	if	269	180	.
	iθ	269	180	.
	is	333	179	.
4	ɑb	26	230	139
	ɑd	58	230	116
	ɑg	42	256	106
	ɑv	218	256	.
	ɑʒ	179	294	.
	ɑz	235	284	.
	ɑm	197	230	.
	ɑn	200	224	.
	ɑŋ	320	192	.
5	pɑ	89	349	.
	tɑ	99	330	.
	kɑ	129	323	.
	fɑ	209	346	.
	θɑ	169	337	.
	sɑ	234	349	.
	hɑ	161	332	.

ST	Syl.	CD	VD	CLD
7	bɑ	13	307	.
	dɑ	20	372	.
	gɑ	29	350	.
	vɑ	166	333	.
	ʒɑ	166	311	.
	zɑ	207	319	.
8	pu	64	269	.
	tu	95	248	.
	ku	93	259	.
	fu	119	247	.
	θu	195	269	.
	su	170	247	.
	hu	80	234	.
9	pi	67	234	.
	ti	70	307	.
	ki	113	290	.
	fi	93	208	.
	θi	157	247	.
	si	123	303	.
	hi	55	285	.
10	ib	52	285	106
	id	70	316	96
	ig	71	307	76
	iv	179	287	.
	iʒ	205	265	.
	iz	218	339	.
	im	237	301	.
	in	210	300	.
	iŋ	291	237	.
11	ub	62	268	95
	ud	64	320	109
	ug	76	294	84
	uv	144	285	.
	uʒ	171	344	.
	uz	214	332	.
	um	270	228	.
	un	244	268	.
	uŋ	374	195	.

Note: All measures are reported in msec; Dot (.) = not measureable.

REFERENCES

Abbs, M.S. and Minifie, F.D. Effect of acoustic cues in fricatives on perceptual confusions in preschool children. J. Acoust. Soc. Amer., 46:1535-1542, 1969.

Allen, G.D. Vowel duration measurement: A reliability study. J. Acoust. Soc. Amer., 63:1176-1185, 1978.

American National Standards Institute, American National Standards Specifications for Audiometers. ANSI S3.6-1969, New York, 1970.

Atal, B.S. and Schroeder, M.R. Predictive coding of speech signals. Proc. 1967 Conf. Commun. and Process., 360-361, 1967.

Barr, A.J., Goodnight, J.H., Sall, J.P. and Helwig, J.T. A User's Guide to SAS.76. Raleigh: SAS Institute, Inc., 1976.

Bergland, G.D. A guided tour of the fast Fourier transform. IEEE Spectrum, 6:41-52, 1969.

Bilger, R.C. and Wang, M.D. Consonant confusions in patients with sensorineural hearing loss. J. Speech Hearing Res., 19:718-740, 1976.

Blackman, R.B. and Tukey, J.W. The Measurement of Power Spectra. New York: Dover, 1959.

Brownlee, K.A. Statistical Theory and Methodology in Science and Engineering. New York: John Wiley and Sons, 1965.

Busch, A.C. and Eldredge, D. The effect of differing noise spectra on the consistency of identification of consonants. Lang. and Speech, 10:194-202, 1967.

Campbell, G.A. Telephonic intelligibility. Phil. Mag., 19:152-159, 1910.

Cole, R.A. and Cooper, W.E. Perception of voicing in English affricates and fricatives. J. Acoust. Soc. Amer., 58:1280-1287, 1975.

Cole, R.A. and Scott, B. Toward a theory of speech perception. Psych. Rev., 81:348-374, 1974.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer., 24:597-606, 1952.

Danaher, E.M., Osberger, M.J. and Pickett, J.M. Discrimination of formant frequency transitions in synthetic vowels. J. Speech Hearing Res., 16:439-451, 1973.

Danaher, E.M. and Pickett, J.M. Some masking effects produced by low-frequency vowel formants in persons with sensorineural hearing loss. J. Speech Hearing Res., 18:261-271, 1975.

Delattre, P.C., Liberman, A.M. and Cooper, F.S. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Amer., 27:769-773, 1955.

Denes, P. Effect of duration on the perception of voicing. J. Acoust. Soc. Amer., 27:761-764, 1955.

Egan, J.P. Articulation testing methods. Laryngoscope, 58:955-991, 1948.

Fant, G.C.M. Acoustic Theory of Speech Production. The Hague: Mouton and Co., 1960.

Fertig, R.H. Temporal interrelations in selected English CVC utterances. Speech Communications Research Laboratory. SCRL Monograph No. 12, 1976.

Fletcher, H. Speech and Hearing in Communication. Princeton: D. Van Nostrand Co., 1953.

Fletcher, H. and Steinberg, J.C. Articulation testing methods. Bell Syst. Tech. J., 8:806-854, 1929.

French, N.R. and Steinberg, J.C. Factors concerning the intelligibility of speech sounds. J. Acoust. Soc. Amer., 19:90-119, 1947.

Fujimura, O. Analysis of nasal consonants. J. Acoust. Soc. Amer., 34:1865-1875, 1962.

Gabor, D. Theory of communication. J. Inst. Elect. Engrs., 93:429, 1946.

Halle, M., Hughes, G.W. and Radley, J.P. Acoustic properties of stop consonants. J. Acoust. Soc. Amer., 29:107-116, 1957.

Harris, K.S. Cues for the discrimination of American English fricatives in spoken syllables. Lang. and Speech, 1:1-7, 1958.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C. and Cooper, F.S. Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust. Soc. Amer., 30:122-126, 1958.

Heinz, J.M. and Stevens, K.N. On the properties of voiceless fricative consonants. J. Acoust. Soc. Amer., 33: 589-596, 1961.

Hirsh, I.J., Reynolds, E.G. and Joseph, M. Intelligibility of different speech materials. J. Acoust. Soc. Amer., 26:530-538, 1954.

Horii, Y., House, A.S. and Hughes, G.W. A masking noise with speech envelop characteristics for studying intelligibility. J. Acoust. Soc. Amer., 49:1849-1856, 1971.

House, A.S. and Fairbanks, G. The influence of consonant environment upon the secondary acoustic characteristics of vowels. J. Acoust. Soc. Amer., 25:105-113, 1953.

House, A.S., Williams, C.E., Hecker, M.H.L. and Kryter, K.D. Articulation testing methods: Consonantal differentiation with a closed response set. J. Acoust. Soc. Amer., 37:158-166, 1965.

Hughes, G.W. and Halle, M. Spectral properties of fricative consonants. J. Acoust. Soc. Amer., 28:303-310, 1956.

Kryter, K.D. Methods for the calculation and use of the articulation index. J. Acoust. Soc. Amer., 34:1689-1697, 1962a.

Kryter, K.D. Validation of the articulation index. J. Acoust. Soc. Amer., 34:1698-1702, 1962b.

Lawrence, D.L. and Byers, V.W. Identification of voiceless fricatives by high frequency hearing-impaired listeners. J. Speech Hearing Res., 12:426-434, 1969.

Levitt, H. Analysis of acoustic characteristics of deaf speech. Ann. Rep., NIH Grant NS09252, 1971.

Levitt, H., Collins, M.J., Dubno, J., Resnick, S.B., and White, R.E.C. Development of a protocol for the prescriptive fitting of a wearable master hearing aid. Communication Sciences Laboratory Report #11, City University of New York, 1978.

Liberman, A.M., Delattre, P.C., Cooper, F.S. and Gerstman, L.J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monog., 68:1-13, 1954.

Liberman, A.M., Delattre, P.C. and Cooper, F.S. Some cues for the distinction between voiced and voiceless stops in the initial position. Lang and Speech, 1:153-167, 1958.

Lindblom, B. Spectrographic study of vowel reduction. J. Acoust. Soc. Amer., 35:1773-1781, 1963.

Lisker, L. and Abramson, A. A cross-language study of voicing in initial stops: Acoustic measurements. Word, 20: 384-422, 1964.

Malecot, A. Acoustic cues for nasal consonants. Language, 32:274-284, 1956.

Markel, J.D. Formant trajectory estimation from a linear least-squares inverse filter formulation. SCRL Monograph No. 7, Speech Communications Research Laboratory, Santa Barbara, Calif., 1971.

Markel, J.D. Digital inverse filtering - A new tool for formant trajectory estimation. IEEE Audio. Electroacoust. AU-20, 129-137, 1972

Markel, J.D. and Gray, A.H. Linear Prediction of Speech. Berlin: Springer-Verlag, 1976.

Miller, G.A. The masking of speech. Psychol. Bull., 44: 105-129, 1947.

Miller, G.A. The perception of speech. In, For Roman Jakobson, M.Halle (Ed.), The Hague: Mouton and Co., 1956.

Miller, G.A. and Nicely, P.A. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer., 27:338-352, 1955.

Nábělek, I., and Hirsh, I.J. On the discrimination of frequency transitions. J. Acoust. Soc. Amer., 45:1510-1518, 1969.

Owens, E., Benedict, M. and Schubert, E.D. Consonant phonemic errors associated with pure tone configurations and certain kinds of hearing impairment. J. Speech Hearing Res., 15:308-322, 1972.

Owens, E., Talbott, C.B and Schubert, E.D. Vowel discrimination of hearing impaired listeners. J. Speech Hearing Res., 11:648-655, 1968.

Oyer, H.J. and Doudna, M. Structural analysis of word responses made by hard of hearing subjects on a discrimination test. Arch. Otolaryng., 70: 357-364, 1959.

Pederson, O.T. and Studebaker, G.A. A new minimal-contrasts closed-response set speech test. J. Aud. Res., 12:187-195, 1972.

Peterson, G.E. and Barney, H.L. Control methods used in a study of the vowels. J. Acoust. Soc. Amer., 24:175-184, 1952.

Peterson, G.E. and Lehiste, I. Duration of syllabic nuclei in English. J. Acoust. Soc. Amer., 32:693-703, 1960.

Pickett, J.M. Perception of vowels heard in noises of various spectra. J. Acoust. Soc. Amer., 29:613-620, 1957.

Pickett, J.M., Martin, E.S., Johnson, D., Smith, S.B., Daniel, Z., Willis, D. and Otis, W. On patterns of speech feature reception by deaf listeners. In, Fant, G. (Ed.) Proc Int. Symp. on Speech Communication and Profound Deafness. Washington, D.C.:A.G. Bell Assoc., 1970.

Pickett, J.M. and Martony, J. Low-frequency vowel formant discrimination in hearing impaired listeners. J. Speech Hearing Res., 13:347-359, 1970.

Pickett, J.M. and Rubenstein, H. Perception of consonant voicing in noise. Lang and Speech, 3:155-163, 1960.

Pollack, I. Effects of high-pass and low-pass filtering on the intelligibility of speech in noise. J. Acoust. Soc. Amer., 20: 259-266, 1948.

Pollack, I. and Pickett, J.M. Effect of high sound levels on the intelligibility of speech in noise. J. Acoust. Soc. Amer., 29:781 (A), 1957a.

Pollack, I. and Pickett, J.M. Effect of noise and filtering on speech intelligibility at high levels. J. Acoust. Soc. Amer., 29:1328-1329, 1957b.

Pollack, I. and Pickett, J.M. Masking of speech by noise at high sound levels. J. Acoust. Soc. Amer., 30:127-130, 1958.

Pollack, I., Rubenstein, H. and Decker, L. Intelligibility of known and unknown message sets. J. Acoust. Soc. Amer., 31:273-279, 1959.

Raphael, L.J. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Amer., 51: 1296-1303, 1972.

Reed, C. Identification and discrimination of vowel-consonant syllables in listeners with sensori-neural hearing loss. J. Speech Hearing Res., 18:773-744, 1975.

Resnick, S.B., Dubno, J.R., Hoffnung, S and Levitt, H. Phoneme errors on a nonsense syllable test. J. Acoust. Soc. Amer., 58:114 (A), 1975.

Rosenthal, R.D., Lang, J.K. and Levitt, H. Speech reception with low frequency speech energy. J. Acoust. Soc. Amer., 57:949-955, 1975.

Saito, S. and Itakura, F. The theoretical consideration of statistically optimum methods for speech spectral density. Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo, 1966.

Sher, A.E. and Owens, E. Consonant confusions associated with hearing loss above 2000 Hz. J. Speech Hearing Res., 17:669-681, 1974.

Stevens, K.N. and House, A.S. Speech perception. In, Foundations of Modern Auditory Theory, Vol II, J. Tobias (Ed.), New York: Academic Press, 1972.

Stevens, K.N. and Klatt, D.H. Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer., 55:653-659, 1974.

Strevens, P. Spectra of fricative noise in human speech. Lang. and Speech, 3:32-49, 1960.

Studdert-Kennedy, M. The perception of speech. In, Current Trends in Linguistics, Vol XII, Sebeok, T. (Ed.), The Hague: Mouton and Co., 1975.

Walden, B.E. and Montgomery, A.A. Dimensions of consonant perception in normal and hearing impaired listeners. J. Speech Hearing Res., 18:444-455, 1975.

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

Winitz, H., Scheib, M.E. and Reeds, J.A. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. J. Acoust. Soc. Amer., 51:1309-1317, 1972.