ACOUSTIC CHARACTERISTICS AND INTELLIGIBILITY OF CLEAR AND CONVERSATIONAL SPEECH AT THE SEGMENTAL LEVEL

by

Francine Robina Chen

B.S.E., University of Michigan
(1978)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREES OF

MASTER OF SCIENCE

(and)

(ELECTRICAL ENGINEER)

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1980

Signature redacted

Signature of Author .....
Department of Electrical Engineering and Computer Science
May 23, 1980

Signature redacted

Certified by .....
Thesis Supervisor

Signature redacted

Accepted by .....
Chairman, Departmental Committee on Graduate Students

ACOUSTIC CHARACTERISTICS AND INTELLIGIBILITY

OF CLEAR AND CONVERSATIONAL SPEECH

AT THE SEGMENTAL LEVEL

by

FRANCINE ROBINA CHEN

Submitted to the Department of Electrical Engineering and Computer Science on May 23, 1980 in partial fulfillment of the requirements for the degrees of Master of Science (and Electrical Engineer.)

ABSTRACT

This thesis was directed towards understanding improvements in intelligibility that result from attempting to speak clearly in terms of changes in segmental acoustic properties. The corpus consisted of 18 CV syllables (six stop consonants and three vowels) embedded in a carrier phrase; each phrase was spoken several times by three male speakers in a "conversational" and "clear" manner. The CV's were excised from the carrier phrase and presented, with various levels of masking noise, to four listeners in identification intelligibility tests. Acoustic parameters of the CV's, such as formant frequencies, VOT, CV ratios, and burst frequencies were measured. Intelligibility confusions were analyzed in terms of articulatory features. In some cases, good correlation between articulatory features and the measured acoustic parameters was observed.

Thesis Supervisor: Victor W. Zue

Title: Lecturer and Research Associate in Electrical Engineering

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my thesis advisor, Dr. Victor Zue, for his constant guidance, encouragement, and generous sharing of ideas and knowledge. I am extremely grateful not only for the time and effort he spent in supervision, but also for the many hours which he spent phonetically labeling the speech waveforms. Deep appreciation goes to Nat Durlach, Lou Braida, and Michael Picheny for donating their time to discuss issues of this thesis with me; to Pat Peterson for the use of his software for intelligibilty testing; and to Mark Davis for giving his time and help in technical matters. To all the unmentioned members of the Communications Biophysics Laboratory who served as subjects when conducting pilot studies and who offered many useful suggestions, I acknowledge with much appreciation.

I would especially like to thank John Makhoul and Richard Schwartz for making available the facilities at BBN for analyzing acoustic parameters; without the use of such facilities, this thesis would have been almost impossible.

Finally, I would like to thank my parents, Robin and June Chen, for their continuing love, encouragement, and support.

TABLE OF CONTENTS

FIGURES AND TABLES

1. INTRODUCTION

In everyday, conversational speech, people speak easily and freely, without paying much attention to their enunciation. However, under various conditions, people tend to speak with more care in an attempt to make their speech "clearer". One condition in which this might occur is when people cannot hear well, such as when talking in a noisy subway or when speaking to someone who has impaired hearing. Another condition in which people try to speak more clearly is when one is talking to someone with low linguistic competence, such as when speaking to foreigners or when speaking to small children.

When attempting to speak clearly, different people will employ different techniques to try to make themselves more intelligible. The techniques used range from speaking more loudly, to speaking more slowly, to changing other characteristics of speech or using a combination of techniques. Not all of these techniques are equally effective. The subject of this study are the changes in acoustic parameters from techniques (other than an increase in level) which are most effective in making speech more intelligible.

In this thesis, clear speech is defined to be the style of speech which results from people attempting, with the help of feedback on their intelligibility, to make their speech more intelligible. The paradigm which was employed to accomplish this is described in Section 2.2. The term conversational speech will be used to denote speech which is spoken in a style similar to conversational speech.

The goal of this research is to determine improvements in intelligibility that result from attempting to speak clearly and to understand these improvements in terms of changes in the acoustic parameters of speech. Such understanding will contribute not only to basic speech science, but also, hopefully, to the development of improved hearing aids, speech training aids, speech synthesizers and speech recognition systems. That is, speech synthesizers which are more intelligible may be produced by utilizing knowledge of acoustic characteristics important to intelligibility and how these characteristics are important. Information on the acoustical characteristics of clear speech may be similarly applied to improve the intelligibility of speech received by an impaired listener. And finally, this knowledge may also be used to help train people with impaired hearing to speak more clearly by providing feedback on acoustical parameters important to speech intelligibility.

1.1 PREVIOUS RESEARCH ON CLEAR SPEECH

Literature aimed at determining the acoustical characteristics governing speech clarity is relatively sparse and sometimes contains conflicting results. In several studies, acoustic characteristics which might be responsible for clear speech were examined. However, in many of these studies, only one or two parameters related to interspeaker, not intraspeaker, differences were investigated. Also, the subject of interest was word intelligibility in sentences, rather than acoustic characteristics at the segmental level.

Tolhurst (1955) and Fairbanks et. al. (1957) investigated intraspeaker changes which may be correlated with more intelligible speech. Tolhurst (1955) found that the instruction to speak more intelligibly resulted in increased intelligibility scores, indicating that intraspeaker changes in speech may affect intelligibility. Fairbanks et. al. (1957) used a VU meter to control the vocal efforts of their speakers. Their results showed a correlation of large consonant-vowel (CV) ratios (the difference in dB of the magnitude of the rms energy in a consonant to that in an adjoining vowel) with more intelligible speech. They also found that increased loudness when speaking resulted in decreased CV ratios.

Tolhurst (1957) conducted further experiments which showed that there was a slight (86% vs. 84% intellibility) numerical advantage in the intelligibility of "prolonged" speech over "normal" speech and of both of these styles over "staccato" speech (77% intelligibility). On the other hand, House et. al. (1965) studied interspeaker differences and suggested that a greater vowel length resulted in less intelligible speech. He also suggested that a greater CV ratio resulted in more intelligible speech.

Salmon (1970) examined interspeaker differences between the four most and four least intelligible speakers selected from 20 male speakers. He suggested that durational differences did not have a significant effect in intelligibility tests, in conflict with the results of both Tolhurst (1957) and House (1965). He also found that a small CV amplitude difference (larger CV ratio) resulted in increased intelligibility in accordance with previous work. More recently, Hecker (1974) investigated the relationship between CV ratios and intelligibility by manually varying the amplitude of the consonant segments and performing intelligibility tests on the modified and original CV's. His results of a large CV ratio being correlated with increased intelligibility are in agreement with Salmon (1970) and House et. al. (1965).

In a study concerned with diagnostic articulation testing, Griffiths (1967) conducted an experiment in which speakers exaggerated the differences between two pairs of sounds. Identification tests were run, but the results were inconclusive as to whether speech is made clearer by exaggeration.

In 1979, Picheny and Durlach conducted a study which is more closely related to this thesis work than the works which were previously examined. The results indicated promising trends in intelligibility and in acoustic characteristics. They found that the intelligibility of one speaker increased an average of 18 percentage points when speaking Harvard sentences clearly as opposed to conversationally. Intelligibility tests were conducted on four listeners with sensorineural hearing loss. The listeners were allowed to select their own presentation levels in these tests. Measurements on the conversational and clear speech waveforms showed that in clear speech, durations of speech elements increased, CV ratios for plosives increased, and first and second vowel formant (F1 and F2) target values were approached more closely.

A pilot study (Chen, 1979) investigating trends in acoustical characteristics of stop consonants when speaking clearly, conversationally and quickly was conducted. The

corpus consisted of 18 CV combinations embedded in a carrier phrase. One male and one female speaker read each of the carrier phrases containing the CV's several times in each speaking style. One token of each CV in each style of speech was acoustically analyzed. Each subject spoke clearly in the manner which they interpreted as speaking to make oneself more intelligible. Intelligibility tests were not run in the probe experiment to verify that the clear style of speech was more intelligible.

Analysis of the acoustic characteristics of the recorded CV's indicated that in clear speech, when compared to normal and fast speech, formant targets of the vowels were more closely approached, CV ratios showed a tendency to increase, and durations of all speech segments increased, but in a nonuniform manner. Durations measured included that of the vowel, the burst and aspiration of the consonant, and the silence period preceeding the consonant. Although formant transition rates were measured, the effect on them was uncertain due to the rapid transition rate characteristic of stops and precision limitations of the formant tracker.

1.2 OVERVIEW OF THE PRESENT INVESTIGATION

A long term research goal is to determine for every phoneme the segmental intraspeaker parameters which contribute significantly to its intelligibility and to the intelligibility of speech in general. This large task set forth by the overall research goal has been reduced in this thesis by considering a subset of speech sounds in a controlled environment. In this study, many acoustic properties of speech which may be important to the intelligibility of the chosen subset of speech sounds were investigated more carefully than in previous works.

This research consisted of two complementary and distinct parts: investigation of acoustic properties distinguishing a subset of clear and conversational speech, and determination and comparison of the intelligibility of the "clear" speech versus the intelligibility of the "conversational" speech. It should be noted that what is called conversational speech is actually speech read in a style which approximates conversational speech, and that what is called clear speech is actually speech read while using the paradigm described in Section 2.2 for eliciting speech which is hopefully more intelligible. In both styles of speech, the environment of the CV was similar. In each case, the CV was embedded in the same carrier phrase to form

a nonsense sentence. Because of this and because the same corpus was utilized in both parts, comparisons between the results of the two parts could be performed.

In particular, the work included developing methods and techniques for eliciting the two different styles of speech and examining a more comprehensive selection of acoustical characteristics of clear speech than in previous investigations. Also, rather than investigate word intelligibility in sentences, as done previously, segmental characteristics were examined through the use of stop consonant phonemes followed by a vowel (a CV).

Identification type intelligibility tests of CV's pronounced conversationally and clearly were performed. The intelligibility results were grouped and analyzed to identify voicing errors, place errors, effect of speaker, effect of listener, and effect of signal-to-noise ratio (SNR). Specific acoustic characteristics were selected for examination because their contribution to the intelligibility of speech was thought to be significant. For example, a longer VOT in the voiceless stops but not in voiced stops could be hypothesized to be significant in voiced/voiceless distinctions. Correlations analogous to this example were hypothesized and discussed.

2. DATA ACQUISITION

2.1 DATA BASE

There are many classes of phonemes: vowels, stop consonants, fricatives, and glides, to name a few. Each class of phonemes has its own characteristics, and within each class, each phoneme has its own acoustic characteristics. The ultimate goal of this research is to understand the segmental acoustic characteristics which make each phoneme intelligible. However, it is an enormous task to analyze all segmental features in all phonemes. To reduce the size of this task, a subset of speech sounds, CV syllables, were chosen for this first study of clear speech at the segmental level. In particular, one class of consonants was chosen for analysis since consonants are often confused with one another, especially within the same class (Miller and Nicely, 1950). The selected data base was comprised of stressed consonant-vowel (CV) syllables spoken by three male speakers. The CV's were formed from combinations of one of the six stop consonants (/p/, /t/, /k/, /b/, /d/, or /g/) followed by one of the three point vowels (/i/ (high, front), /a/ (low, back), or /u/ (high, back)). Stop consonants were selected for examination because they exhibit many dynamic acoustic characteristics, resulting in increased efficiency of the number of acoustic

measurements made per token. Since a consonant is influenced by its vowel environment, the following vowel in a stressed CV should be carefully determined. In this investigation, the three point vowels were chosen for study because they represent the extremes in combinations of the first and second vowel formants (as seen in the F1 versus F2 plane).

For reasons related to acoustic anaylsis, stressed consonants were chosen for study. They are usually articulated with greater care and effort (Zue, 1976), resulting in a more robust signal. This allows easier and more reliable determination of acoustic parameters. Employing only male speakers eliminated an additional parameter involving the differences between male and female speakers (Fairbanks et al., 1957) and simplifies extraction of signal parameters since male voices have lower pitch and are less noisy (breathy). Five conversational and five clear tokens of each of the 18 CV's from each of the three male speakers were analyzed, resulting in a total of 540 tokens.

Initially, five native American, male, test speakers were recorded. They were instructed to speak test sentences conversationally, clearly, and clearly under an interactive paradigm (described in the next section) with the

experimenter serving as the listener. From the five test speakers, three speakers were selected on the basis of their pronunciation, the ease with which they learned the task, and the seriousness with which they worked.

2.2 SPEECH ELICITATION TECHNIQUES

An important part in the elicitation of speech is the carrier in which the CV is embedded. The carrier affects the ease with which one may attempt to say something conversationally or clearly. It also influences the local and global environment of the CV. Therefore, the same carrier was used when speaking clearly and conversationally to minimize differences in local or global environments; these differences, if present, could affect the intelligibility and acoustic characteristics of the phonemes (Kreul et al., 1969). Various types of carrier phrases were examined for use when eliciting clear and conversational speech. From the results of the investigation, it was decided to use the sentence "Say /hə'CVp/ again." as the carrier for both styles of speech. The use of a sentence carrier was a compromise between a paragraph carrier (which encouraged a conversational style of speech but caused difficulty when trying to speak clearly), and the isolated CV's (which prompted a clear style of speech while detering a conversational style of speech). The /hə'CVp/ context was

chosen to provide a uniform environment for the CV and to allow easier excision of the string from the carrier.

Conversational speech was elicited through the use of instructions and by having the speaker practice reading 300 sentences which contained all 18 CV's comprising the corpus. The speakers were instructed to read the sentences as they would normally talk. After practices, ten conversational style tokens of each CV combination embedded in the carrier sentence were recorded by each speaker. The sentences were spoken in groups of five, with the speaker pausing for a breath between each sentence group.

As mentioned in the introduction, different people will react differently to the explicit instruction to speak clearly. They employ different techniques which they believe would make themselves more intelligble. However, there is no assurance that the speech from someone attempting to speak more clearly has a higher intelligiblity score than their speech when speaking conversationally. Therefore, it was desirable to use an objective method which would give some assurance that the speech was in fact intelligible. In addition, the method should encourage speakers to try techniques other than increasing their volume to make themselves sound clearer. Several methods attempting to acheive this were examined. (Appendix 1

describes methods tried but not outlined below). From these methods, an interactive paradigm was chosen for eliciting the clear speech. This paradigm, diagrammed in Figure 2.1 employed a normal hearing listener who served as feedback to the speaker on the intelligibility of his speech. The same listener was employed for all three speakers. The listener had masking noise and the speaker's speech applied binaurally. Each time the speaker read a sentence, the listener, who had been instructed not to outguess the speaker, responded with the embedded CV which he heard. The speaker repeated the sentence until the listener perceived the correct CV. The listener was not given feedback on whether or not he replied with the correct CV. An automatic volume control (AVC) circuit was used to normalize the speaker's volume, thereby detering the speaker from becoming more intelligible by speaking louder (see Appendix 2 for details on the AVC). Fairbanks et al. (1957) had used a VU meter to control their speaker's volume, but this proved to be a difficult and distracting method. The AVC allowed the speaker to concentrate on speaking clearly without worrying about volume. The listener binaurally heard 80 dB of masking noise plus the speech at a SNR set for each speaker. The SNR level was determined by the lowest level used in the practice session where the listener perceived about one third of the CV's correctly the first time, and also by running through 20 test sentences at the beginning of the

session to check that the listener wasn't perceiving most of the CV's correctly. The speaker read the sentences from a list which consisted of the CV's embedded in the same carrier sentence as in the conversational style. The CV's were randomized such that they each appeared six times sentences on the list.

Practice sessions prior to recording were employed. In one session, the speaker was prompted by the experimenter on how he could make his speech sound more clear to the experimenter. A practice session using the interactive paragigm with the listener was then held. A separate list of sentences containing each CV three times was used. Recording of the clear speech followed in the next session.

2.3 SPEECH ACQUISITION

2.3.1 Recording Sessions

Five sessions were held with each speaker, but actual recording of speech material occurred only in two of the five sessions. In the first session, the speakers practiced saying the sentences in a manner that approximates conversational speech and became acquainted with the recording environment. In the second session, the speakers were recorded speaking the sentences in a manner similar to

conversational speech. Then, in the third session, the speaker was instructed by the experimenter on how to speak more clearly. A fourth session was held in which the subject practiced with the listener under the interactive paradigm. Finally, in the fifth session, the clear speech was recorded in a session in which the interactive paradigm was utilized. After recording the speech, it was processed for use in intelligibility tests and acoustic measurements. The processing involved digitization, normalization, and smoothing of the ends of the waveforms.

The actual recording sessions were broken into smaller subsessions to reduce fatigue on the part of the speakers and the listener. In the conversational style, two subsessions were employed. In the clear style, two major sessions with five one minute breaks between each of six subsession were used. The speech for each speaker was recorded in twe sessions of about one and one half hours each. Two separate days were used because the effect of fatigue on the speakers was thought to be greater than differences due to speaking on different days.

2.3.2 Recording Setup

Recordings were made in an anechoic chamber. The speaker was seated in the center of the chamber on a chair resting on a 3'x3' board. To reduce sound reflections produced by the presense of the board, it was covered with foam. A stand was used to hold the speech material to eliminate extraneous paper rattling. Another stand was used to position an Electrovoice RE55 microphone about nine to twelve inches to the side of the speaker's mouth. A speaker box was placed on the floor facing the wall behind the speaker to allow communication with the speaker while minimizing the amount of extraneous noise which it would produce and to avoid feedback. An Otari MX5050 tape recorder was positioned in the anteroom, and another microphone was present for communication with the chamber.

The setup for recording clear speech is the same as for recording conversational speech, except that a listener (as described in Section 2.2 under the interactive paradigm) was presented with the speaker's speech processed by an AVC plus white masking noise. The speech was recorded on tape prior to processing by the AGC. The listener was seated in the anteroom and used the microphone to the chamber for communication. This setup was diagrammed in Figure 2.1 (Section 2.2).

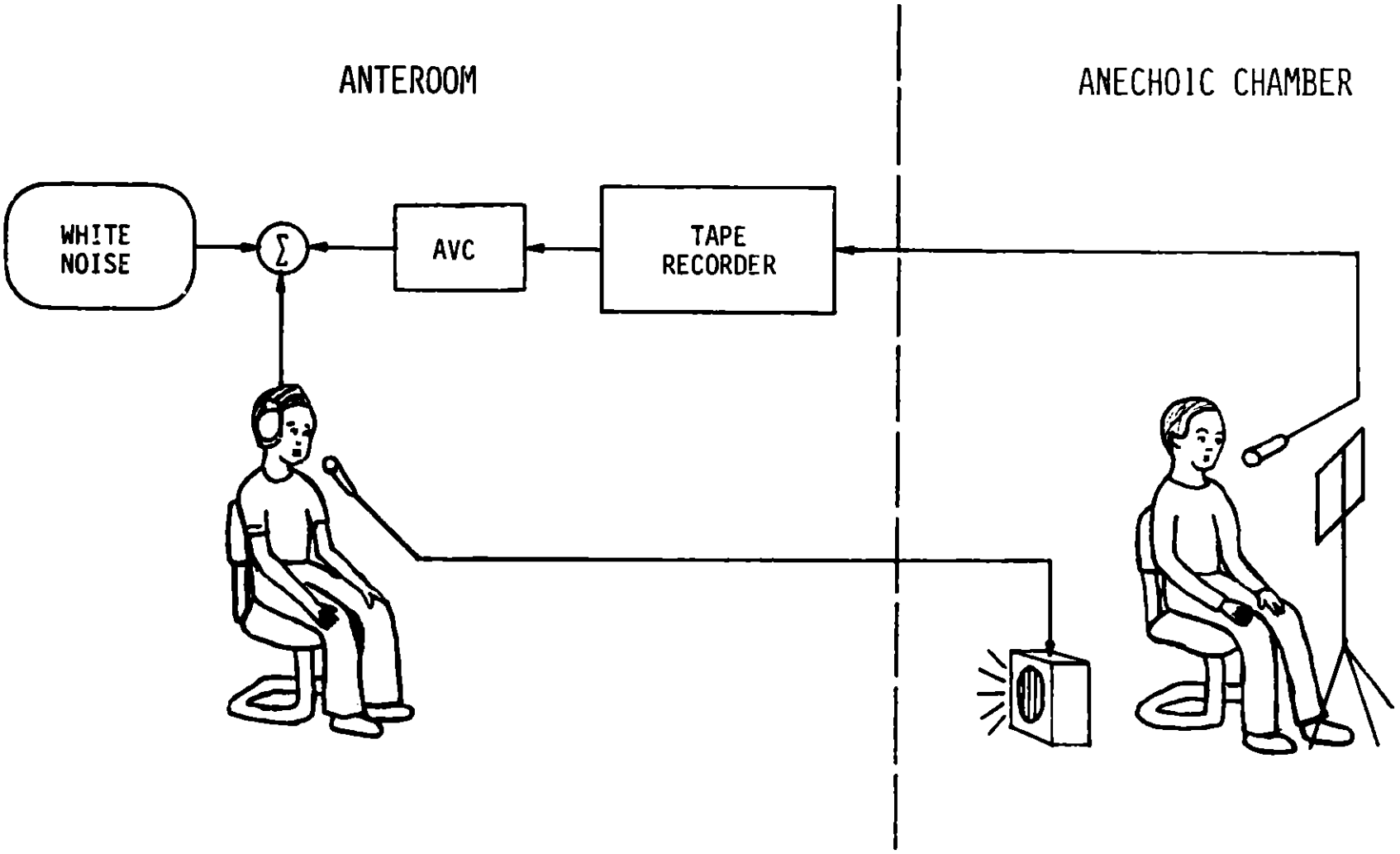


Figure 2.1 Paradigm for elicitation of clear speech

2.4 PROCESSING OF DATA

The same database of waveforms were used in acoustic analysis and in intelligibility tests. Five tokens of each CV recorded by each speaker were processed. Of the ten conversational tokens per CV recorded by a speaker, tokens two, three, four, seven, and eight, were used for analysis, resulting in five tokens per CV analyzed. In cases of mispronunciation or extraneous noise detected in a token after the recording session, the ninth token was used for analysis. The first token was not used because it is characteristically more precise in style, and the sixth token was not used because it is at the beginnning of a sentence group, unlike the tokens which were analyzed. The fifth and tenth tokens were not used because the effects of rapidly decreasing fundamental frequency (F0) characteristic at the end of sentences may have been present.

The recorded analog speech waveforms were first band-limited at 4.5 kHz with a 96 dB/octave rolloff variable lowpass filter to prevent aliasing. They were then digitized to 12 bits at 10 kHz. This sampling rate was chosen because the significant acoustic and perceptual features of stop consonants are below 5 kHz (Cooper et al., 1952). The digitized speech was segmented to extract the phrase /ə'CVp ə/ for use in acoustic analysis and

intelligibility tests since the carrier was not necessary and because the segmentation allowed greater ease in intelligibility testing and reduced the amount of filespace needed to store waveforms. The location of segmentation was chosen to permit formant transitions to be observed and to permit easy and more uniform sectioning of the sentences.

The dc offset of each waveform was removed prior to normalization to eliminate its effect on power calculations. The average power in each token waveform was then normalized to a set value of 1/10 full scale on the D/A. Next, the waveforms were smoothed at each end for 12.5 msec with a raised cosine window to eliminate sudden onsets and offsets in the signal. Copies of the waveforms were transferred to a DECsystem-20 computer, where existing programs were used to compute acoustic parameters from the signal and to phonetically label each waveform by hand in an interactive environment (Woods et al., 1976). A consistent set of rules were used to mark the boundaries of acoutic events. Also available on the DECsystem-20 computer was a data base program which can be used to tabulate and display the computed acoustic parameters (Schwartz, 1976). The data base program provided an interactive environment in which one could run acoustic-phonetic experiments on selected portions of a database of speech waveforms, and then display or tabulate the results as needed.

3. ACOUSTIC ANALYSIS

There are many acoustic characteristics of speech which could be selected for examination. However, the relative importance of each characteristic in its contribution to the intelligibility of speech is not equal. Therefore, selection of parameters for measurement and analysis was done by hypothesizing how important each is in relation to intelligibility. Parameters in both the time domain and the frequency domain were examined. The speech of each speaker was analyzed separately for two reasons. First, the intelligibility results for each speaker differed, and second, each speaker may have used different strategies to make himself clearer.

3.1 DURATIONAL MEASUREMENTS

The acoustic events of the CV's composed of a stop consonant followed by a vowel are of the following form: (1) a period of silence preceeding the release of the consonant in which pressure is built up; (2) the release of the constriction at which fricative noise is generated, typically followed by a period of aspiration for voiceless stops; (3) onset of voicing for the following vowel. The duration of each of these events were measured for the CV. The first segment is called the silence duration or "stop

gap", the second segment is called the voice onset time (VOT), and the third segment is the vowel duration. In voiced consonants, aspiration does not occur; instead, voicing occurs soon after or concurrent with the release of the burst. In this thesis, the voice onset time is measured from the beginning of the burst until the time in which voicing of the vowel begins for both voiced and voiceless consonants. Pre-voicing sometimes occurred when the speakers were trying to speak clearly, but this was ignored in measuring the VOT.

The duration of the stop gap, VOT, vowel, and syllable (defined as VOT duration plus vowel duration) were measured in all 540 tokens. In general, the durational measurements on conversational speech were consistent with past data. In clear speech, the durations of all segments were longer than in conversational speech. The amount by which each segment increased was nonuniform, and in some segments the amount of increase varied with the type of consonant. More detailed results will now be presented.

3.1.1 Silence Duration

A bar diagram of the average value and the range, as marked by the 10 and 90 percentiles of the population, of the stop gap of each stop consonant is shown in Fig. 3.1

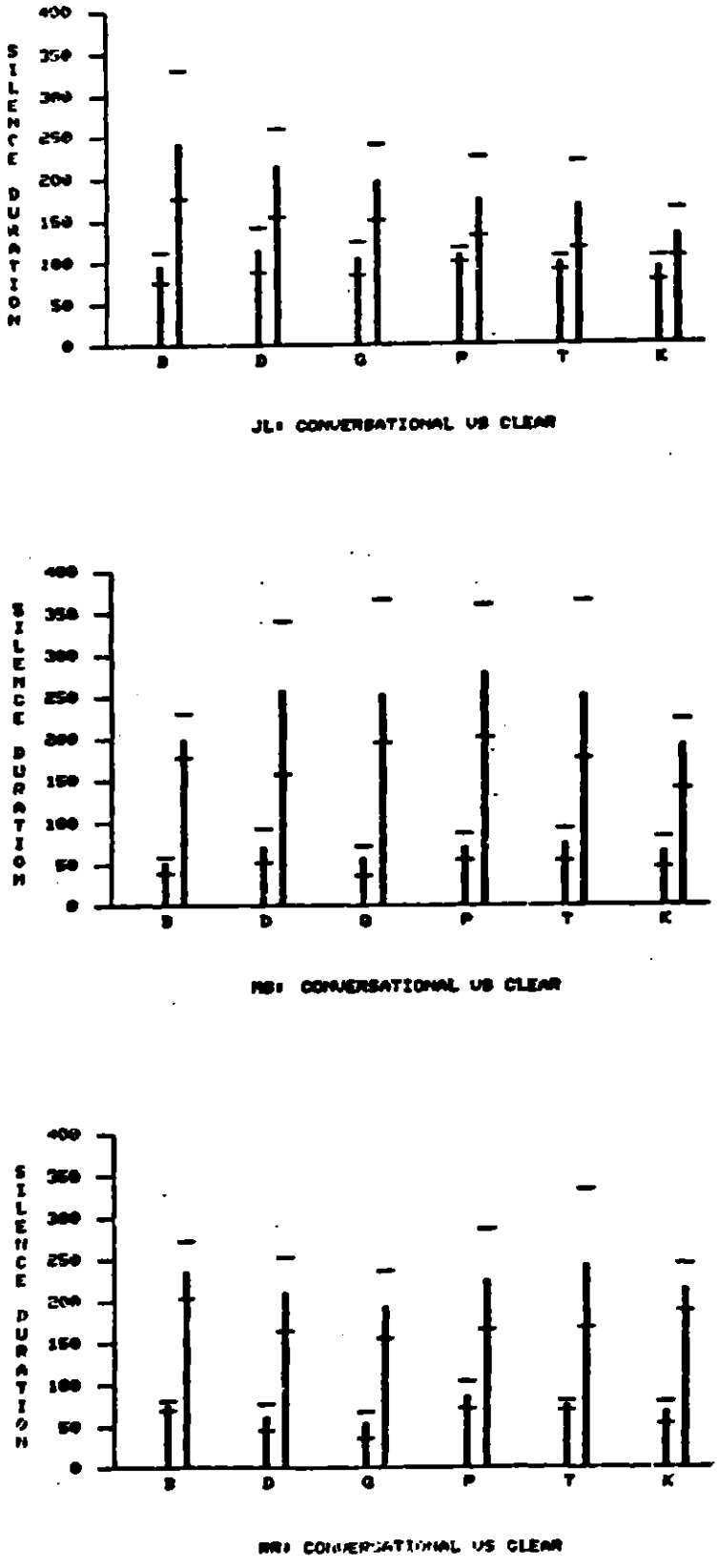


Figure 3.1 Silence duration preceding stop consonants of conversational (left) and clear (right) speech as a function of consonant context

for each speaker. It can be observed that the time preceeding the release of the burst in the stop increased when speaking clearly. In speakers MS and RR, the silence duration increase ranged from 271% to 400%, without any pattern regarding voicing. This great lengthening of the period preceeding the burst may be partially due to pausing at the end of the /hə/ before pronouncing the CV. This would lead to varying silence durations and account for the larger variances observed in the silence duration of clear speech.

One of two patterns was observed in the silence duration of clear and conversational speech by all speakers. In the first pattern, the silence duration preceeding stops decreased from velars to alveolars to labials within each voicing group. This pattern was also observed in voiceless stops by Zue (1976). In the second pattern, the silence duration of labials was about equal to the silence duration of the velars, and the silence duration of the alveolars was larger than both. This pattern was observed in some of the speech from all three speakers. More specifically, this occurred in conversationally spoken voiced consonants by JL, conversationally spoken voiced and voiceless consonants by MS, clearly spoken voiceless stops by RR, and clearly spoken voiced stops by MS.

3.1.2 Syllable Duration

The CV syllable duration is defined to be the time from the beginning of the stop burst until the end of the vowel. In all three speakers, the syllable duration when speaking conversationally was roughly the same among all CV's. However, when speaking clearly, all syllable durations increased in a nonuniform manner as shown in Fig. 3.2. In both speakers RR and MS, the duration of the syllables containing voiceless stops increased much more than the duration of syllables containing voiced stops. Also, the average duration of syllables containing labials was shorter than those containing alveolars, which was shorter than those containing velars. Speaker JL, on the other hand, showed a larger increase in voiced stops than in voiceless stops, this time with the labials having the longest average duration and the velars having the shortest average duration.

3.1.3 Voice Onset Time

The voice onset time of voiced consonants is generally shorter than the VOT of voiceless consonants. Klatt (1975) found that in voiced consonants the average ranged from 11 to 27 msec and in voiceless consonants the average ranged from 47 to 70 msec. In general, the VOT of labials was

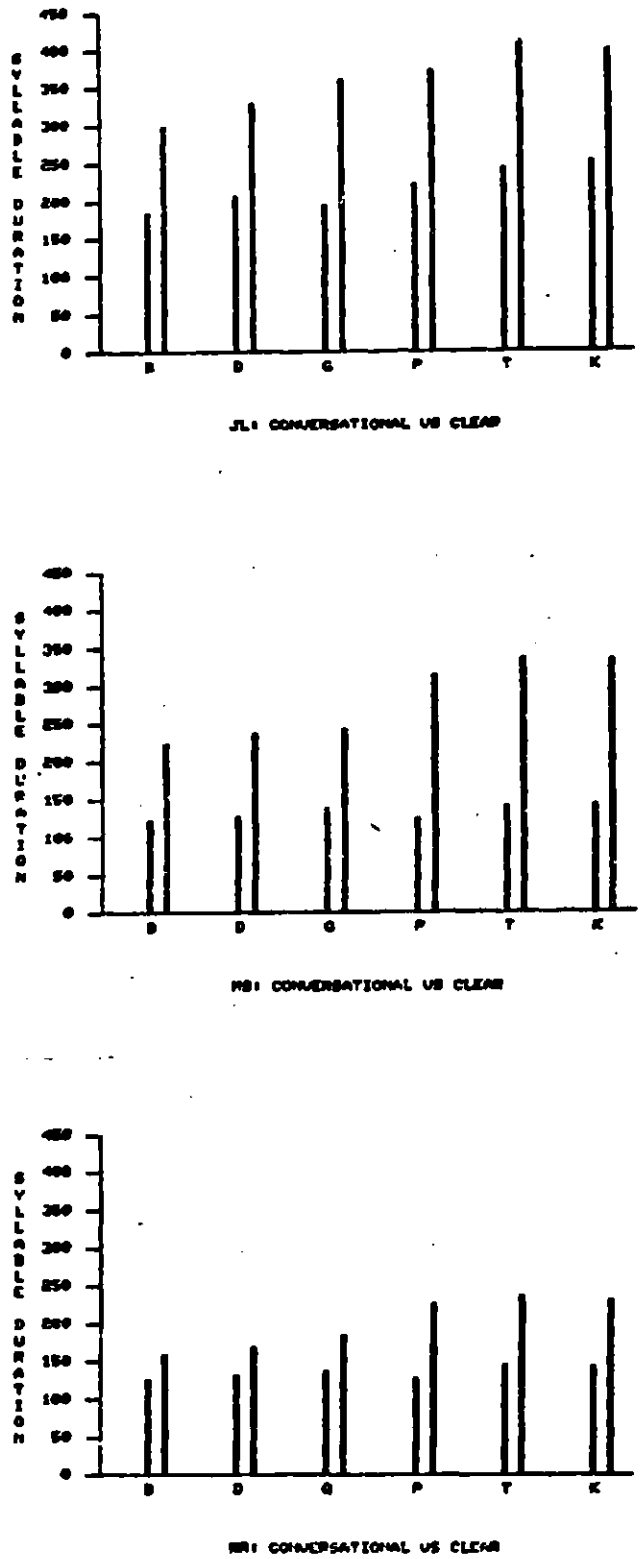


Figure 3.2 Average syllable duration of conversational (left) and clear (right) speech as a function of consonant context

found to be shorter than that of alveolars which in turn was shorter than that of velars. Histograms comparing the voice onset time (VOT) of each speaker are shown in Figs. 3a, 3b, and 3c. In each histogram, the VOT's of voiced (solid line) and voiceless (dotted line) stops were binned separately. The average VOT of the voiced stops was observed always to be shorter than the average VOT of the voiceless stops for all speakers and for each style, in agreement with Klatt's (1979) results. The average VOT of voiced consonants spoken conversationally by JL was in the upper range of VOT values measured by Klatt (1975), while the voiceless consonants had an average VOT value (99.3 msec) which was far greater than that measured by Klatt (1975). The average VOT of voiced consonants (35 msec) spoken conversationally by MS was greater than the average of 18.3 msec measured by Klatt (1975). However, the average VOT of voiceless consonants (55 msec) was concurrent with Klatt's finding of 61 msec. The average VOT of voiced consonants spoken conversationally by RR (23 msec) was slightly greater than the average value measured by Klatt (1975) while that of voiceless consonants was somewhat less (43 msec). In all speakers, the VOT's of the voiced consonants spoken conversationally had a smaller standard deviation than the voiceless consonants. Also, in all speakers, the distributions of the voiced and voiceless consonants spoken conversationally overlapped to a certain extent. This overlap could be due partially to influence of

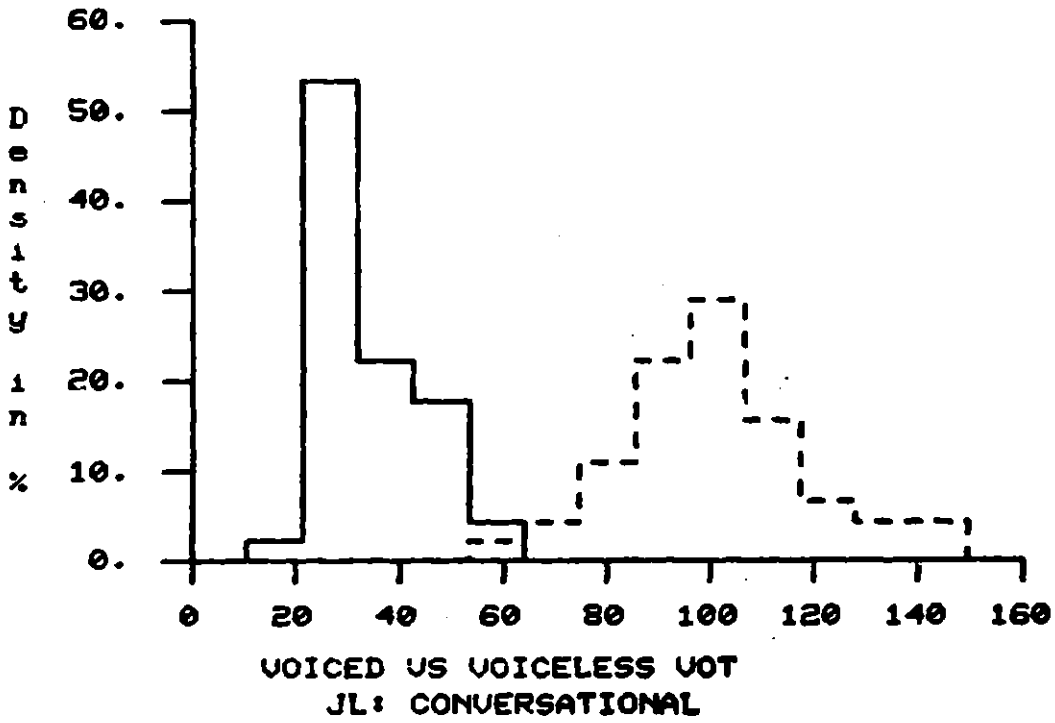
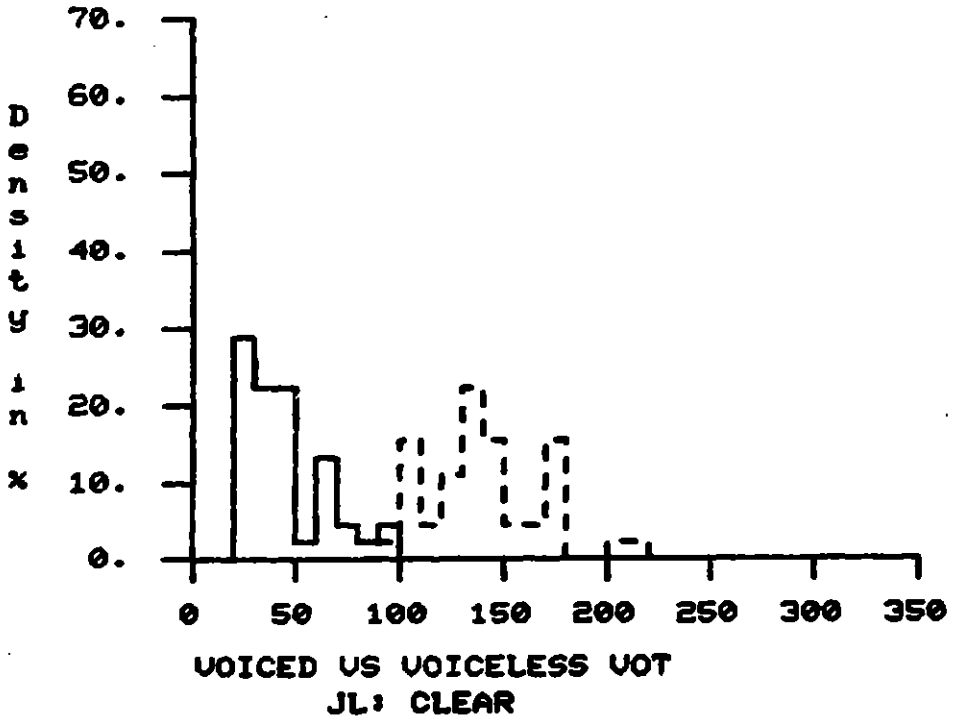


Figure 3.3a Distribution of VOT from voiced (solid line) and voiceless (dotted line) stops for speaker JL



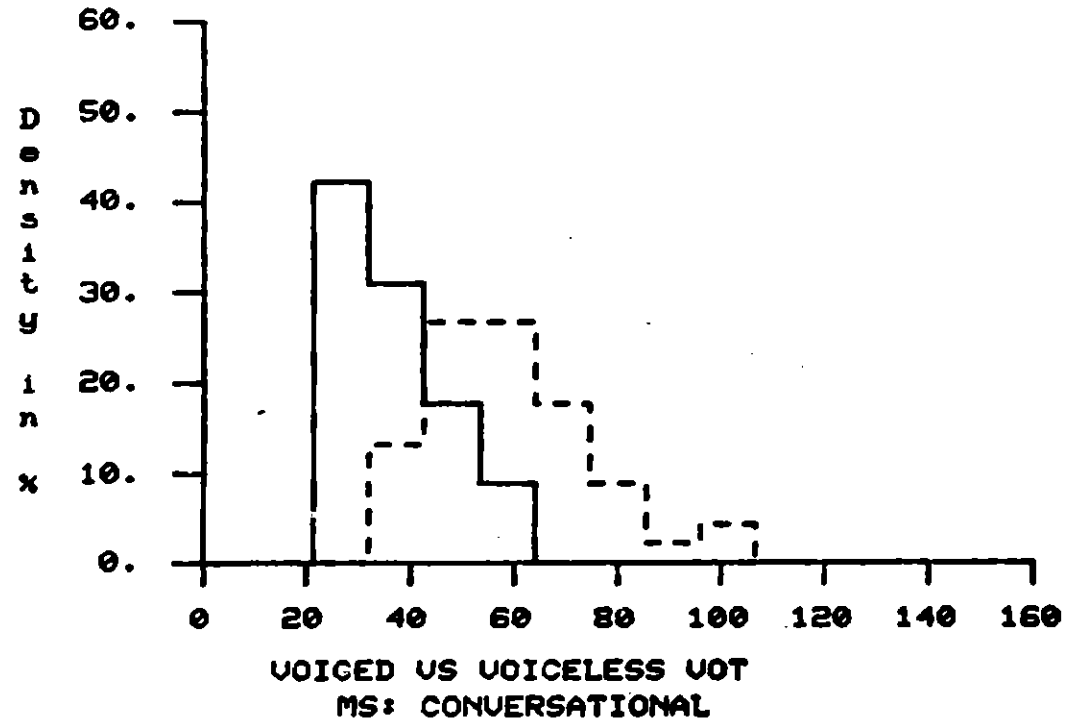
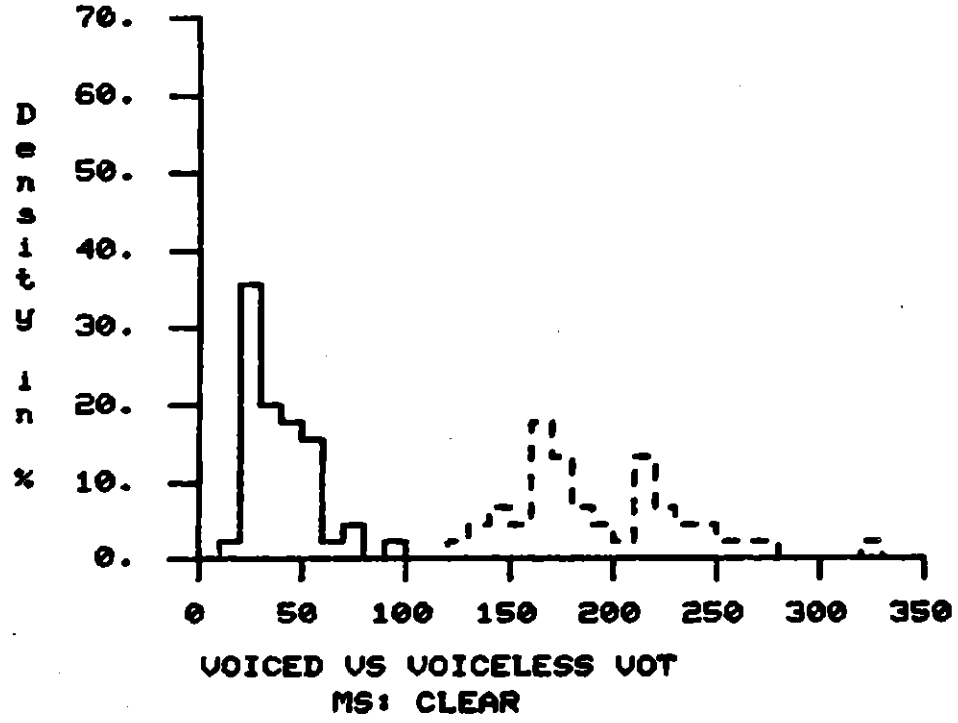


Figure 3.3b Distribution of VOT from voiced (solid line) and voiceless (dotted line) stops for speaker MS



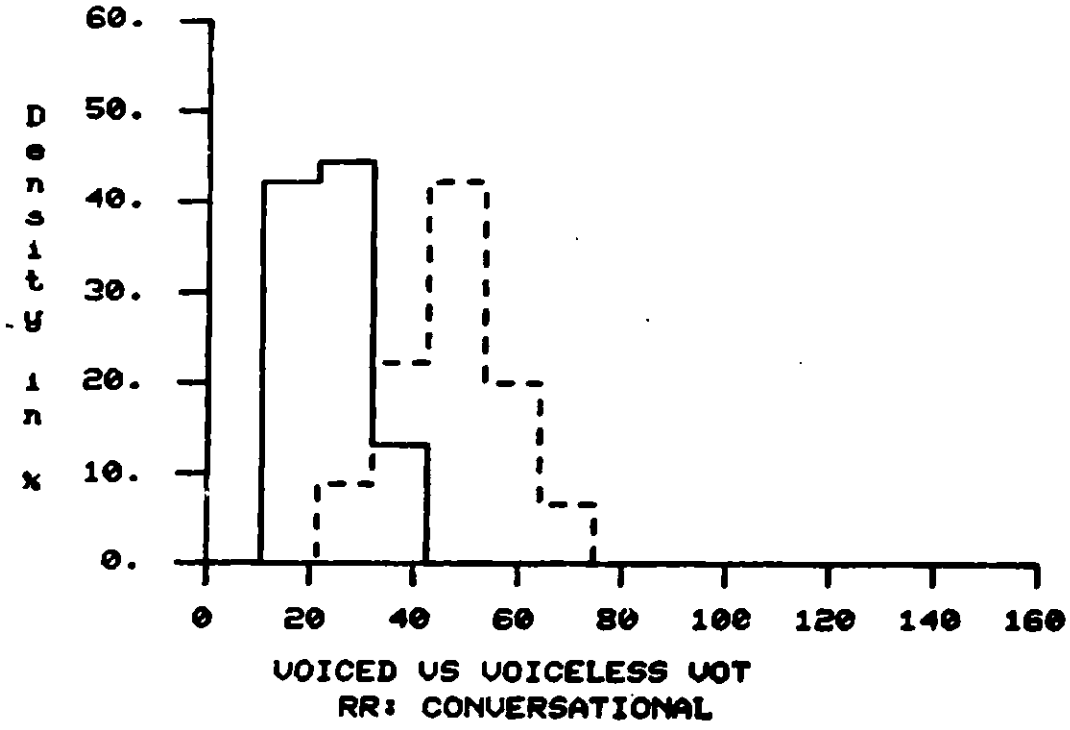
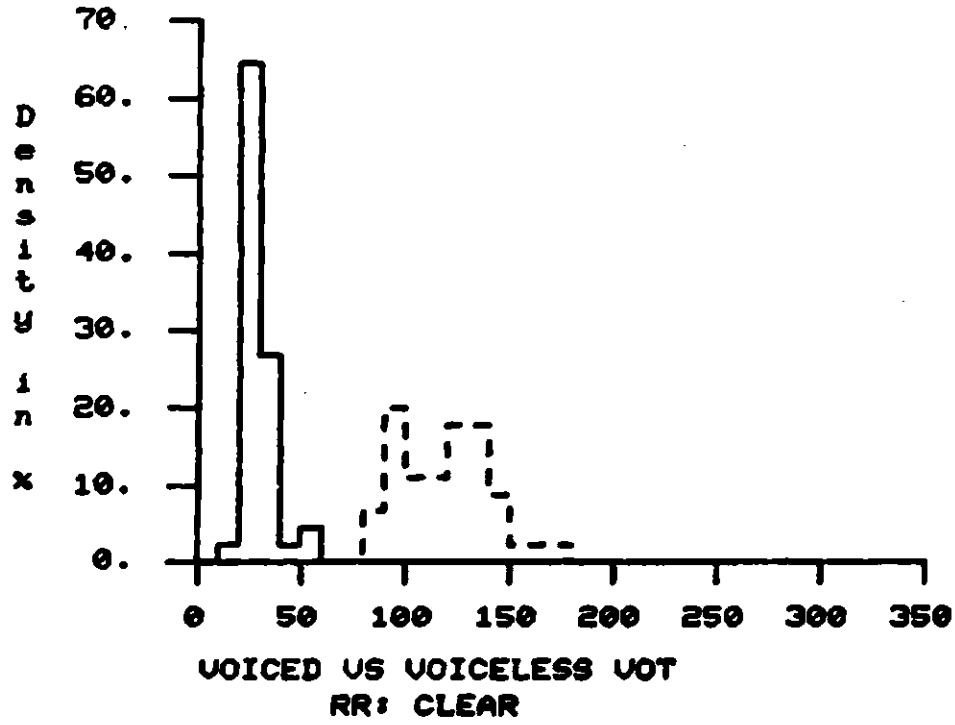


Figure 3.3c Distribution of VOT from voiced (solid line) and voiceless (dotted line) stops for speaker RR



the vowel context. The VOT's of voiced and voiceless consonants spoken clearly were greater than those spoken conversationally in all speakers. The distributions of voiced and voiceless consonants spoken clearly by JL overlapped to about the same extent as the distributions for conversationally spoken consonants. But the distributions of the voiced and voiceless consonants spoken clearly by MS and RR are nonoverlapping.

Fig. 3.4 shows the average value and range of the VOT for each speaker. Klatt's results of labials having a shorter VOT than alveolars, and alveolars having a shorter VOT than velars, provided one compares within a voiced class or voiceless class, was generally upheld in the results on all three speakers, including the clear speech cases. Small exceptions were observed in the conversational speech of MS where the VOT of /t/ and /k/ were about the same, and in the clear speech of RR, where the VOT of /p/, /t/, and /k/ were all about the same.

3.1.4 Vowel Duration

Intrinsically, the duration of /a/ is longer than the duration of /i/ or /u/. Peterson and Lehiste (1960) measured /a/ to be longer than /u/, which was longer than /i/. Zue (1976) measured /i/ and /a/ to be about 14% and

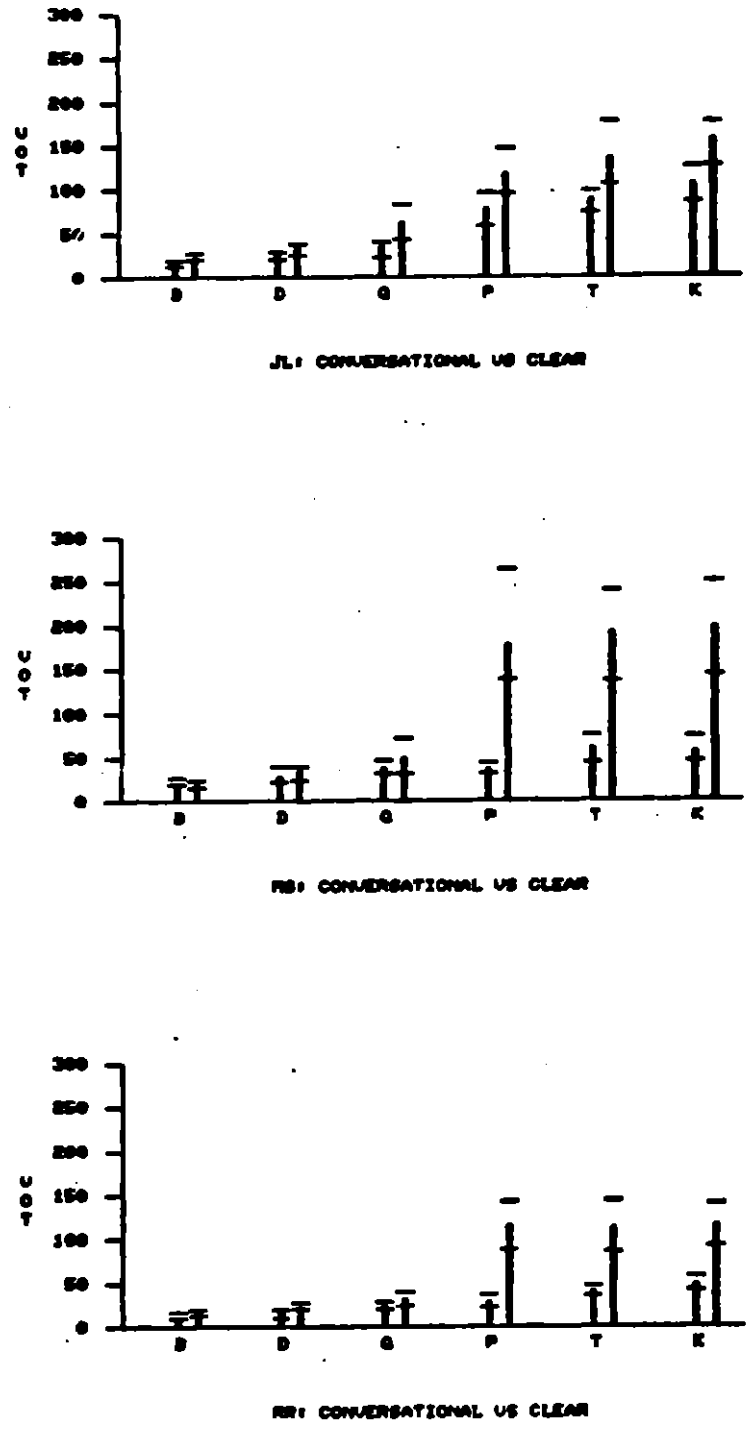


Figure 3.4 VOT of conversational (left) and clear (right) speech as a function of consonant context

67% longer, respectively, than /u/. The measured average vowel durations and corresponding 10 and 90 percentile ranges are shown in the bar diagram of Fig. 3.5. In this thesis, trends similar to past measurements were observed in conversational speech. In all three speakers, the average duration of conversationally spoken /a/ was 30 to 50% longer than the average durations of /i/ and /u/, with /i/ and /u/ durations approximately equal. These same relative orderings of average duration lengths were observed in the clearly spoken vowels of RR and JL. In the clear speech of MS, however, the average duration of /i/ was longer than the average duration of /a/ and /u/, with the /a/ and /u/ duration about the same. The duration of the clear speech vowel increased approximately proportionally to the duration of the conversational vowel in JL's speech. Durations of vowels spoken by JL and RR following voiced consonants were slightly longer than those following voiceless consonants. The duration of vowels following voiced consonants spoken clearly by MS were almost double of those spoken conversationally, while vowels following voiceless consonants increased, but to a lesser extent. The duration of RR's clearly spoken vowels following a voiced stop consonant was substantially longer than those following a voiceless stop consonant.

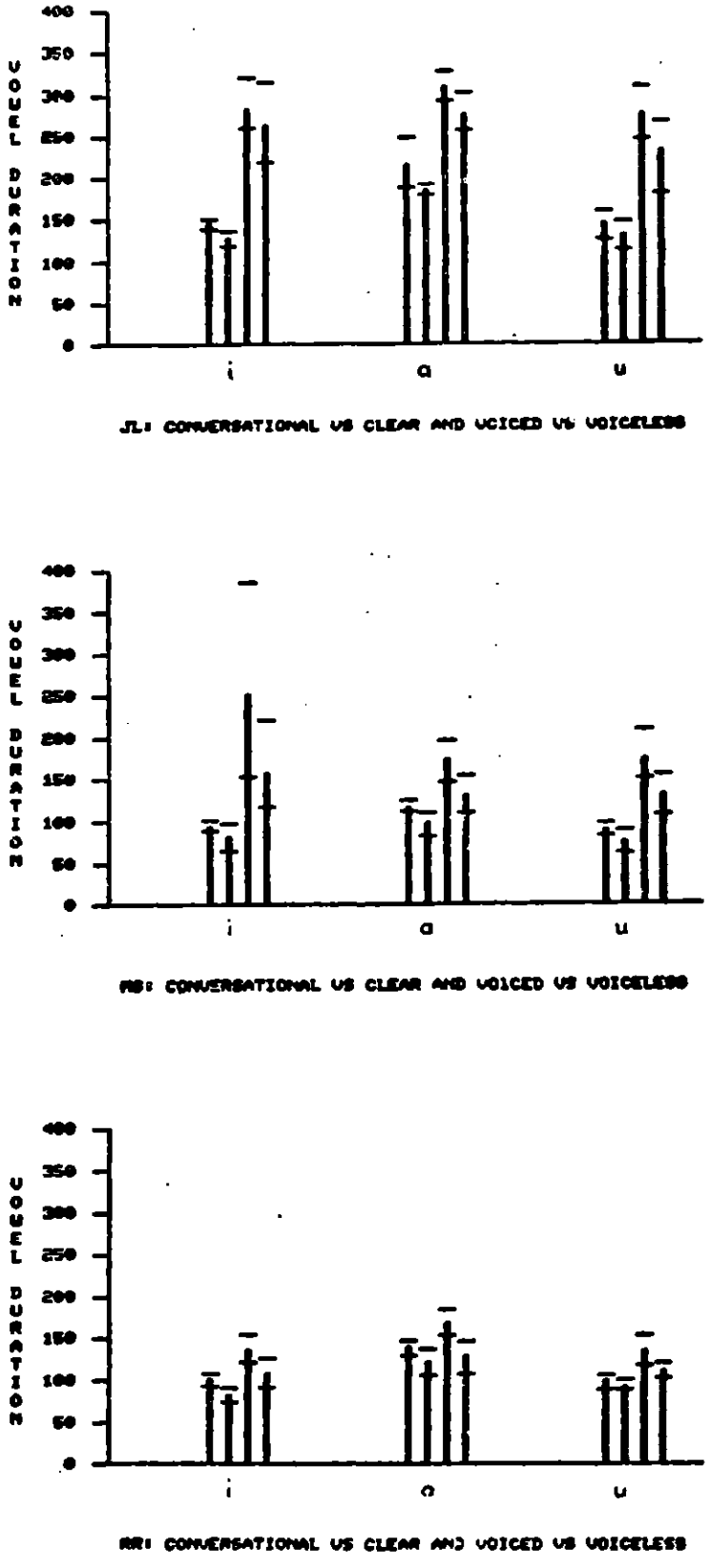


Figure 3,5 Duration of vowels following voiceless conversational stops (leftmost), voiced conversational stops (second from left), voiceless clear stops (second from right), and voiced clear stops (rightmost)

3.1.5 Formant Transition Duration

Following release of the burst, the articulators move towards the targets of the following vowel. Because the articulators have a finite amount of inertia, this movement requires a finite amount of time and usually is not complete until some time into the vowel. When someone attempts to speak clearly, several strategies involving the rate at which this occurs could be hypothesized. One possible strategy may be to keep the transition rate the same or to increase it, which would give the listener more time to hear the vowel. Another strategy would be to decrease the transition rate so the listener hears the transition better. As an estimate of the transition rate, the transition duration was measured. The formant transition rate was not measured because it is very sensitive to small errors in calculating formant frequencies. In rapid speech, the articulators may never reach their "target" value; instead, they may begin to move towards their next "target" position. For this reason and for consistency, the target value of the second formant frequency in a vowel is defined to be the average of the formant frequencies within 30 msec of the maximum of the first formant. The formant frequencies were calculated using a linear prediction formant tracker in which the values were median smoothed. The duration of the formant transition was then defined to be the time from the

onset of voicing until the time in which the formant was within 10% of the formant target of the vowel. The duration of the second formant transition was measured because it is more robust and usually shows a greater change in frequency, allowing a more accurate measurement.

Only voiced consonants were examined for formant transition rates because voicing begins sooner, allowing one to observe the formant transitions more easily. In voiceless consonants, most of the transition occurs during the following aspiration. The formant tracker usually has a great deal of difficulty during this period (because aspiration is noisy) resulting in unreliable measurements of the formant transition.

As shown in Fig. 3.6, in conversational speech, average formant transition durations for vowels following /b/ were shorter than those following other consonants and averaged less than 20 msec. The transition duration of /g/ was longer than that of /d/ in MS and RR, but in JL's speech the transition duration of /d/ was longer than /g/.

In clear speech, average formant transition durations of all vowels were longer than those in conversational speech. Speaker's JL and RR increased the duration of the formant transition in proportion to the original values, so

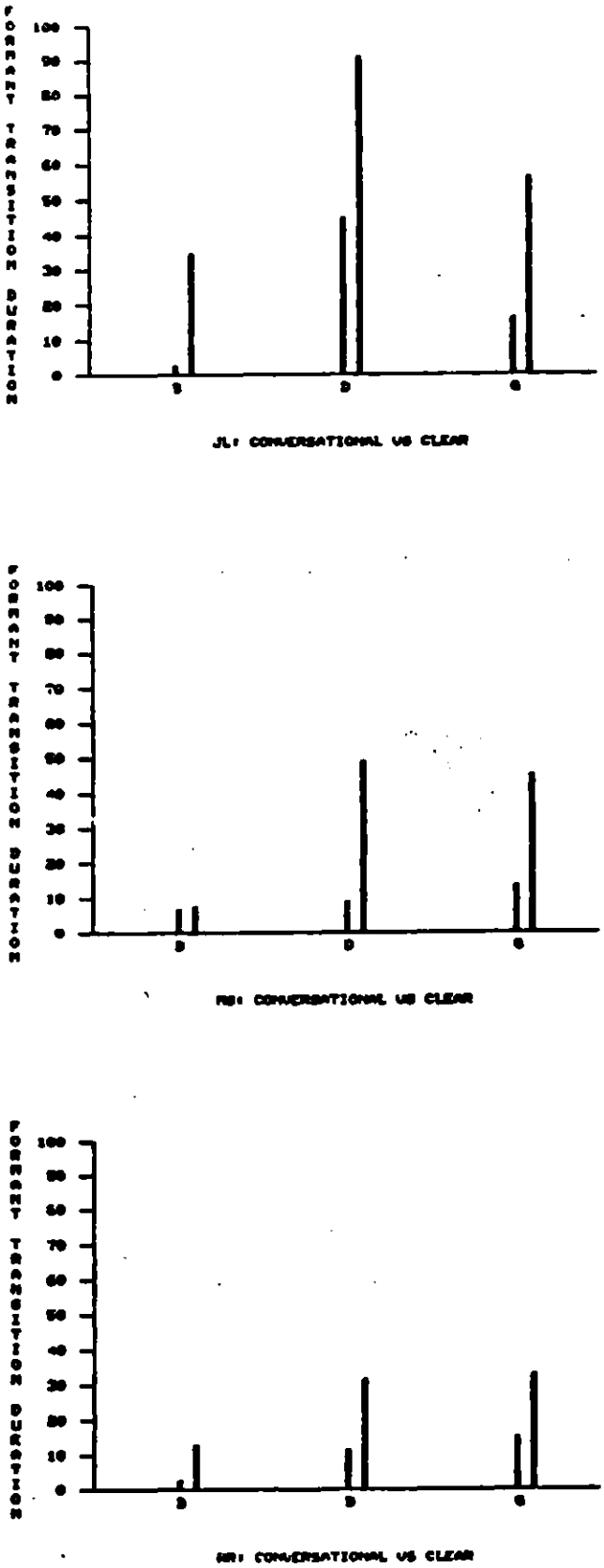


Figure 3.6 Formant transition duration of conversational and clear speech as a function of consonant context

that for JL, the transition duration increases for vowels following a /b/ to /g/ to /d/ and for RR increases from /b/ to /d/ to /g/. Speaker MS, whose clear speech had relative transition lengths similar to that of RR's, now in clear speech exhibits relative transition lengths similar to that of JL.

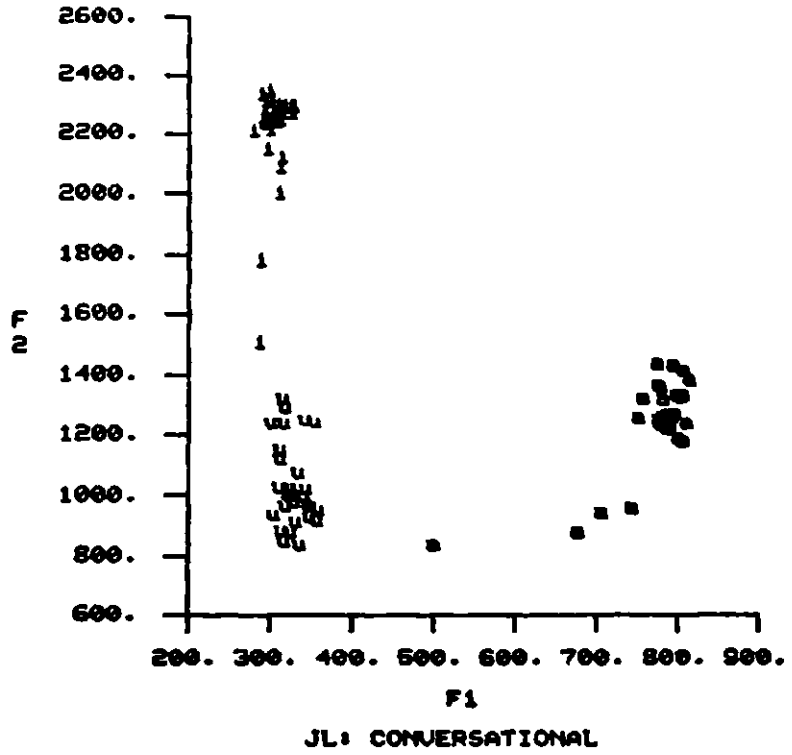
In summary, durational measurements on the conversational speech correspond well with previously reported results (Klatt (1975), Zue (1976)). The increases in duration in clear speech were nonuniform. More specifically, in clear speech the stop gap exhibited wide variances, probably due to pauses. The increase in syllable duration seemed to be dependent upon whether the consonant was voiced, but a consistent pattern was not observed among the speakers. The distributions of voiced and voiceless VOT's were observed to become more distinct and the average value of the VOT's to increase. Although the overall duration of the vowels increased, the relative durations of the vowels were maintained in the clear speech of all speakers except MS. Finally, formant transition durations were observed to increase.

3.2 FREQUENCY AND ENERGY MEASUREMENTS

Consonants and vowels may also be characterized by their spectral properties. Vowels, which are voiced and roughly periodic, exhibit resonant, or formant frequencies which are particular to each vowel. These formant frequencies occur roughly every 1000 Hz, depending upon the length of the vocal tract, but may shift up or down in the production of each of the various vowels. The vowels, when plotted on a second formant (F2) versus first formant (F1) plane, fall approximately within a triangular region spanned by the vowels /i/, /a/, and /u/.

3.2.1 First and Second Formant Frequencies

Figs. 3.7a, 3.7b, and 3.7c show F2 vs F1 for all speakers and for each style of speech. The F1-F2 value of each token is marked by the phonemic symbol for each vowel. All figures show that /i/ has a low first formant and high second formant, that /a/ has a high first formant and low second formant, and that /u/ has a low first and second formant. The conversational speech of MS shows much variance in the formant frequency values. Inspection of his clear speech indicates that the clustering of the vowels have become tighter and also that the vowel triangle, formed by the average value of each vowel, has spread out. The



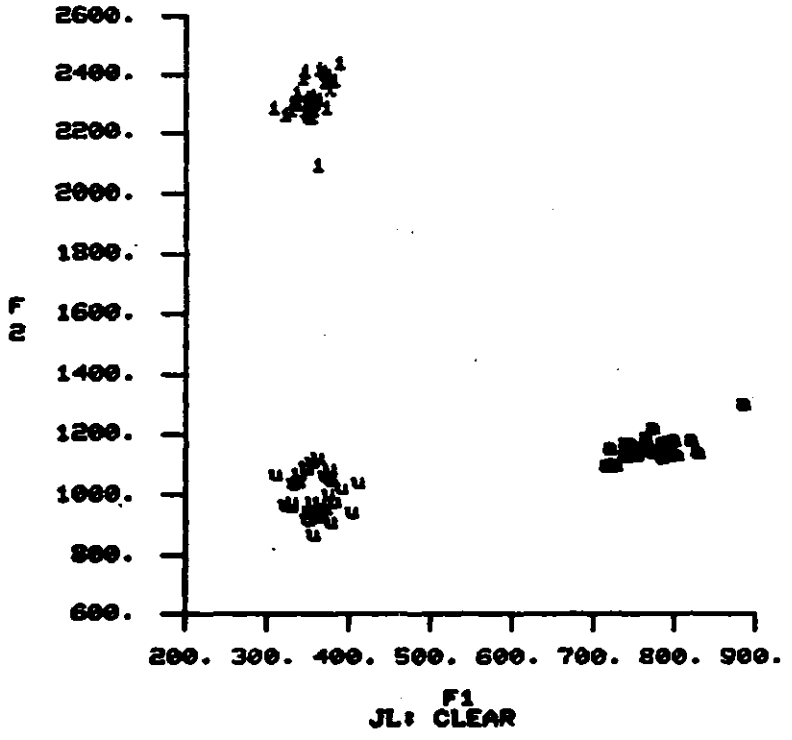
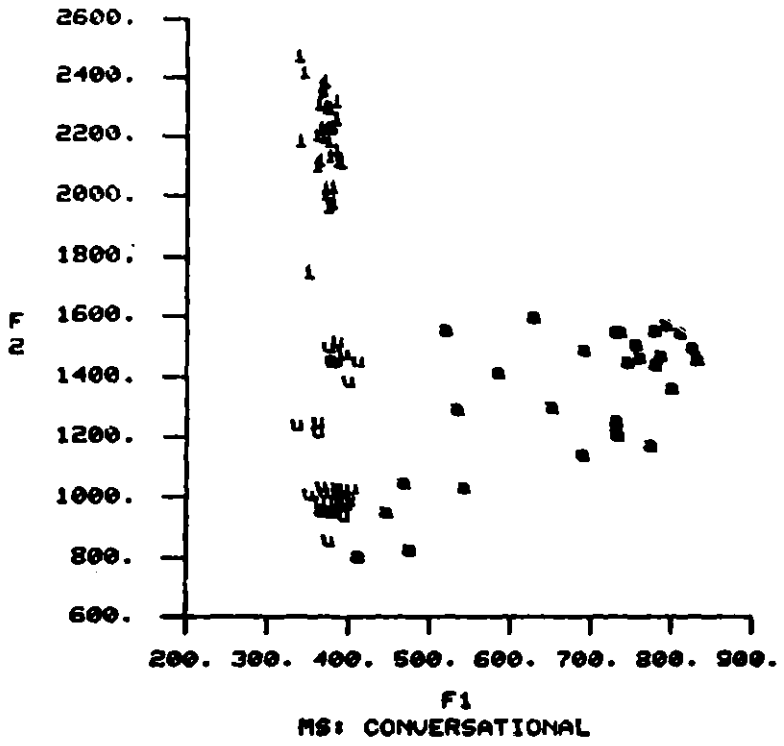


Figure 3.7a F2 vs F1 of vowels conversationally and clearly spoken by JL



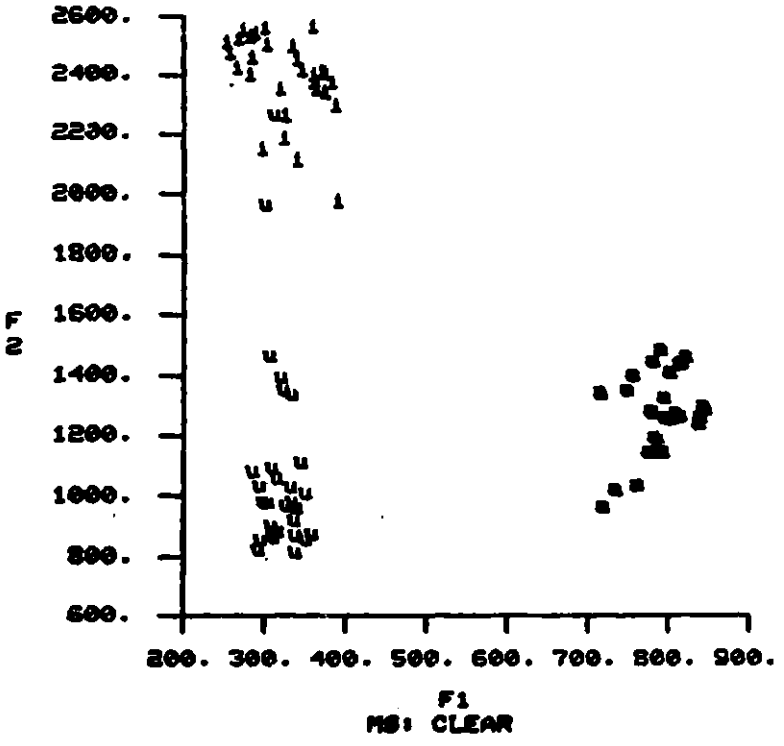
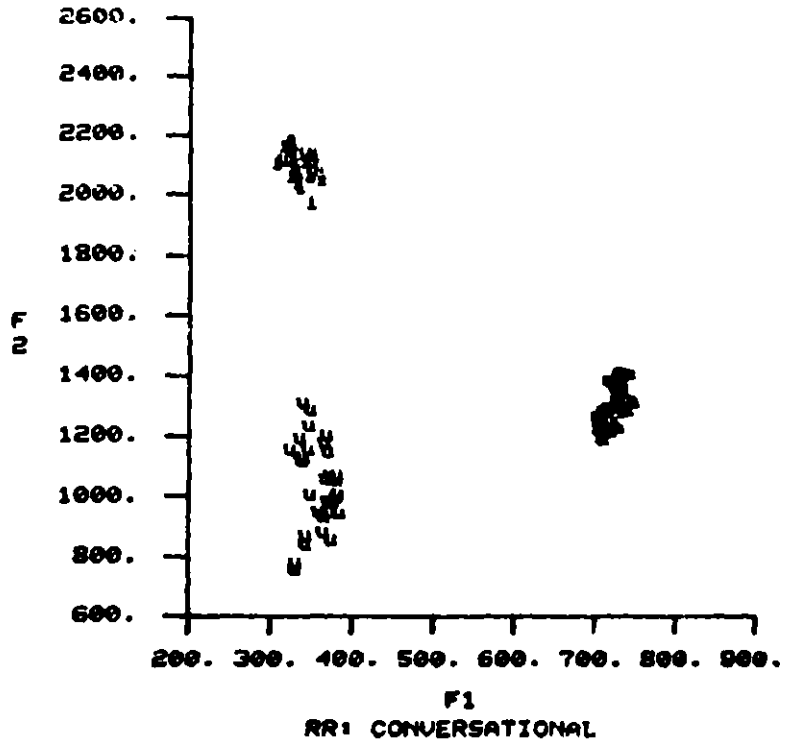


Figure 3.7b F2 vs F1 of vowels conversationally and clearly spoken by MS



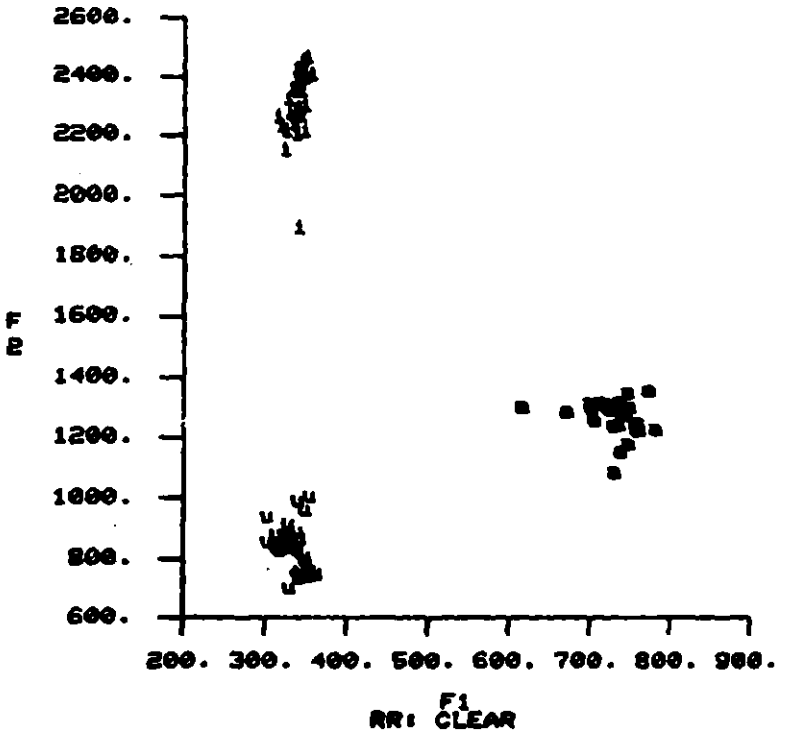
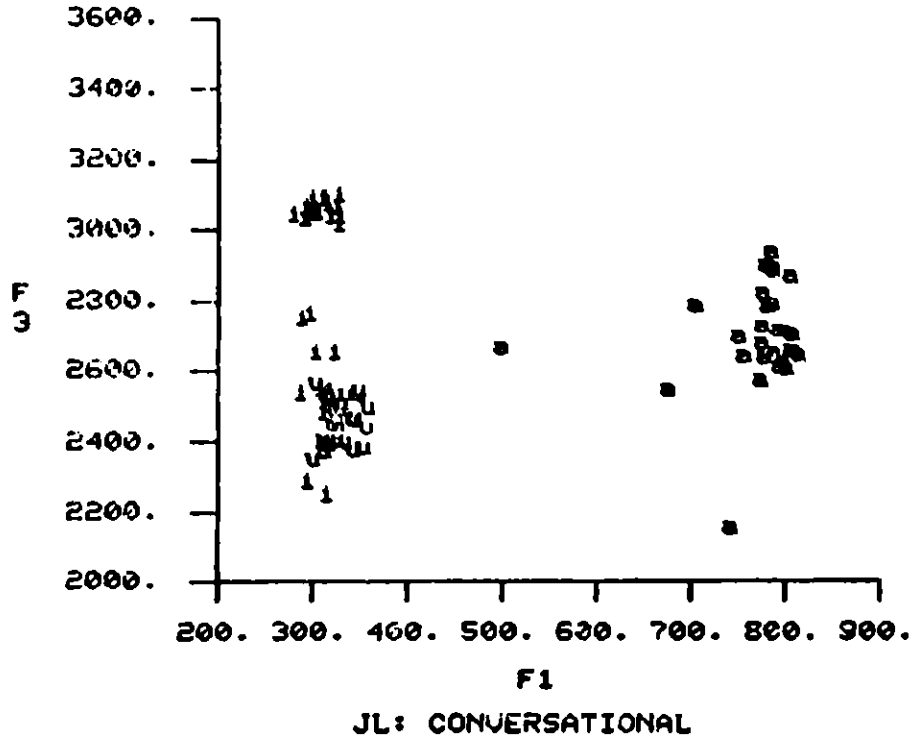


Figure 3.7c F2 vs F1 of vowels conversationally and clearly spoken by RR

conversational speech of JL also exhibits more variablity in formant frequency values than does his clear speech. The vowel triangle also spreads out a bit in clear speech, although the bulk of the change in average values of point vowels is due to the variance of formant values near the center of the vowel triangle. The first formant of /i/ and /u/ in speaker JL has increased. The second formant value of /a/ in speaker JL has decreased while the first formant value remained about the same, resulting in a change in shape of the vowel triangle. Speaker RR also exhibited a spread in his vowel triangle, but the clustering of his vowels decreased.

3.2.2 Third Formant Frequency

The third formant frequency (F3) may be plotted as a function of F1 or as the difference between it and F2 versus the difference between F2 and F1. In Figs. 3.8a, 3.8b, and 3.8c, F3 vs F1 is plotted for conversational and clear speech. In the conversational speech of all speakers, it can be observed that the third formants of /i/, /a/, and /u/ are all within the same relative region. The average values of the third formant decreased from /i/ to /a/ to /u/ in all speakers. Examination of the third formant in clear speech reveals that the average values for the three vowels have separated. The average value of F3 in /i/ has significantly



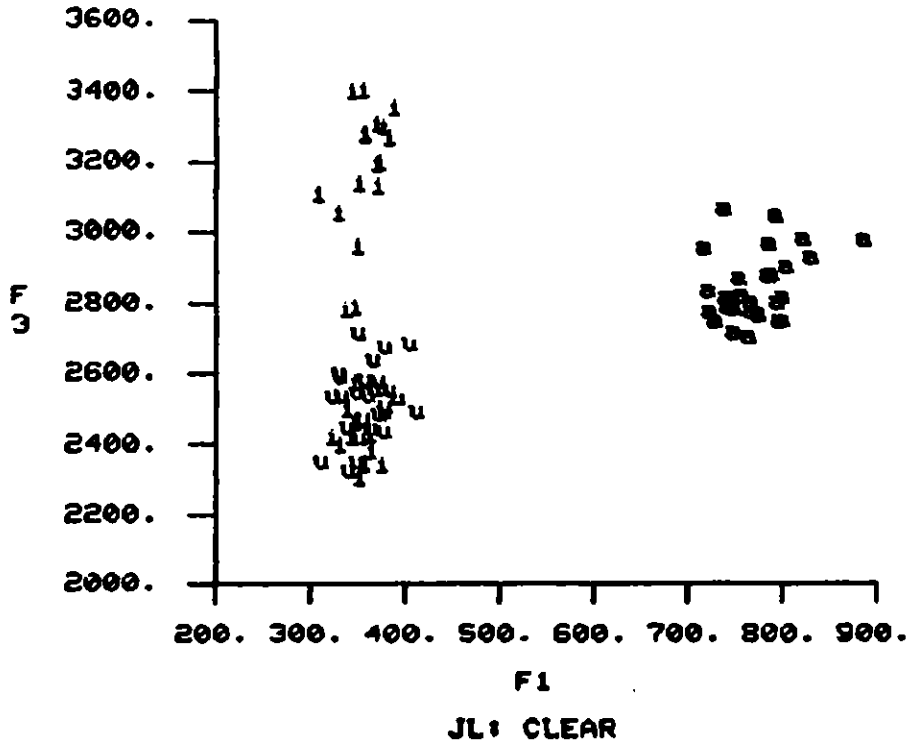
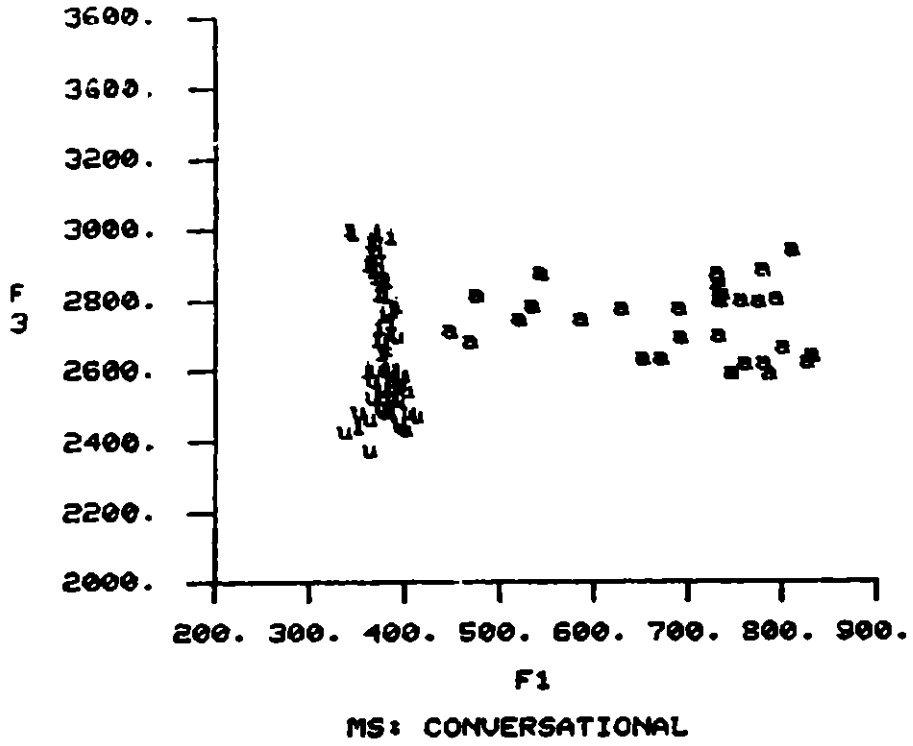


Figure 3.8a F3 vs F1 of vowels conversationally and clearly spoken by JL



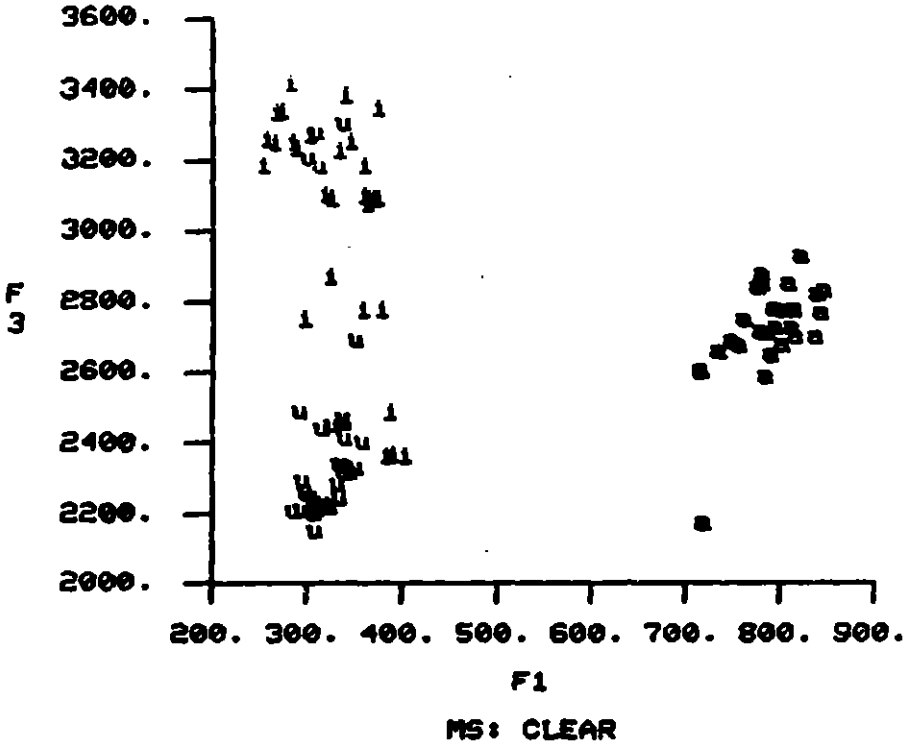
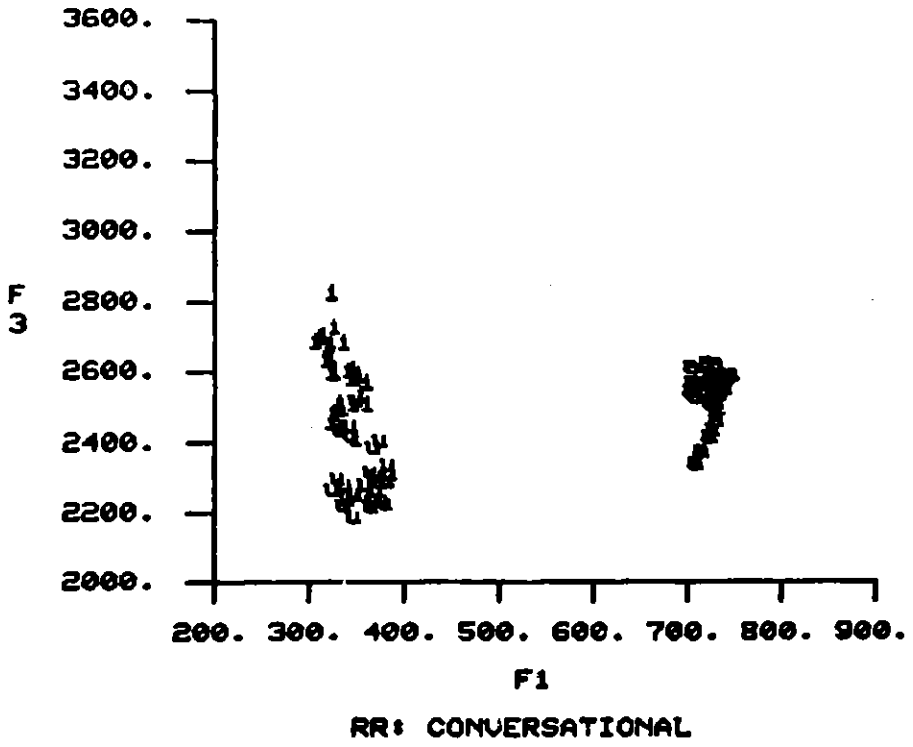


Figure 3.8b F3 vs F1 of vowels conversationally and clearly spoken by MS



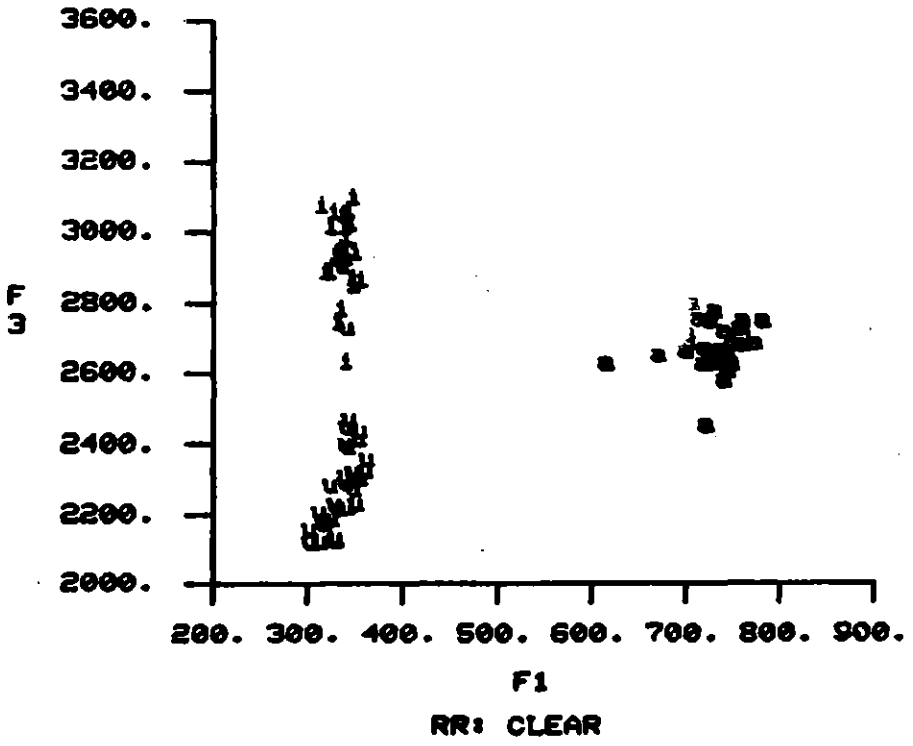
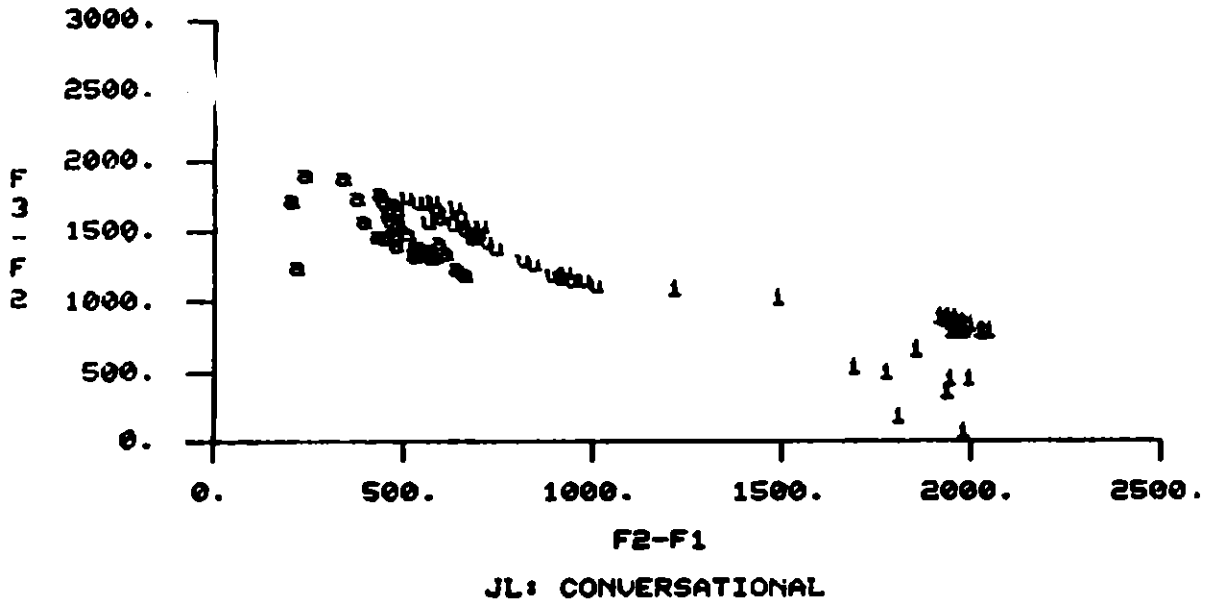


Figure 3.8c F3 vs F1 of vowels conversationally an clearly spoken by RR

increased in all speakers. It is larger than the value of /a/ and /u/ in clear speech. In RR, F3 in /a/ and /u/ remained approximately the same. In MS's speech, the average value of F3 in /a/ remained about the same, while it decreased in /u/. In speaker JL, the third formant frequency of /a/ increased to the value of /i/ in conversational speech, but /u/ remained about the same. There also was tighter clustering of the formant frequencies, resulting in much less overlap of frequency values.

In the figures on F3-F2 vs F2-F1 (Figs. 3.9a, 3.9b, and 3.9c), it can be observed that /i/ is set apart from /a/ and /u/ by a large difference between F1 and F2. It also exhibits a smaller difference between F2 and F3. /a/ and /u/, however, are less distinguishable. Because they both have low second formants, but /a/ has a high first and /u/ a low first, the difference between the first two formant frequencies should be greater for /u/. This is observed to some degree in all three speakers when speaking conversationally. Note that the difference between the second and third formant frequencies is about the same in /a/ and /u/ for all three speakers when talking conversationally. In the clear speech of JL and MS, the values of /a/ and /u/ are slightly more distinct, while the speech of RR becomes even less distinct.



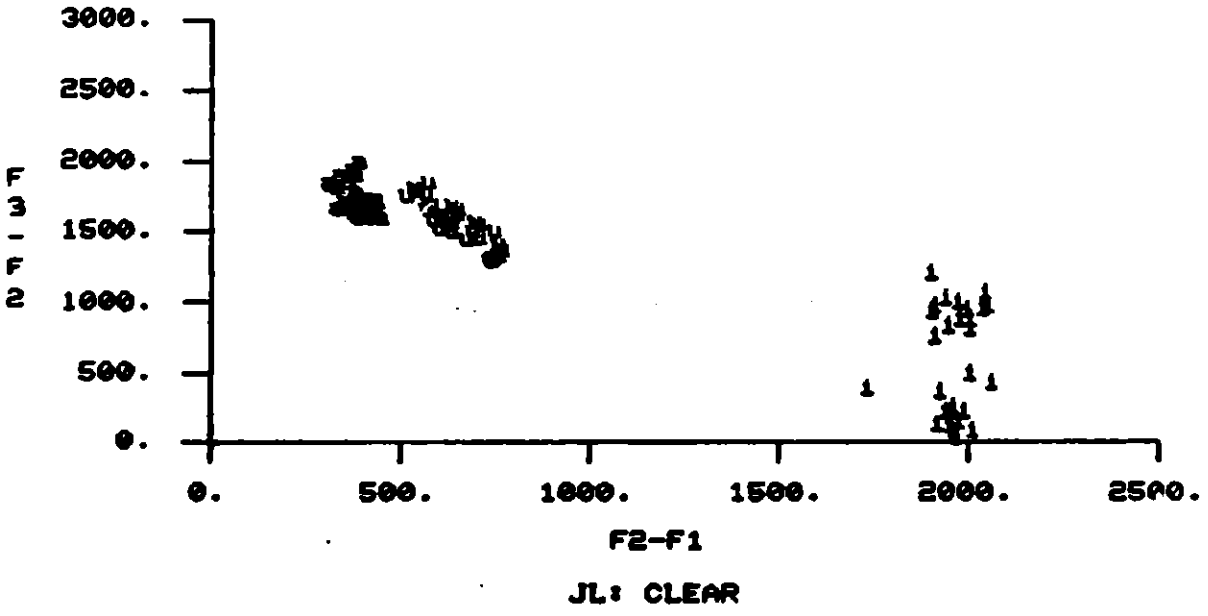
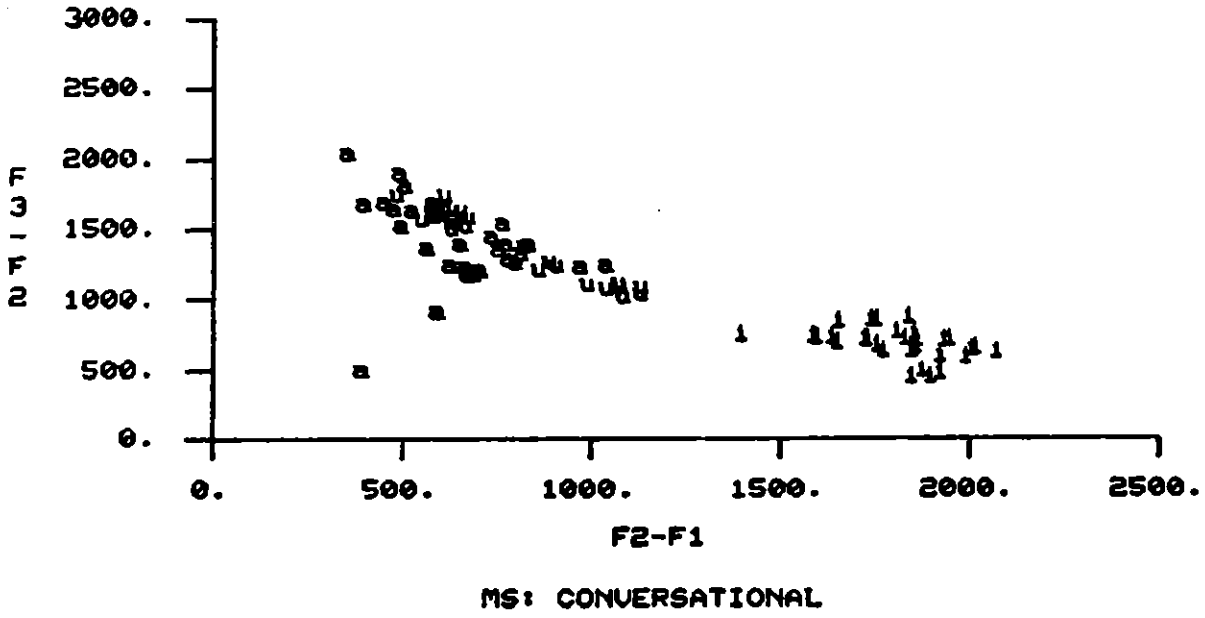


Figure 3.9a F3-F2 vs F2-F1 of vowels conversationally and clearly spoken by JL



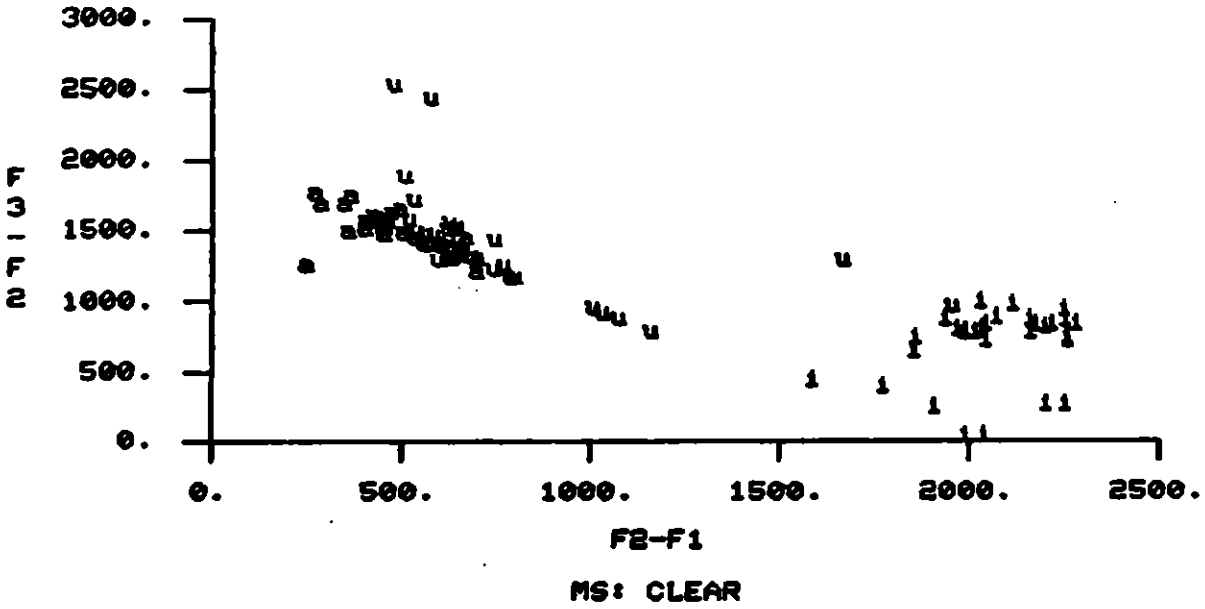
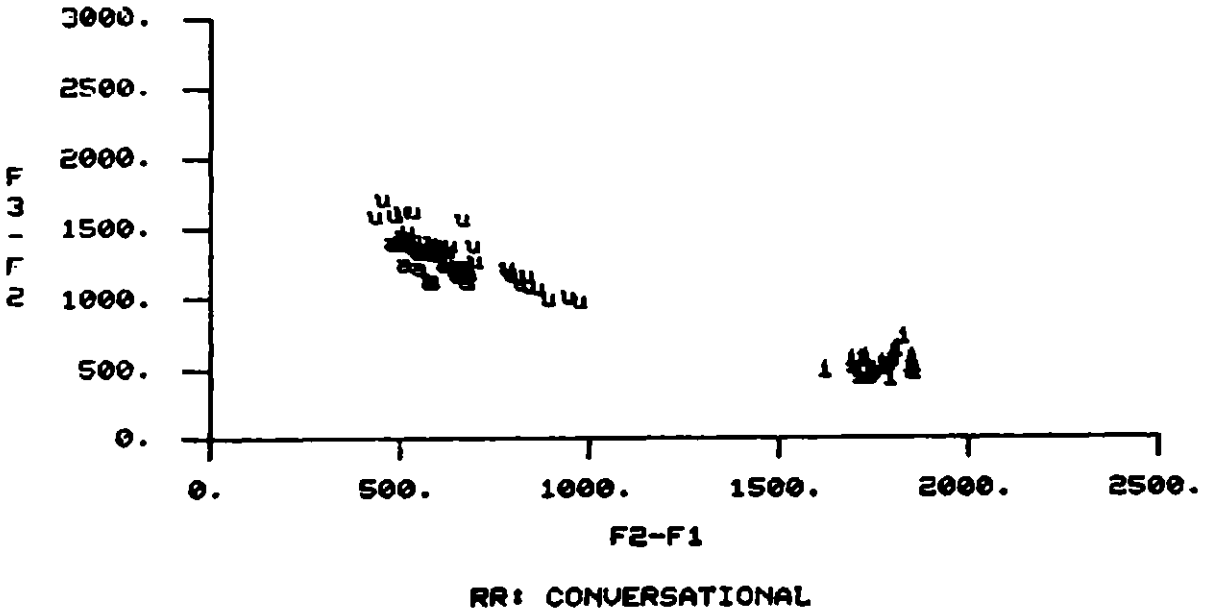


Figure 3.9b F3-F2 vs F2-F1 of vowels conversationally and clearly spoken by MS



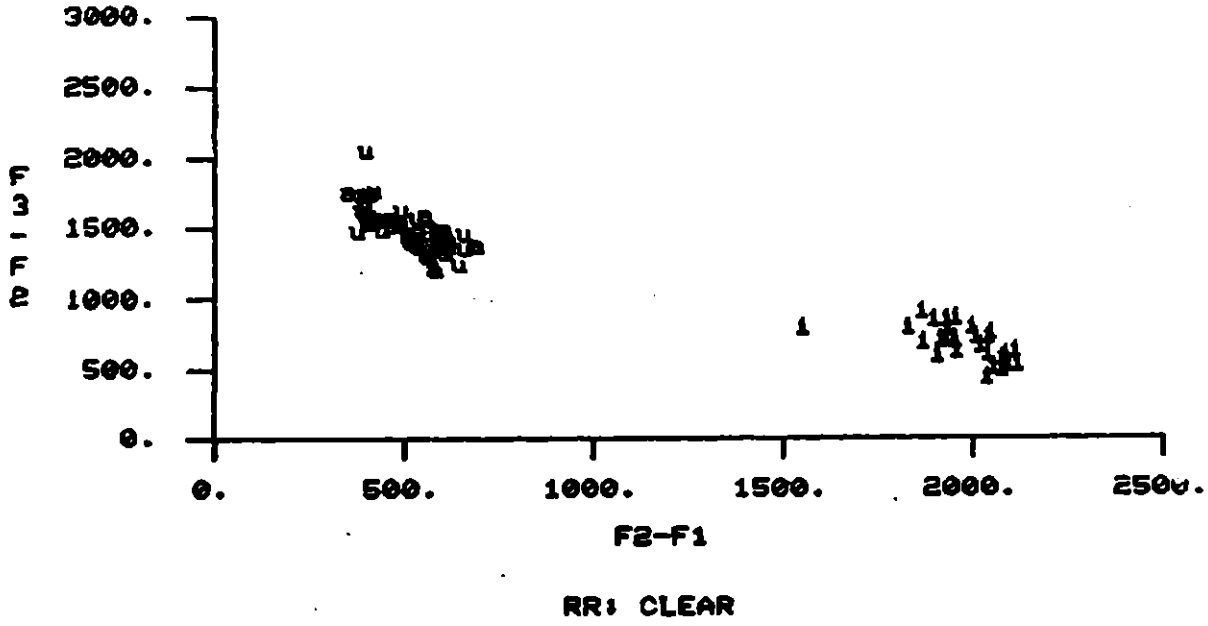


Figure 3.9c F3-F2 vs F2-F1 of vowels conversationally and clearly spoken by RR

3.2.3 Fundamental Frequency

The frequency with which our vocal cords vibrate may be varied through control of the muscles near the larynx. In the conversational speech of all three speakers, the fundamental frequency (F0) of the vowel decreased from /u/ to /i/ to /a/, as shown in Fig. 3.10. When speaking clearly, F0 dramatically increased, and F0 for /i/ became closer to that of /u/. The amount of increase in pitch was dependent upon the speaker. MS exhibited the largest increase in pitch and RR the least.

An increase in pitch is usually effected by raising the larynx, thus decreasing the length of the back cavity of the vocal tract. This causes a corresponding increase in the frequency of the formants associated with the back cavity. However, the upward shift in the larynx is usually small, resulting in a small change in the formant frequency of the back cavity. Consequently one must be cautious in attributing shifts in formants from conversational to clear speech entirely to the change in pitch. In speaker JL, the first formant of the high vowels /i/ and /u/ are observed to shift upwards when speaking clearly. Because /i/ and /u/ are high, the area of the back cavity is unconstricted, resulting in a Helmholtz resonance which is the first formant and which increases in frequency when the pitch

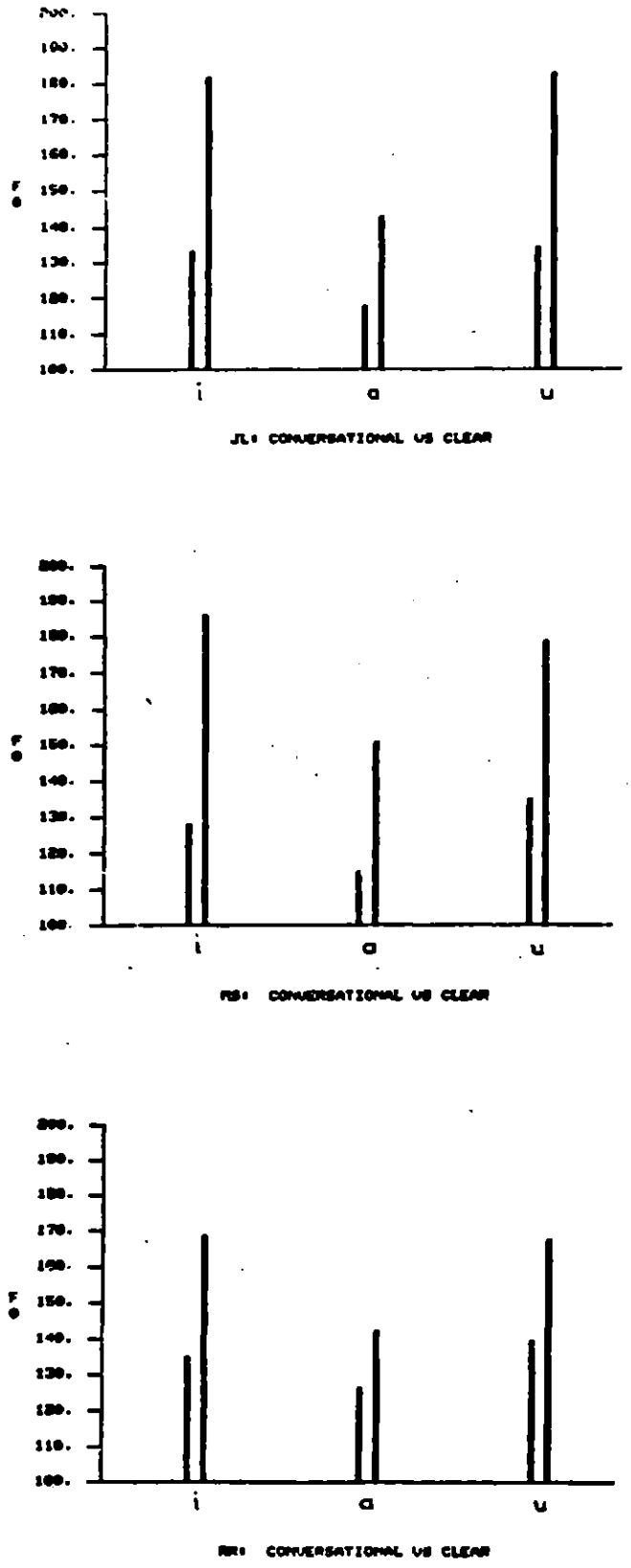


Figure 3.10 Fundamental frequency of conversationally (left) and clearly (right) spoken vowels

increases. This increase in F1 for JL and less of a shift in speaker RR corresponds to the lesser amount of increase in pitch for RR than JL. On the other hand, MS exhibited a decrease in F1 with increased pitch. This indicates that MS is changing something else which has a stronger effect in lowering F1 than raising the pitch has in raising F1.

3.2.4 CV Ratios vs. Average Burst Frequency

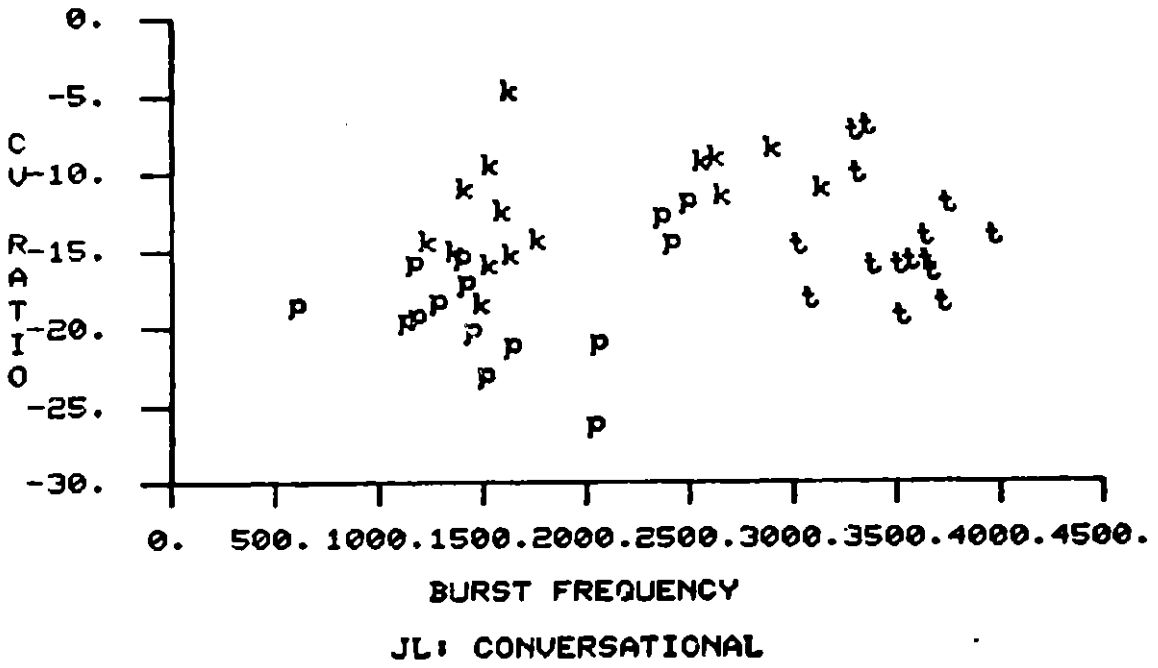
The difference between the amplitude of the energy in the burst and the maximum amplitude of energy in the vowel, in dB, was called the CV ratio. The "average value of the burst" was computed by finding the point on the burst spectrum in which 75% of the "mass" of the first moment is above that frequency. This value is related to the burst frequency because the "mass" of the spectrum is most concentrated at the value of the burst frequency. Because the 75% center of mass point is used, rather than the 50% center of mass point, low to mid burst frequencies are well represented, but high burst frequencies tend to appear lower. Despite such limitations, the relative order of the burst frequencies is preserved.

Burst frequencies were computed for voiceless stops only. Computation of burst frequencies for voiced stops may include transition information because the VOT in voiced

stops is very short and because the program constrained the window used in computing the burst spectrum to be 20 msec. A window this long may include the beginning of a vowel following a voiced stop. However, the VOT in voiceless stops is longer, allowing computation of the burst spectrum without including the formant transitions of the vowels.

Overall, the measurements on conversational tokens again corresponded well with Zue's (1976) results. The burst frequency of /t/ was generally higher than that of /p/, and the burst frequencies of /k/ formed two groups which were dependent upon whether the following vowel was a front or back vowel. The average CV ratio of /p/'s also was observed to be lower than the average CV ratio of the other voiceless stops in conversational speech.

The CV ratio vs. burst frequency of speaker MS is shown in Fig. 3.11b. It is seen that the burst frequencies are closer together in clear speech than in conversational speech. The burst frequency of /k/ is observed to form two clusters, one from 900 to 1700 Hz and another from 2500 to 3300 Hz in conversational speech. The /k/'s which form the cluster of high burst frequencies are all part of the CV /ki/ while the /k/'s with the lower burst frequency are part of the CV's /ka/ and /ku/. These results are in agreement with Zue (1976). Because articulatory movements to prepare



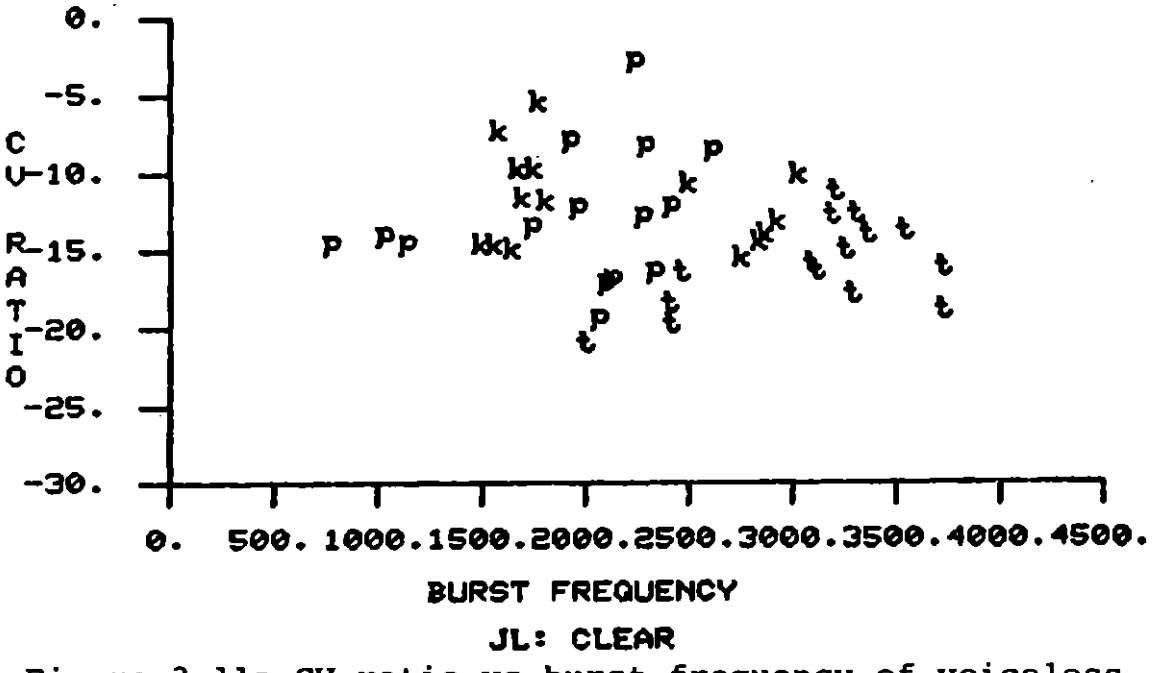


Figure 3.11a CV ratio vs burst frequency of voiceless consonants conversationally (left) and clearly (right) spoken by JL

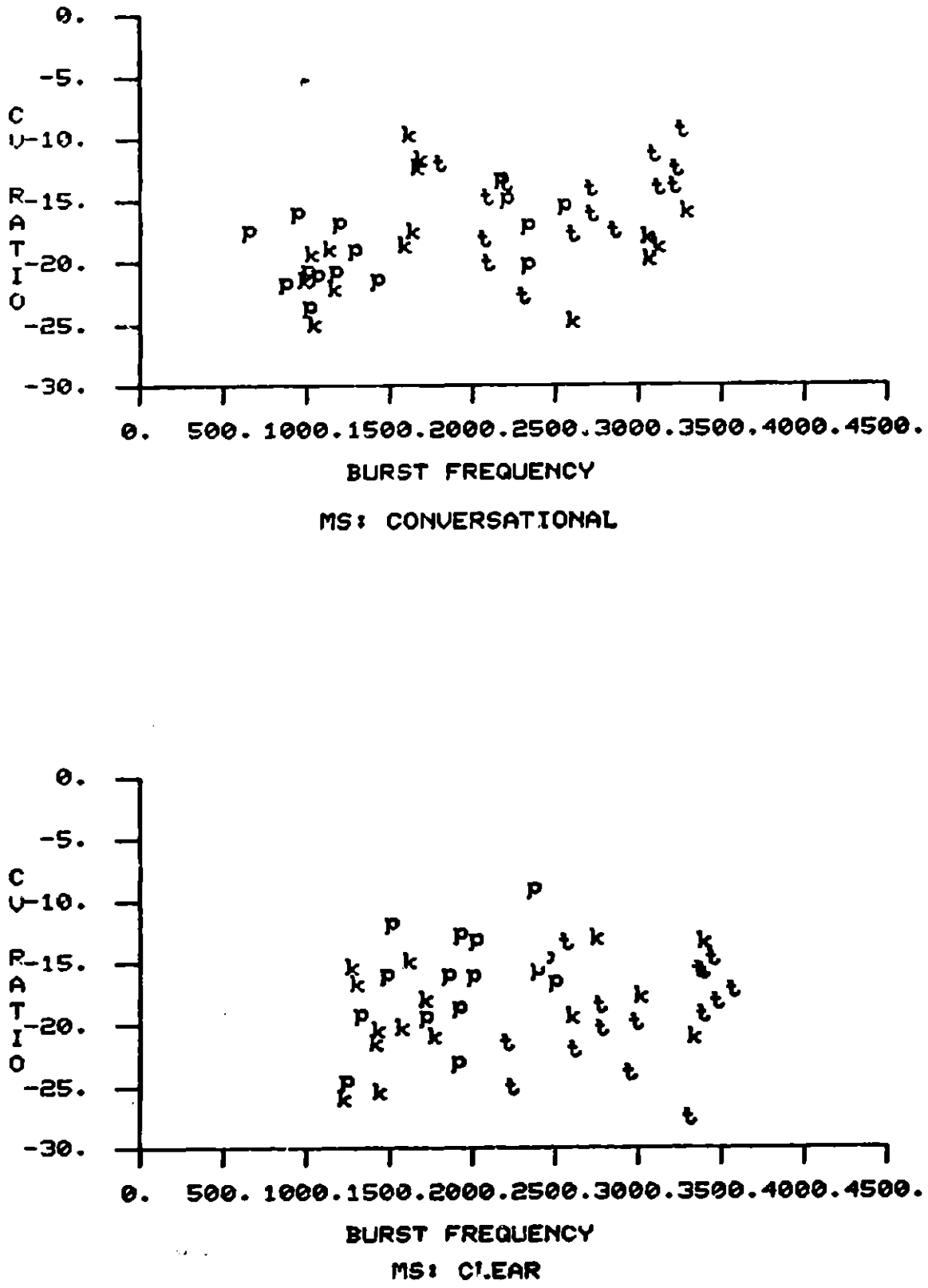
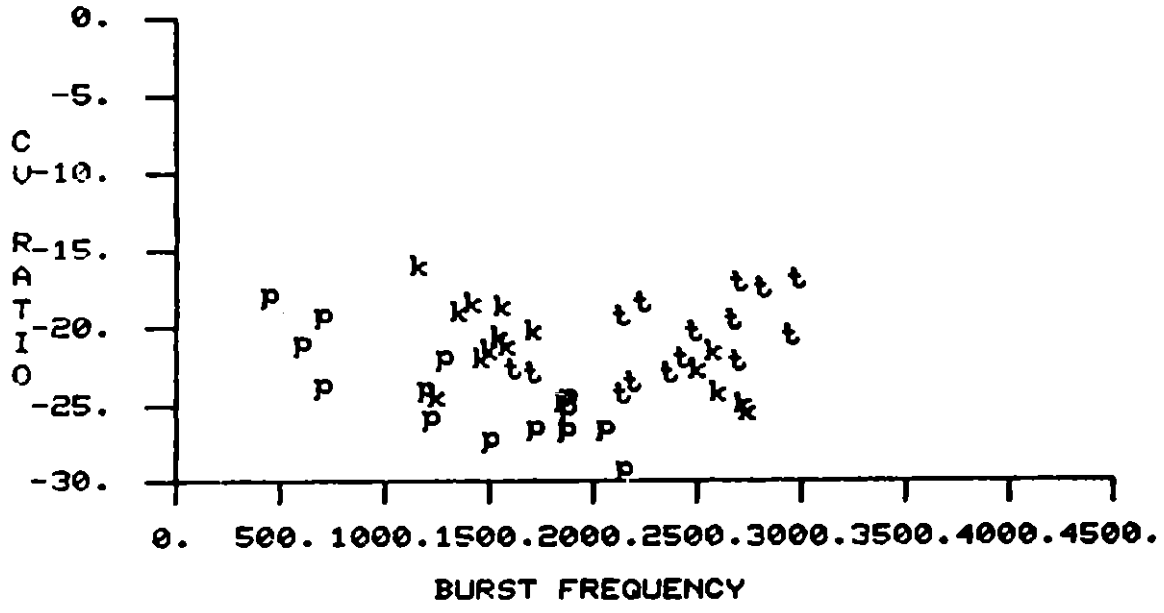


Figure 3.11b CV ratio vs burst frequency of voiceless consonants conversationally (left) and clearly (right) spoken by MS



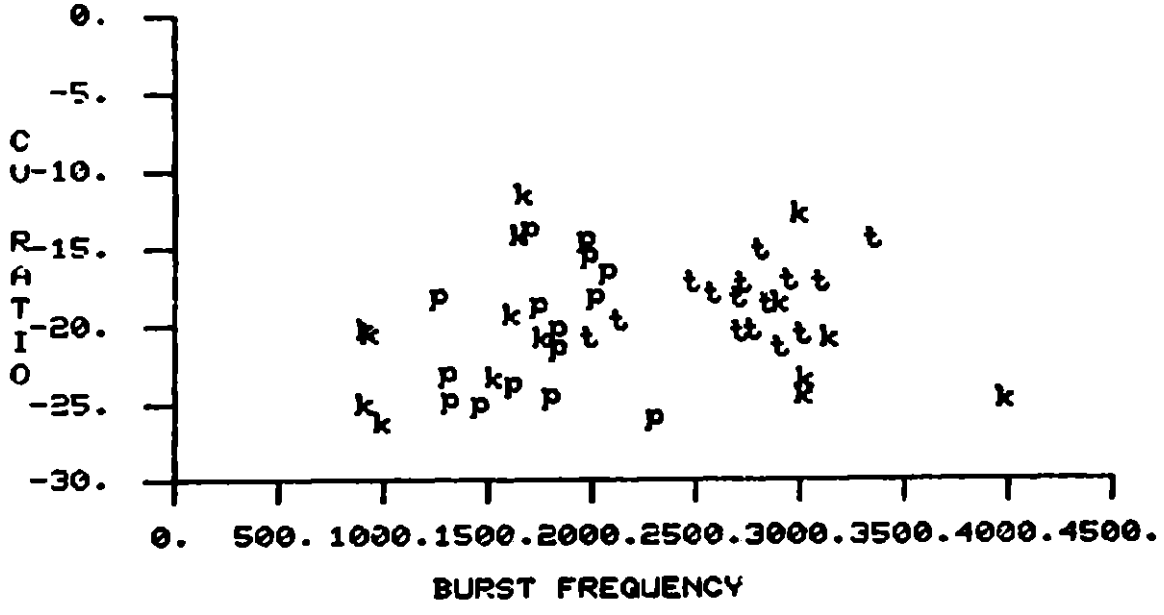


Figure 3.11c CV ratio vs burst frequency of voiceless consonants conversationally (left) and clearly (right) spoken by RR

for the vowel proceed during production of the consonants, the following vowel influences the consonant. Since /i/ is a front vowel, the constriction to produce /k/ is closer to the lips than when /a/ or /u/ (both of which are back vowels) follow. A constriction nearer to the lips decreases the length of the "tube" formed from the constriction to the opening, thereby increasing the burst frequency. All other frequencies are cancelled by the presense of zeros. The /p/'s may also be seen to cluster in two groups--those whose burst frequency is above 2 kHz and those whose is below 1.5 kHz. The /p/'s whose burst frequencies are above 2 kHz are again followed by an /i/. The /t/'s also roughly cluster in two groups; this time the dividing frequency is 3 kHz where /t/'s followed by an /a/ have a burst frequency above 3 kHz.

In MS's clear speech, the burst frequency of the /k/'s again cluster in two groups. The /ki/'s range from 2500 to 3400 Hz while the /ka's and /ku/'s range from 1100 to 1800. This distribution is basically the same as that of conversational speech. The /pi/'s are again separated from /pa/'s and /pu/'s. This time, however, the separation is less than 500 Hz, while it was previously more. The burst frequencies of /t/'s have moved more towards the higher values, although the maximum range value is still 3500 Hz. Also the values of /tu/ are now in the upper range with /ta/, where in conversational speech they were lower. The

overall distribution of CV ratios appears to be about the same, but close inspection reveals that the CV ratio of the /p/'s and /k/'s with high burst frequencies have increased, while those of /t/ have decreased.

The CV ratio vs burst frequency of speaker RR is shown in figure (3.11c). Once again, the burst frequencies of /k/ formed two groups: those above 2400 Hz and those below 1700 Hz. Clear cut distinctions concerning burst frequencies and the following vowel are not observed. The burst frequency of /p/'s followed by an /i/ are generally higher, although one token of /pu/ was even larger.

In RR's clear speech, the /k/'s show three distributions. The /ki/'s range in frequency from 2850 Hz to 3100 Hz, while the /ka/'s range from 1450 to 1750 Hz, and the /ku/'s range from 850 to 1000 Hz. The /t/'s once again seem to cluster from 2300 to 3300 Hz, except for two /ta/'s at 1951 and 2091 Hz. The burst frequncies of the /p/'s range from 1100 Hz to 2400 Hz with the bulk between 1500 to 2000 Hz. No clear distinctions of the burst frequency and following vowel were observed. The CV ratio was generally larger in clear speech than in conversational speech. Although the burst frequency of the /k/'s were separate, the CV ratio of each group varied widely. On the whole, the CV ratio of /ka/ and /ki/ each increased, while that of /ku/

actually decreased. In conversational speech, the CV ratio of /pi/ was in the lower section of the range of CV ratios of /p/, but in clear speech, the CV ratio of /pi/ is now in the upper section. The range of the CV ratios of /t/ is just about 2 dB higher.

As shown in Fig. 3.11a, the overall range of burst frequencies and CV ratios of speaker JL was about the same in clear and conversational speech. Once again, the /k/'s exhibited two distinct distributions, with the /ki/'s having a higher burst frequency. In the clear speech, a /ku/ lies between both distributions. In conversational speech, the /t/'s are relatively well clustered in the upper frequencies and relatively high CV ratios. The /p/'s are in the lower frequency range, non-overlapping with the /t/'s, and also exhibit a lower CV ratio.

In clear speech, the /t/'s are still in the upper frequency range and the /p/'s are still in the lower frequency range, but a group of /tu/ and /ti/'s have a lower burst frequency which is in the region of some /p/'s. The burst frequencies of /k/ overlap less with other consonants in clear speech than in conversational speech. The clustering of /k/'s along the frequency axis also is tighter.

4. INTELLIGIBILITY TESTS AND RESULTS

Two types of intelligibility tests were conducted on the clear and conversational speech: detection threshold tests and identification tests. From the identification experiments, the intelligibility of the two styles of speech were evaluated, and from the detections experiments, the relative thresholds of the clear and conversational styles of speech were determined. Evaluation of the relative loudness of the tokens was desirable because the normalization method was not a standardized procedure.

4.1 NORMALIZATION AND INTELLIGIBILITY

The "loudness" of a speech signal affects its intelligibility, especially when degraded by noise. Since the speakers tended to say tokens at different levels, some type of normalization process to equalize the loudness was needed. Presently there are no standard methods of normalization, although many options in normalization methods do exist. Each type of normalization has its advantages and disadvantages. For example, normalizing with respect to maximum vowel amplitude may destroy natural cues of different loudness levels between vowels. With regard to this work, however, it is felt that differences in intelligibility which would be large enough to be of

interest should still be apparent if a reasonable type of normalization is employed.

The method of normalization was chosen from among several methods tried. The methods of normalization investigated ranged from scaling each token so that the maximum absolute value of the samples was represented as the maximun value on the D/A (a window of 1 sample or 0.1 msec) to scaling each waveform so that the average rms power was a predetermined value (a window the length of the waveform). Windows of 25.6 and 102.4 msec and normalization only to the extent of setting gains when recording and digitizing the speech were also investigated. Three listeners informally listened to a sampling of tokens, each of which were "normalized" by the five methods previously mentioned. For each normalization method, they rated the variation in loudness within the group of tokens without knowing by which method each group was normalized. All methods (other than "minimal normalization") seemed to be about the same in their effectiveness. Therefore, the averaging of power was employed because it would better preserve the relative intensities of vowels among each other. A smaller window would tend to normalize the vowels to the same amplitude since the amplitude of vowels are usually greater than the amplitude of consonants.

4.2 INTELLIGIBILITY TESTING

In both intelligibility experiments, the same four normal-hearing listeners were used. These listeners were different from the listener used to elicit the clear speech. All subjects underwent a routine audiological examination to check that their hearing was normal. Experimental stimuli and masking noise were presented to each subject's right ear through headphones, while the left ear did not receive any sound. Since the experiments were conducted in a sound proof room, extraneous noise which would disturb the subject's concentration was minimal, making the use of masking noise to the left ear unnecessary.

All 540 segments of speech waveforms which were processed as described in Section 2.4 were used in both intelligibility experiments. The tokens were stored on a disk accessed by the CBG's PDP-11 computer.

4.2.1 Detection Experiments

One practice and two data collecting sessions of detection experiments were conducted. The experiments consisted of five trials of each of three sets of stimuli (30 trials total). The three sets of stimuli were: 1) all tokens of conversational speech 2) all tokens of clear

speech and 3) one phrase conversationally spoken by MS, /ə'gup ə/, and continuously repeated. The third set was included as a control against which the variance of the other two styles of mixed phrases could be compared. Throughout the detection measurements, the level of the 20 kHz bandwidth "white" masking noise was maintained at 65 dB SPL. The level of the stimuli was varied in 1 dB increments by the subject to determine the detection threshold level. The experimental setup is shown in Fig. 4.1. The PDP-11 continuously played out the stimuli with a short pause between each token. The experimenter used her attenuator to attenuate the stimulus signal to a level unknown to the subject. The subject then determined the upper and lower detection threshold for each trial as follows: she began at a level on her attenuator such that she could not detect the signal and slowly decreased the amount of attenuation (increased the level of the signal) by one dB increments until she could just detect the signal. This level was recorded as the upper detection level. The subject then decreased the attenuation 10 to 15 dB more so that she could easily hear the signal, and then slowly increased the amount of attenuation until she just could not detect the signal; this level was recorded as the lower detection level.

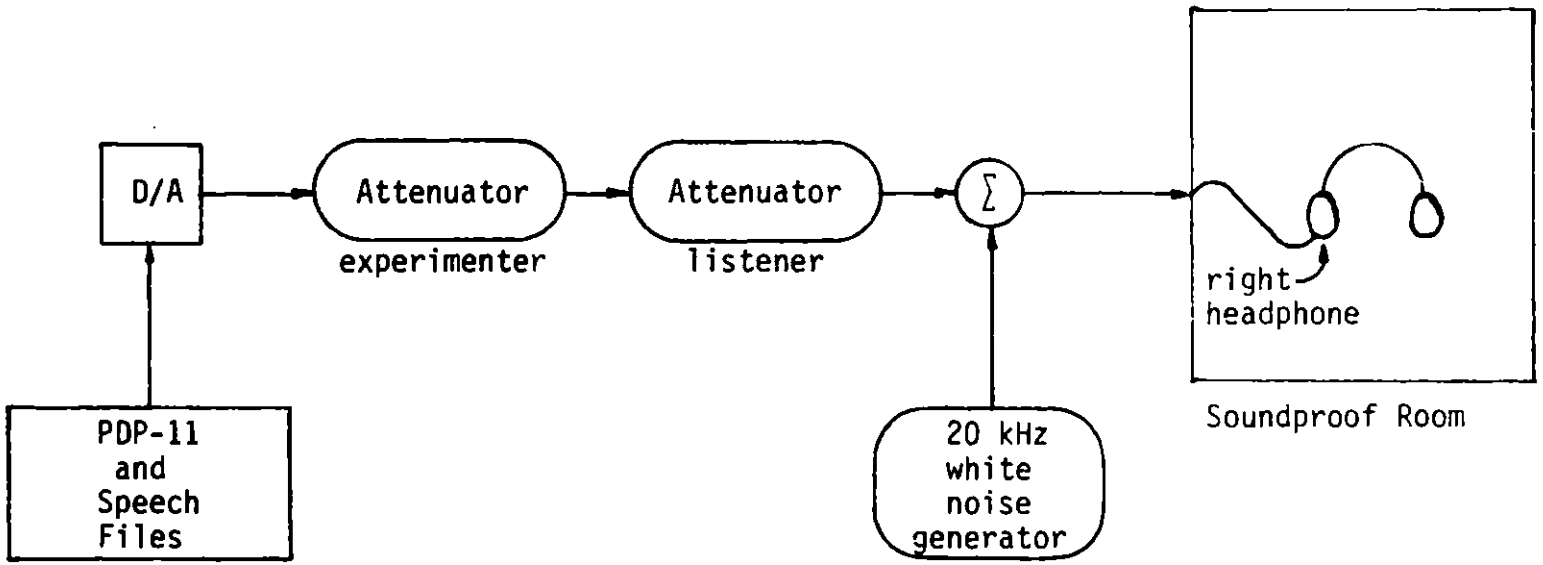


Figure 4.1 Equipment diagram for threshold detection experiment

4.2.2 Identification Experiments

The identification experiments were organized into 20 runs in which data was collected plus 4 practice runs. A run consisted of the presentation and identification of all 270 conversational tokens or of all 270 clear tokens. The clear and conversational tokens were not mixed in testing because the two types of speech are not naturally mixed. Table 4.1 shows the order of the runs. Prior to the data collection runs, four practice runs consisting of two signal-to-noise ratios not used in the data collection runs were carried out. Each SNR was run once with all conversational speech tokens and once with all clear speech tokens. In the practice runs, the listeners became acquainted with the experimental setup and speech material, thereby minimizing errors due to unfamiliarity with the experiment. Training was not employed because the relative intelligibility of the two styles of speech were of interest. Feedback as to the correctness of their responses was not given to the subjects at any time during the experiment. For each SNR, a conversational run was presented first, and then a clear run of the same SNR was presented in the same session. From Table 4.1, it can be observed that the SNR's were presented symmetrically: they began at the largest SNR, decreased to the smallest, and then increased back to the largest. This symmetrical

TABLE 4.1
SCHEDULE OF IDENTIFICATION RUNS*

SESSION / RUN	1	2	3	4	5	6	7	8
1	11 O	-12 L	∞ O	-4 L	-15 O	-21 L	-10 O	∞ L
2	11 L	-12 O	∞ L	-4 O	-15 L	-21 O	-10 L	∞ O
3				-10 L	-21 O	-15 L	-4 O	
4				-10 O	-21 L	-15 O	-4 L	

* O: CONVERSATIONAL
L: CLEAR
NUMBERS REPRESENT SNR OF RUN

presentation balanced the possible effects of learning on the results.

During each trial of the identification experiment, the PDP-11 was used to randomly select without replacement a token from a list. The list consisted of either the 270 clear or 270 conversational tokens, depending on the type of run. The selected token was then played out through the listener's right headphone (see Fig. 4.2). The listener responded by typing on the computer terminal the name of the CV from the 18 possible CV combinations listed near her. The PDP-11 accepted only the 18 CV's as a response; if something else was typed, it would request another response from the listener. The acceptable response to each stimulus was recorded by the computer for later use in analysis.

4.3 INTELLIGIBILITY RESULTS

The results from the detection threshold tests furnished data on the relative thresholds of each style of speech. Results from the identification tests furnished data on the overall intelligibility of the CV's, consonants, and vowels with respect to speaker and SNR. In this study, SNR is defined as the power (in dB relative to 1 V) to which the tokens were normalized minus the rms energy (in dB relative to 1V) of the added 20 kHz white noise. It should

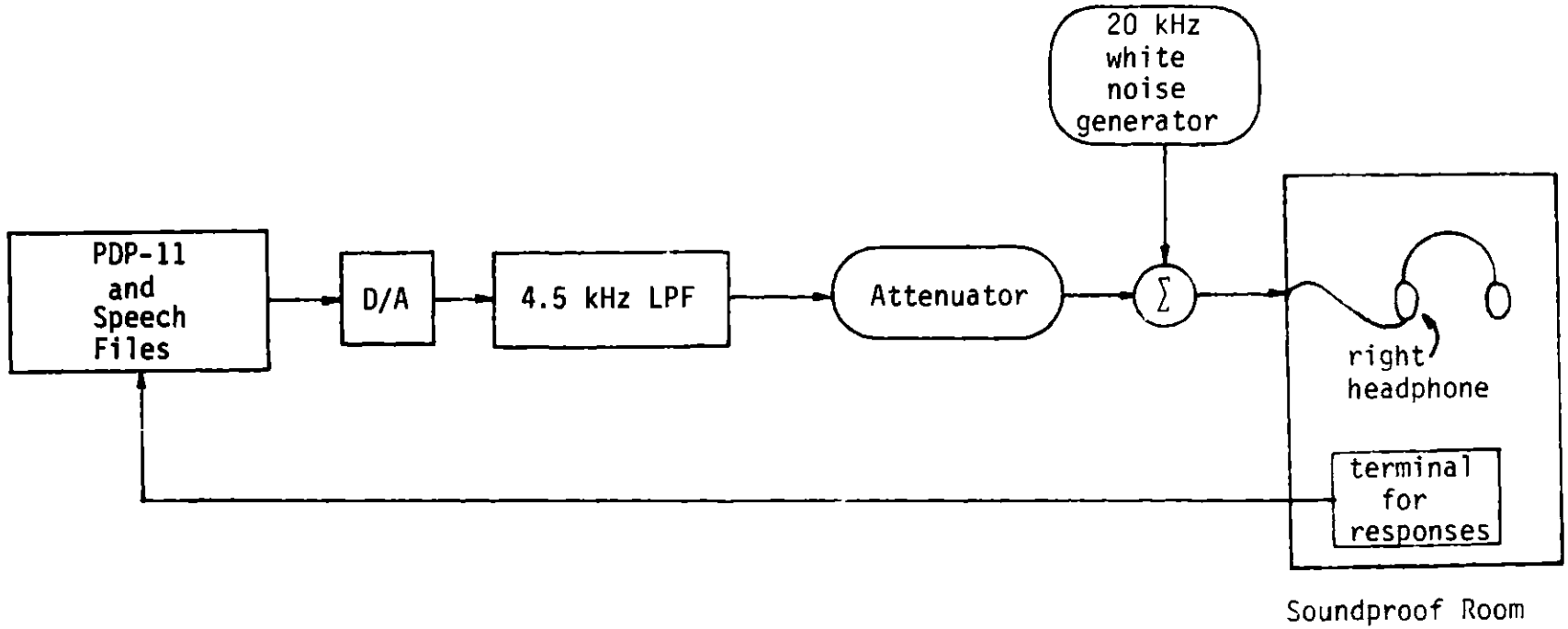


Figure 4.2 Equipment diagram for identification experiment

be noted that the speech was bandlimited to 5 kHz and the headphones used in the intelligibility tests bandlimited the overall signal to about 10 kHz. In the intelligibility tests, each speaker was analyzed separately because differences in the relationship between the intelligibility of clear speech and conversational speech was apparent between speakers. In addition, analysis by speaker allowed correlation between clear and conversational speech for each speaker.

4.3.1 DETECTION RESULTS

Detection measurements showed that the lower detection threshold, which generally had a smaller standard deviation than the upper detection threshold was an average of 2.03 dB higher for the clear speech than for conversational speech. For each style of speech, Table 4.2 shows the means and variances for each listener from the 10 trials of detection measurements. The standard deviationn for the /ə'gup ə/ phrase was less than that of the clear and conversational speech. This may have been caused by variations in level within each speech style and/or the use of different tokens.

TABLE 4.2

DETECTION LEVEL RESULTS*

LISTENER	CONVERSATIONAL		CLEAR		/ə'gup ə/	
	UPPER	LOWER	UPPER	LOWER	UPPER	LOWER
FC	37/1.8	40/2.4	38/1.6	40/1.8	39/0.8	42/0.7
DD	33/1.6	38/1.5	36/3.2	40/1.4	35/1.7	40/1.2
SM	34/1.3	35/1.5	35/1.6	39/1.3	35/1.3	38/1.1
RU	35/2.2	38/1.5	37/1.2	39/0.7	38/1.2	41/1.3

* RESULTS REPORTED IN THE FORM: AVERAGE/STANDARD DEVIATION

4.3.2 IDENTIFICATION RESULTS

4.3.2.1 Overall Intelligibility

Identification-type intelligibility tests indicated that overall, the clear speech was more intelligible than the conversational speech. But examining the speakers separately over the 5 SNR's shows (Fig. 4.3) that the intelligibility of the clear speech of speaker MS ranged from 15 to 38 percent greater and JL's speech ranged from 9 to 31 percent greater than the intelligibility of their conversational speech, while the intelligibility of speaker RR was approximately the same in both styles. The smaller difference in intelligibility at SNR's of -21, -4, and ∞ (no noise) were probably due to bottoming out effects. That is, at SNRs of -4 and ∞, the clear speech of JL exhibits very high intelligibility scores approaching the saturation level of 100%. At the SNR of -21, the conversational speech has saturated at about "chance", where "chance" is at a level where the subject can still hear the stimulus. A priori, chance should be equal to 1/18, or about 6%, since 18 CV's were used. However, if other factors (such as the intelligibility of vowels being high and voiced/voiceless distinctions of consonants generally being perceived) are assumed to be always correctly perceived, then this puts and upper limit of 33% correct as the chance value. Therefore,

if "chance" is limited to a range of 6 to 33% correct, then the intelligibility results at the lower SNR's are in the correct range to be approaching chance. Although no statistical tests were performed, the variation of listener responses was small. Therefore, the responses of all the listeners were grouped together.

Accounting for the differences in threshold of clear and conversational speech (by shifting the clear speech curve to the right 2.03 dB or the conversational speech curve to the left by 2.03 dB in Fig. 4.3) indicates that speakers JL and MS are still more intelligible, overall, while the intelligibility of the clear speech of RR became less than the intelligiblity of his conversational speech at higher SNR's. The difference in thresholds were computed by averaging the lower detection threshold across all listeners for each style of speech.

Analysis of the intelligibility tests by examining the intelligibility of the consonants and vowels separately allows a more detailed examination of perceptual errors. Stop consonants may be further categorized as to whether voicing is present and as to their place of articulation. The voiced stop consonants are b, d, and g; the unvoiced stop consonants are p, t, and k. Stop consonants are produced by constricting the vocal tract at one of three

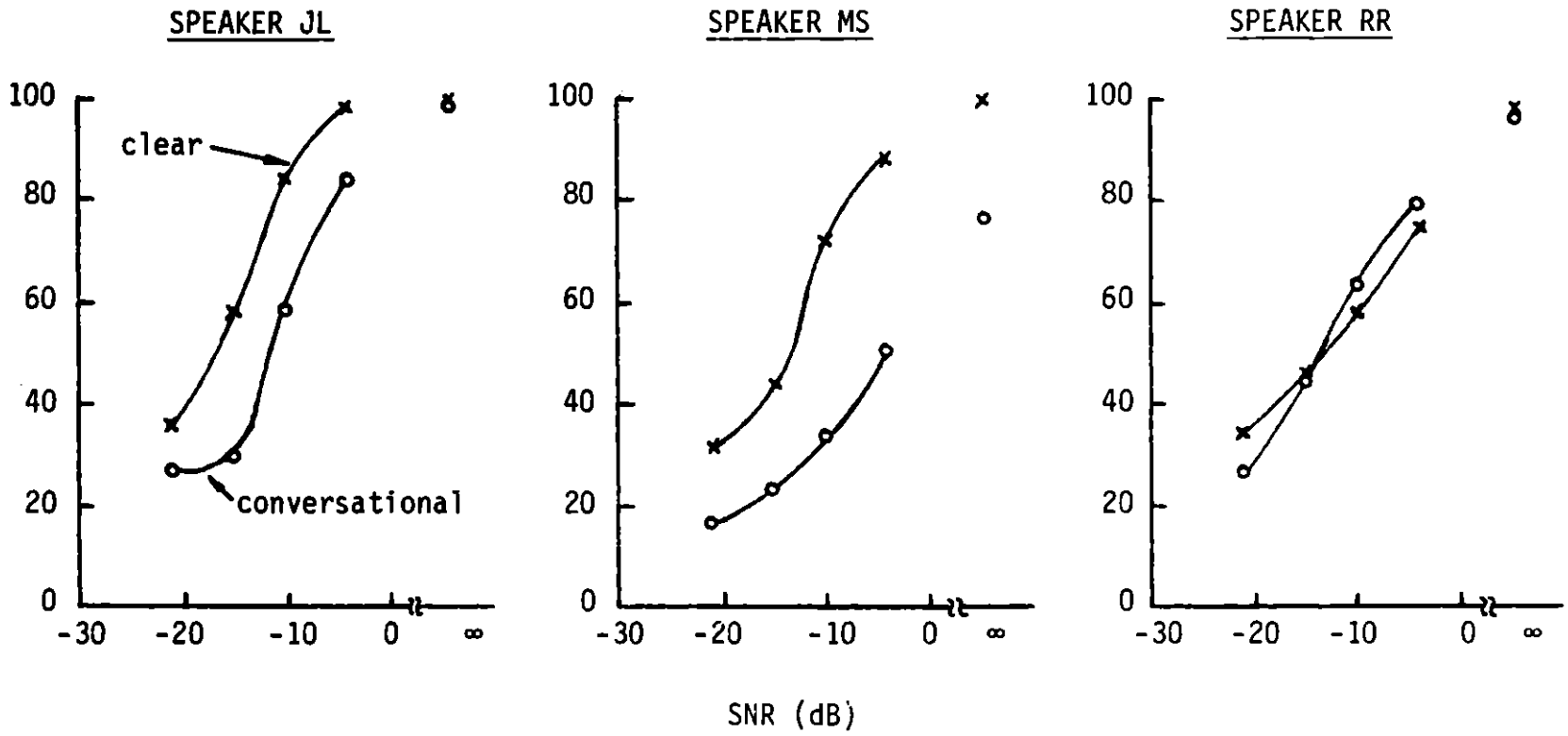


Figure 4.3 CV identification performance vs SNR

locations: the front of the mouth (labials), behind the teeth (alveolars) or near the velum (velars). In stops, the labials are b and p, the alveolars are t and d, and the velars are k and g. The changes in intelligibility of these classes between clear and conversational speech will be examined in the next section.

4.3.2.2 Consonant Intelligibility

The intelligibilty of the consonants at various SNR's followed the same general trends among speakers as in the overall intelligibility results (Fig. 4.4) Again, RR exhibited no significant difference in intelligibility between clear and conversational speech, speaker MS displayed the most dramatic difference, and speaker JL exhibited a higher intelligibility at each SNR for clear speech over conversational speech, but not as great as MS did. An upper saturation of consonant intelligibility was approached at an SNR of -4 for the clear speech and a lower saturation at -21 for the conversational speech where the percent correct was about 1/3. Since six stop consonants were used, a priori, a value of 1/6 shold be approached. Shift of the conversational curve by +2.03 dB still results in the intelligibility of clearly spoken consonants being significantly more intelligible than conversationally spoken consonants.

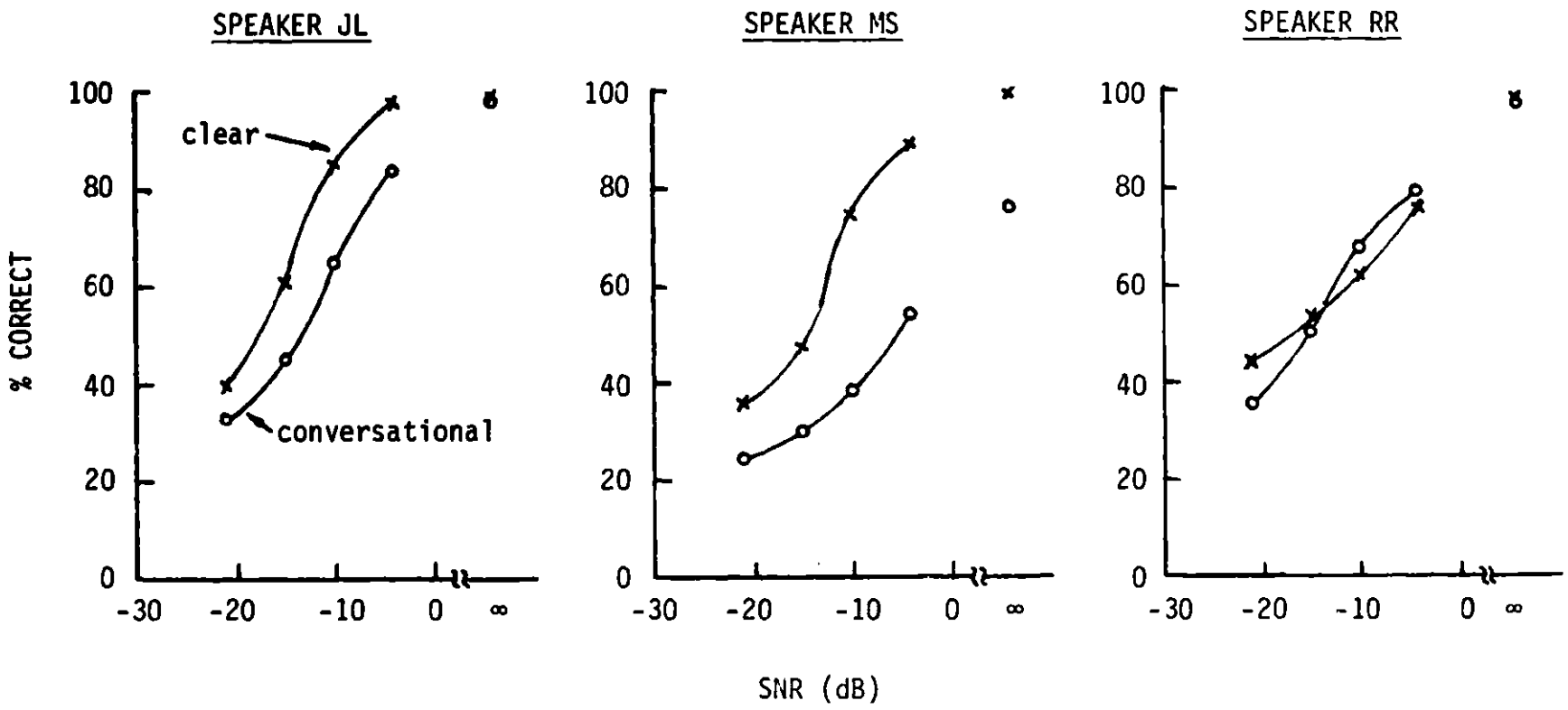


Figure 4.4 Consonant identification performance vs SNR

Without masking noise, the intelligibility of the clear speech of MS was almost perfect (approached 100%), but the intelligibility of his conversational speech under the no noise condition was only 76 percent. Speakers JL and RR showed almost perfect intelligibility under the no noise condition. The approximately same intelligibility of the 2 styles of JL's speech under no noise, but greater intelligibility of his clear speech compared to his conversational speech under noise conditions indicates that JL's clear speech is more resistant to noise degradation than his conversational speech. RR's clear speech is just about the same as his conversational speech. Examining the intelligibility results across the four listeners and three speakers indicates that the errors in perceiving conversational speech under no noise conditions are due mainly to voicing errors, as shown in the confusion matrix of Table 4.3. The label at the left side of each row refers to a stimulus and the label at the top of each column refers to a response for each cell. The number in each cell represents the percentage of times in which the labeled response was given upon presentation of the labeled stimulus.

Below each confusion matrix in Table 4.4, the average percent correct when perceiving voicing in each speech style is shown. The percent correct in perceiving voicing is

TABLE 4.3

CONSONANT CONFUSIONS

CONVERSATIONAL SPEECH: NO MASKING NOISE

	B	D	G	P	T	K
B	89	-	-	11	-	-
D	3	79	-	-	18	-
G	-	1	82	-	1	16
P	1	-	-	98	1	1
T	-	1	-	1	98	1
K	-	-	1	-	3	96

TABLE 4.4

VOICING CONFUSIONS

SPEAKER

CONVERSATIONAL

CLEAR

JL

	U*	V
U	93	7
V	5	95

94

U	V
92	8
4	96

94

MS

	U	V
U	90	10
V	49	51

71

U	V
93	7
5	95

94

RR

	U	V
U	86	14
V	3	97

92

U	V
98	2
2	98

98

* U: UNVOICED OR VOICELESS
V: VOICED

equal to 94% in both styles of speech of speaker JL. Voicing in MS's conversationl speech was perceived correctly 71% of the time, while in his clear speech, it was perceived correctly 94% of the time, an increase of 23%. RR exhibited only a slight increase of 6% from conversational to clear speech. RR and JL exhibited high intelligibility scores in both styles of speech while MS exhibited high intelligibility scores only in his clear speech.

More details on the identification performance of the stop consonants are summarized in confusion matrices, as shown in Tables 4.4, 4.5, and 4.6. Except in the conversational speech of MS, the intelligibilty scores from the no noise condition are all close to 100%. Therefore, the no noise condition was not included in tabulation to produce the matrices so that differences could be more easily observed. As shown in Table 4.4, in both styles of speech of speaker JL, there were very few voicing errors, and the number of errors between the two styles was comparable. The conversational speech cf speaker MS, on the other hand, exhibited a strong bias towards voiceless consonants (Table 4.4). This bias is present at all SNR's. The intelligibility of voicing in MS's clear speech was comparable to the intelligibility of the speech of JL.

TABLE 4.5

PLACE CONFUSIONS

SPEAKER	CONVERSATIONAL				CLEAR			
		LA*	AL	VE	LA	AL	VE	
JL	LA	74	18	8	80	12	8	
	AL	28	58	14	13	75	12	
	VE	23	28	49	14	21	65	
				60				73
		LA	AL	VE	LA	AL	VE	
MS	LA	66	20	14	67	19	14	
	AL	35	50	15	15	68	17	
	VE	35	36	29	18	28	54	
				48				63
		LA	AL	VE	LA	AL	VE	
RR	LA	82	12	6	78	13	9	
	AL	28	61	11	28	53	19	
	VE	25	30	45	28	26	46	
				63				59

* LA: LABIAL
AL: ALVEOLAR
VE: VELAR

TABLE 4.6

VOWEL CONFUSIONS

SPEAKER

CONVERSATIONAL

CLEAR

JL

	i	a	u
i	87	-	13
a	-	100	-
u	31	-	69

85

i	a	u
90	-	10
-	100	-
11	-	89

93

MS

	i	a	u
i	56	-	44
a	3	95	3
u	20	-	80

77

i	a	u
90	-	10
-	100	-
15	-	85

91

RR

	i	a	u
i	83	-	17
a	-	100	-
u	13	-	87

90

i	a	u
81	-	19
-	100	-
12	-	88

90

RR's speech exhibited yet another pattern: Table 4.4 shows a slight bias towards perceiving voiced consonants in his conversational speech, although the intelligibility of the consonants is relatively high. The confusion matrice on voicing in the clear speech of RR indicates that both voiced and voicless consonants are perceived equally well and their intelligibility is high.

The overall percentage of correctly perceiving place of articulation is shown below each confusion matrice of Table 4.5 Speakers JL and MS both show increases from conversational to clear speech of 13 and 15 percent, respectively. RR, however, exhibited a slight decrease of 4 percent. Confusion matrices on place of articulation are also shown in Table 4.5. A general trend of the labials being perceived the best and alveolars being perceived next best is observed for all speakers and styles except JL when speaking clearly. In this case, the labials and alveolars are perceived approximately equally well.

A general bias away from velars may also be observed. In JL and MS, the labial for alveolar and alveolar for velar confusions are about equal in conversational speech, but in clear speech the labial for alveolar confusions decrease by a much greater amount than the alveolar for velar confusions. (This trend was not observed in RR's results.)

The improvement in alveolar intelligibilty was due mainly to a decrease in labial for alveolar confusions, and the improvement in velar intelligibility was due mainly to a decrease in labial for velar confusions, and to a lesser extent, velar for alveolar confusions. In general, the perception of place is about equally good for clear and conversational speech for speaker RR. The degradation in transmission of voicing and place with decreasing SNR is illustrated in Fiss. 4.5 and 4.6. These results are in agreement with Miller and Nicely's (1955) data; it is seen that voicing is affected much less by masking than is place of articulation.

4.3.2.3 Vowel Intelligibility

The intelligibility of vowels is not degraded by noise as much as the intelligibility of consonants. Comparison of Figs. 4.4 (Section 4.3.2.2) and 4.7 illustrates this point and also illustrates that the clear speech is generally more intelligible for speakers MS and JL. Examination of vowel confusions (Table 4.6) shows that /a/ is seldom confused, and that the confusions exist mostly between /i/ and /u/. More specifically, in speaker JL, /u/ was confused as /i/ much more often than /i/ was confused as /u/ in conversational speech. In JL's clear speech, the /i/ for /u/ confusions decreased to the level of the /u/ for /i/

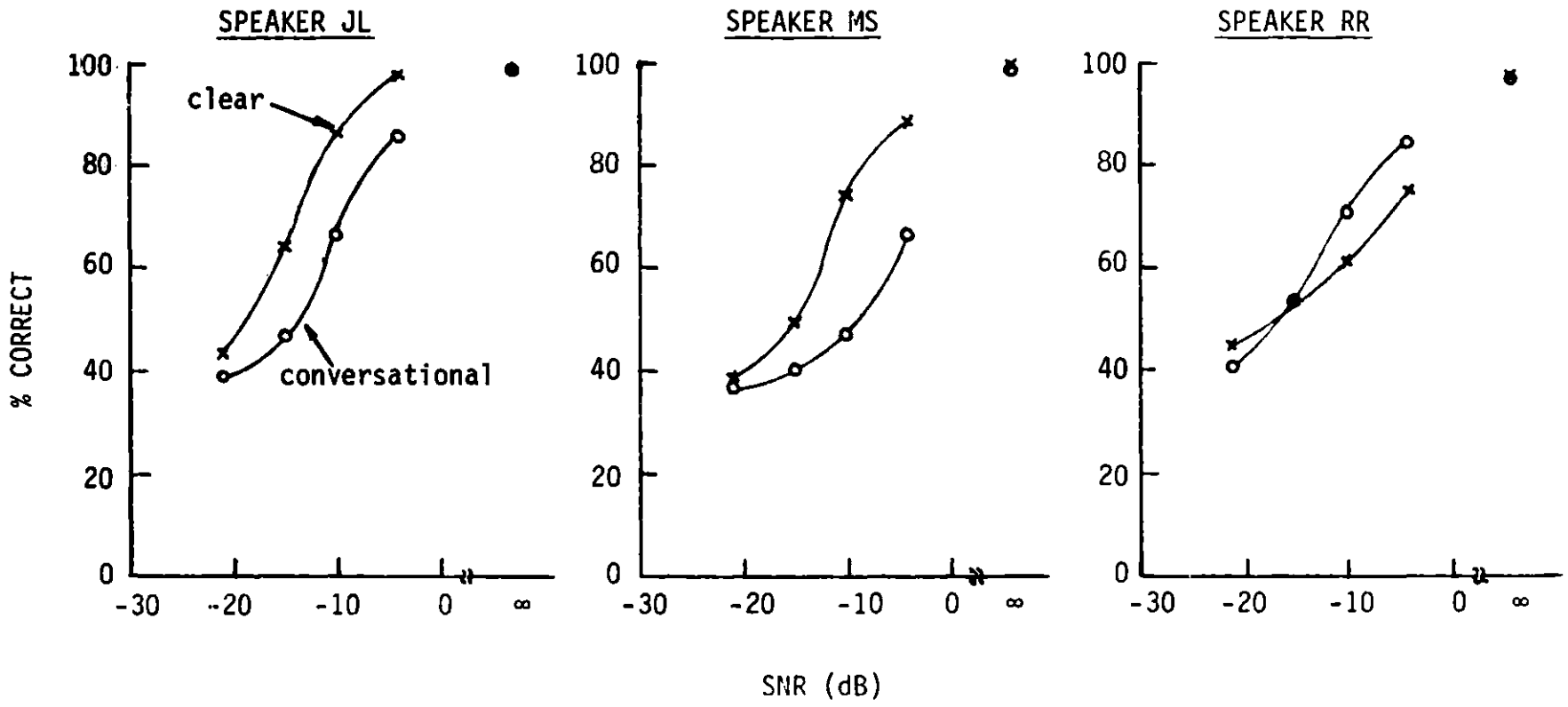


Figure 4.5 Place identification performance vs SNR

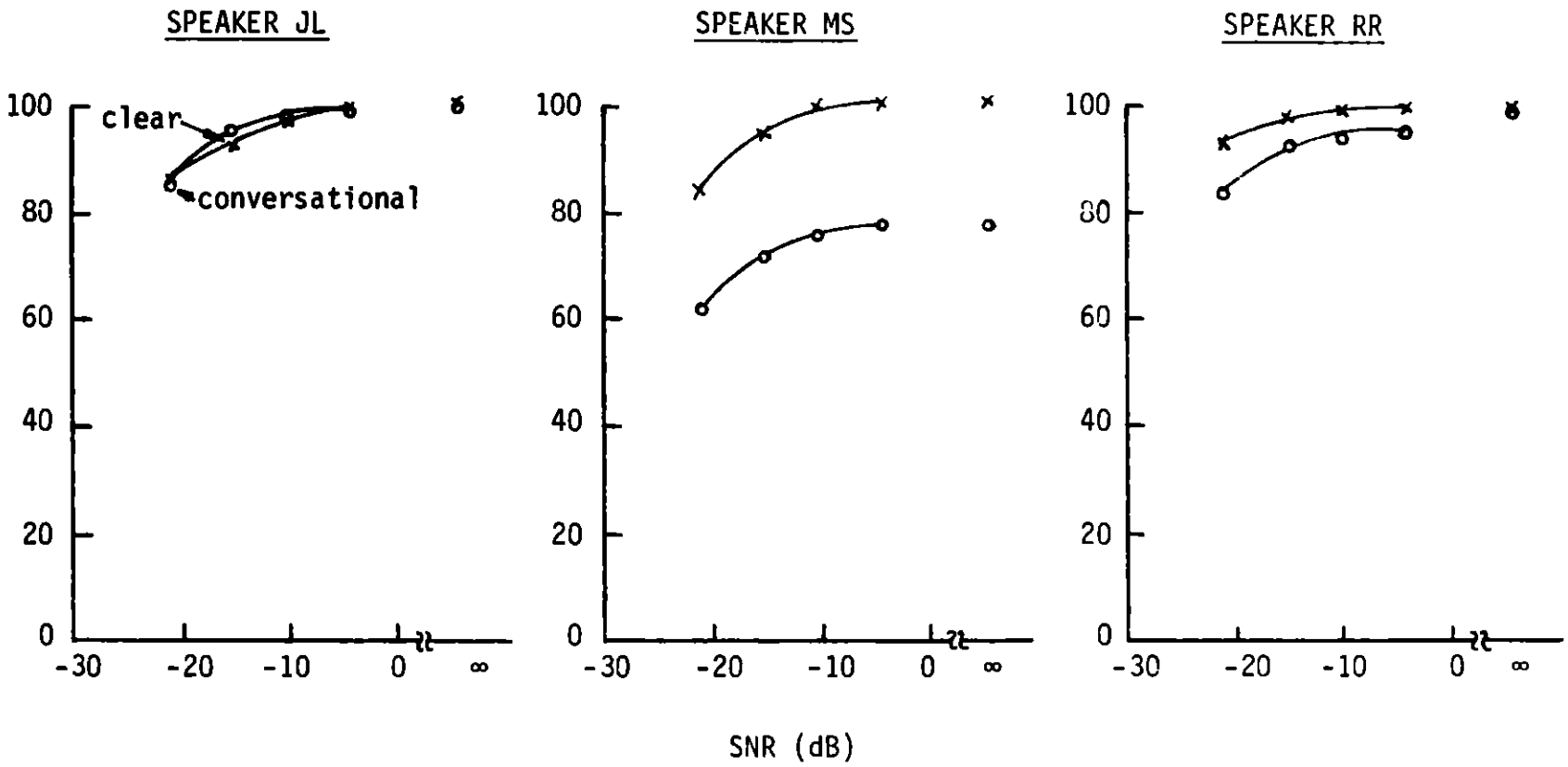


Figure 4.6 Voicing identification performance vs SNR

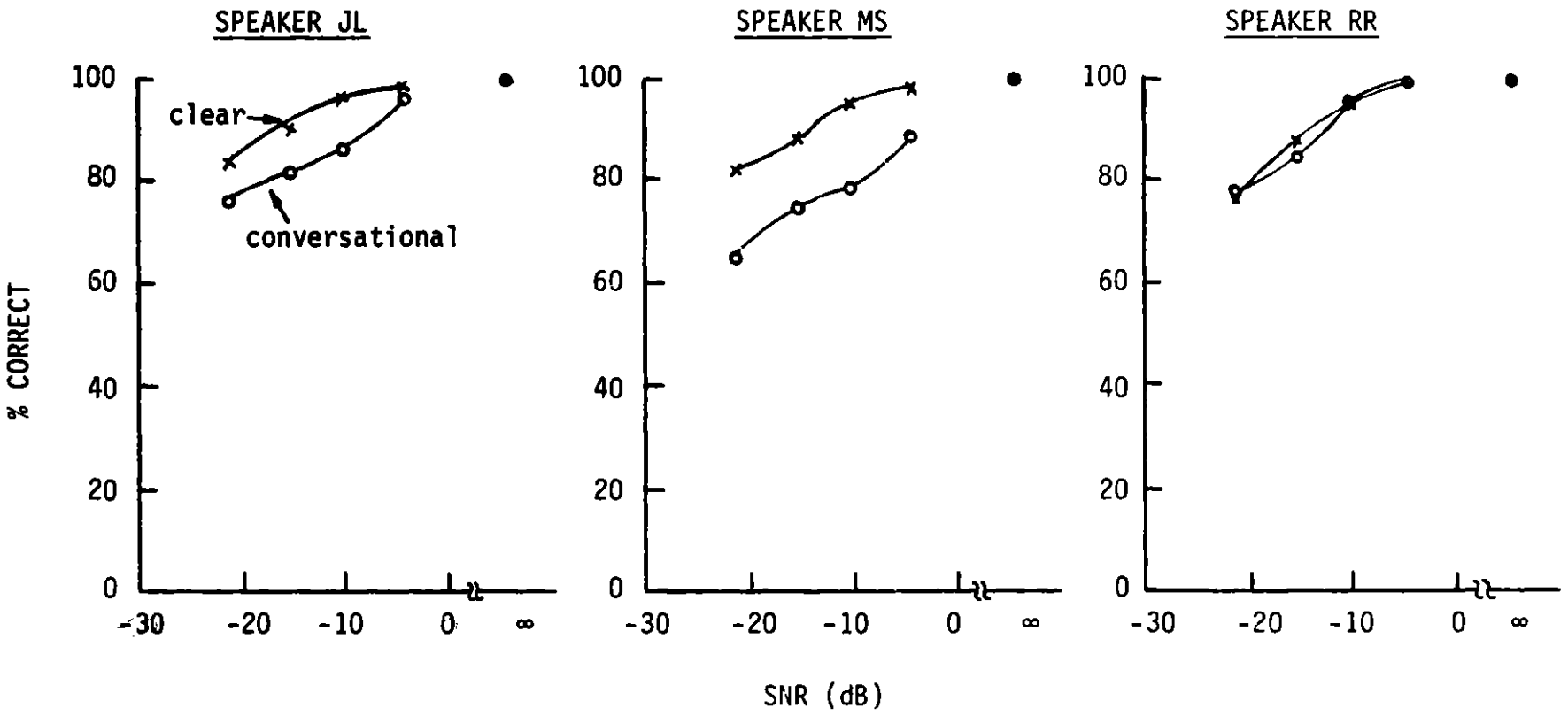


Figure 4.7 Vowel identification performance vs SNR

confusions, which resulted in an error rate for each of about 10%.

Speaker MS, contrary to JL's results, had a strong bias towards /u/ in his conversational speech. In clear speech, while /i/ for /u/ confusions decreased by 5 percentage points, /u/ for /i/ confusions decreased by 33 percentage points. The /i/'s, which were much less intelligible than /u/ in conversational speech (by 24 percentage points) were 4 percentage points greater in percent correct for clear speech. The speech of speaker RR did not show any significant change in the intelligibility of vowels between the 2 styles of speech. Both styles of speech also exhibited almost no confusions of /a/ but showed /i/ and /u/ confusions which were slightly biased towards /u/.

5. DISCUSSION

The results of the intelligibility tests and acoustic analysis indicate that there are significant differences between clear and conversational speech, both perceptually and acoustically. The ability to draw correlations between these acoustic differences and perceptual differences is a significant step towards attempting to process speech to sound more intelligible. In this discussion, perceptual differences and possible corresponding acoustic differences in consonants and in vowels will be examined. As previously mentioned, phonemes may be classified according to whether or not a particular feature is present. Changes in perception of these features will be examined against changes in acoustic parameters. Following this, more general acoustic differences and how they might contribute to the intelligibility of speech will be discussed.

5.1 CONSONANTS

Characteristic features of stop consonants are voicing and place of articulation. Acoustically, voiced consonants are distinguished from voiceless consonants by the presense of voicing and by a shorter voice onset time (VOT). From the results of this study, there seems to be good

correspondence between the perception of voicing and the VOT.

5.1.1 Voicing and VOT

Examination of the intelligibility of stop consonants showed that the speech of different speakers differed in intelligibility and were perceived with different types of errors. Comparison of the number of voicing errors and the VOT's for each style of speech by each speaker seems to indicate a general relation between the two. The correct perception of voicing in MS's conversational speech was much lower than that of the other speaker-style combinations. Histograms of the VOT distribution comparing voiced and voiceless stops from conversationally spoken speech by MS (confer Fig. 3.3b), show significant overlap of the distributions. The more intelligible speech from the other speaker-style combinations shows more distinct VOT distributions. This indicates that the distinctness of the distributions may be more important than the difference in the means in contributing to the correct perception of voicing. From this, one may suppose that the presence of voicing is determined as in a discrimination experiment where the subject uses an optimal criterion along a decision axis to determine the presence or absence of what is being tested. By moving the criterion along the decision axis

(experimentally, one method is to vary the payoff) and computing the probability of a correct detection versus the probability of a false alarm, an ROC curve may be traced out as in signal detection theory. In a similar manner, ROC curves were derived from the distributions of MS's and RR's VOT's from conversational speech. From these curves, it was observed that the difference between the curves corresponded well with the observed difference in intelligibility results.

As previously mentioned, the speech of speaker MS was perceived with many more voicing errors than his clear speech. From MS's confusion matrices on voicing in conversational speech, it was seen that the errors were biased towards perceiving voiceless consonants. Intuitively, one would expect a bias in conversational speech towards the voiced consonants because the VOT is shorter in conversational speech, and a short VOT is characteristic of voiced consonants. Another complexity also arises in that in clear speech the average VOT is increased. From this, one would expect the listener either to perceive most of the consonants as voiceless or to become confused when hearing the much longer than normal VOT's observed in clear speech. Instead, the listener is able to distinguish between voiced and voiceless consonants with higher accuracy than in conversational speech. More

specifically, in MS's speech, the maximum VOT of voiced consonants was observed to increase by less than a factor of two, while the minimum VOT of voiceless consonants was observed to increase by over a factor of two. If the perceived speech rate decreased by a factor of two and the voiced/voiceless criterion scaled accordingly, the VOT of voiced consonants would be perceived to be shorter and the VOT of voiceless consonants would be perceived to be longer, thus increasing the perceived difference between the features. From this, a hypothesis could be made that the listener takes into account the overall rate with which someone speaks and scales the criterion for voiced/voiceless distinctions accordingly. In CV's composed of voiced stops, the increase in syllable duration was greater than the increase in VOT and in CV's composed of voiceless stops, the increase in syllable duration was less than the increase in VOT. Therefore, the VOT of voiced stops was shorter relative to the syllable duration and the VOT of voiceless stops was longer relative to the syllable duration. This possibly would make the voiced/voiceless distinctions greater. Conversational speech has a more rapid pace than clear speech. Therefore, according to the hypothesis, the VOT criterion for voiced consonants would be low, thus possibly accounting for the bias towords perceiving voiceless consonants.

Speaker JL exhibited high intelligibility of voiced/voiceless distinctions (Fig. 3.3a). There was no significant difference between the clear and conversational speech. The voiced/voiceless VOT distributions of clear and conversational speech overlapped very little in both cases. This is in accord with the previous hypothesis that the amount of overlap, rather than the separation of the means of the distributions, is what is important in discrimination of voiced and voiceless stops. In the clear style, the VOT of both voiced and voiceless stops increased somewhat. The longer VOT of the voiced stops again did not result in decreased intelligibility scores in agreement with the scaled criterion hypothesis.

5.1.2 Place of Articulation

English stop consonants are articulated in one of three places. Each place is characterized by a constriction at a different position along the vocal tract. As mentioned in Section 3.3, the burst frequency is a function of the length of the front cavity of the vocal tract. From this, one might conclude that the burst frequency and place of articulation should be intimately related. Furthermore, one may also speculate that one would find changes in the burst frequencies when speaking more clearly. As mentioned in Section 3.3, burst frequencies were computed only for

voiceless stops. Because the window used in computation of the burst frequency exceeded the short VOT of voiced stops, some of the information from the transition is included. This makes it difficult to correlate changes in intelligibility with changes in burst frequencies.

The intelligibility results on consonants showed that speakers did not improve the intelligibility of all consonants equally. In the clear speech of speaker JL, the greatest increase in intelligibility was observed in /k/, the next greatest in /t/, and finally a small overall increase in intelligibility in /p/. In the clear speech of MS, the largest increase in intelligibility was observed again in /k/, then in /t/, and a very small increase in the intelligibility of /p/. In speaker RR, the only consonant that showed what seemed to be a significant increase in intelligibility was /k/, while the intelligibilty of /t/ actually decreased.

The changes in CV ratios and burst frequencies (Figs. 3.11a, 3.11b, and 3.11c) when comparing clear and conversational speech were not as definitive as the changes observed in some of the other acoustic parameters. The energy in a conversationally spoken /p/, normalized to the energy in the following vowel (CV ratio) was generally lower than the energy in other stops, but in the clear speech of

JL, the average CV ratio of /p/'s increased. This increase in CV ratio may have led to an increase in intelligibilty. The clustering of the two groups of /k/'s in JL's speech also became tighter and overlapped less with the lower frequency /i/'s. This tighter clustering may have increased the intelligibility of the /k/'s and the smaller overlap may have resulted in a decrease in /t/ for /k/ confusions. In clear speech, the average burst frequency of the /t/ and average CV ratio decreased, contrary to what would be expected since /t/ is a dental consonant which should have the highest burst frequency.

In the clear speech of MS, the average CV ratio of the /p/'s again increased. However, the burst frequency of the /p/'s also increased. Since /p/'s usually have a low burst frequency, this alone may have contributed to lower intelligibility. However, the increase in CV ratio may have led to an increase in intelligibility, as in JL's speech. It is seen that the burst frequency of the /t/'s has generally increased, which might have contributed to the higher intelligibility score. The burst frequencies of the two groups of /k/'s in clear speech are observed to move towards each other in clear speech. The increase of the burst frequency of /ka/ and /ku/ may make the two consonants more intelligible, and reduce the /p/ for /k/ confusions,

while the decrease in burst frequency of /ki/ would reduce the /t/ for /k/ confusions.

In the clear speech of RR, the burst frequencies of /p/ are observed to cluster more tightly in the region of 1300 to 2300 Hz. This corresponds to the /p/'s that have high burst frequencies in conversational speech. The clustering of /p/'s at higher frequencies and a larger average VC ratio in clear speech is similar to the trends observed in the speech of MS. Also, as with MS, the increase in intelligibility was small, indicating that a high /p/ burst frequency and larger CV ratio may have opposing effects on intelligiblity. The average CV ratio of the /t/'s is observed to increase, which should lead to an increase in intelligiblity. However, the burst frequency of the /t/'s decreased. Since /t/'s should have the highest burst frequency, this may lead to a decrease in intelligibility of the /t/'s. The /k/'s were observed to increase in CV ratio, which would lead to an increase in intelligibility. However, the /ku/'s were of a low CV ratio and had a low burst frequency, making the /k/'s more like a /p/. The intelligibility results are in agreement--/ku/'s are perceived as /ku/'s only 26% of the time, but are perceived as /pu/'s 61% of the time. The /ki/'s, which increased in burst frequency, but exhibited a wide range of CV ratios, were perceived correctly only 29% of the time. The errors

were more often in incorrectly perceiving a /pi/, but sometimes a /ti/ was perceived. The low CV ratio of some of the /ki/'s may account for this error. The high burst frequency characteristic of /t/'s would make one perceive a /t/, rather than a /k/. The /ka/'s that exhibited a burst frequency in the middle range were perceived the best of all /kV/ combinations.

In summary, strong differences in burst frequencies and CV ratios were not observed. However, more detailed examination of changes in these acoustic parameters with changes in intelligibility showed some correlation between the two. Movement of the burst frequencies towards their targets seemed to result in a higher intelligibility, as did an increase in CV ratio. A decrease in CV ratio seemed to result in perception of the consonant as a /p/.

5.2 VOWELS

Vowels are characterized by their resonant, or formant frequencies. From the intelligibility results, it was noted that the speech of MS exhibited the largest difference between clear speech and conversational speech (14 percent), JL the next largest difference of 8 percent, and RR showed no change. It was also noted that the intelligibility of the vowels in the clear speech of all three speakers was

similar. MS and JL, who exhibited lower intelligibility in conversational speech, both exhibited less tightly clustered vowels in F2 vs. F1 plots of conversational speech, while the other speech that exhibited high intelligibility (MS, RR, and JL's clear and RR's conversational speech) showed relatively tightly clustered vowels and a relatively open vowel triangle. JL's vowel triangle changed very little in shape, and the size was about the same for the two styles of speech. Speaker MS exhibited the most dramatic difference in terms of tighter clustering of vowels and a larger vowel triangle in clear speech.

These results suggest that tight clustering of the vowels (the vowels reach their formant targers more closely) and a larger vowel triangle (further separation between the vowels) may lead to greater ease in discrimination of the vowels. Closer examination of RR's F2 vs. F1 plots show some differences between the clear and conversational speech. That is, RR's formant frequencies moved away from each other in his clear speech, as MS's did, but the clustering of his vowels, especially /i/ and /u/ have decreased. These two changes oppose each other and their effects may have balanced out such that the intelligibility of the two styles of speech is about the same. Although the significance may be questionable, it may be observed that the intelligibility of the /i/'s was 2% less in clear speech

and that the /i/'s were less tightly clustered in clear than in conversational speech. Also the /u/'s were 1% more intelligible in clear speech and the /u/'s were more tightly clustered in clear than conversational speech. The intelligibility of the /a/'s was always high and did not change, although the clustering was less tight in clear speech. This may be due to /a/ being intrinsically louder and/or having a high first formant, distinguishing it from /i/ and /u/. Since white noise masks high frequencies more, the difference in second formant frequencies distinguishing /i/ and /u/ may be masked more than the high first formant of /a/ is masked. This could cause /i/-/u/ confusions and leave /a/ comparatively more intelligibile.

Comparison of the fine structure of MS's intelligibility results and F2 vs. F1 plots shows that for all vowels the intelligibility was larger and the clustering was tighter in clear speech. But because the amount by which the clustering became tighter is not proportional to the increase in intelligibility, at least one other factor which is important to intelligibility is suggested.

It is noted that /a/ is never confused for /i/ or /u/ except in MS's speech. In his speech, the F2 vs. F1 plots show that the first formant of /i/ and /u/ are below 400 Hz in both clear and conversational speech, which would

indicate that F1 is important in distinguishing /a/ from /i/ and /u/. That is, because /u/ and /i/ always have a low F1, they are never perceived as /a/. However, sometimes /a/ was incorrectly perceived as either /u/ or /i/. This may be because in some /a/'s, F1 was low and F2 ranged over the region of F2 for /u/ up to where /u/'s and /i/'s were confused. It was also noted that the second formant frequencies of /u/ and /i/ approached each other in conversational speech, which would lead to the /i/ and /u/ confusions. But from the F2 vs. F1 plots it is seen that the second formant of /u/ decreases more from conversational to clear speech than the second formant of /i/ increases from conversational to clear speech. From these plots, one might suppose that /u/'s would be incorrectly perceived as /i/'s more often than /i/'s would be incorrectly perceived as /u/'s. However, this is not the case; the confusion matrices on MS's vowels indicate that /i/'s were incorrectly perceived as /u/'s. This indicates that other factors also influence the intelligibility. Another way to characterize the vowels is by the difference between two formant frequencies, that is, by F3-F2 and F2-F1. From the F3-F2 vs F2-F1 plots of MS's speech, it is observed that in clear and conversational speech, the /u/'s are in about the same region, but generally closer towards the lower values of F3-F2, as are the /a/'s, but only more scattered. The /i/'s, however, show a larger difference between F1 and

F2--that is, they have moved away from the cluster of /a/'s and /u/'s. This difference in the first two formants may be what is important in perceiving /i/. The larger scatter of /a/ and /u/'s in conversational speech may also account for the lower observed intelligibility of these vowels. The difference frequency plots for all speakers show that /i/ is quite distinct from /a/ and /u/. In multiple conversational tokens from JL, F2-F1 for /u/ were relatively high and F3-F2 low so that the stray tokens were in the direction of /i/. This again contradicts the observed intelligibility results. A possible factor may be that white masking noise was used, thereby masking high frequencies more than low frequencies. The masker may have masked the second formant and since /i/ and /u/ are distinguished by the frequency of the second formant, /i/ and /u/ would sound similar, and more like /u/. Subjects did report that /i/ and /u/ sounded alike, especially at lower SNR's.

5.4 OTHER FINDINGS

Overall, the CV ratios did not appear to change significantly from clear to conversatinal speech, although those of the /p/'s appeared to increase and that of /t/ decrease in the clear speech of MS and JL. Therefore, contrary to previous findings, CV ratios did not always increase when one was attempting to speak more clearly.

The durations of all segments of the CV were found to increase from clear to conversational speech, but in a nonuniform manner. Speaking more slowly would, in general, allow the articulators to prepare better for each speech sound, and perceptually, would give a person more time in which to perceive each sound. Thus the longer period of silence preceeding the release of the burst in stop consonants would also allow for a greater buildup of pressure and greater burst, and the longer duration of the vowel would give the articulators more time to reach their target position so that the formant target values are better reached. The duration of formant transitions were also observed to increase. This increase may be due in part to more time taken to reach formant target frequencies with greater accuracy, but, perhaps more importantly, to give a listener more time to perceive the transition and perhaps to perceive better where the burst began.

From these results, it is seen that there are several changes in speech which may occur when people attempt to speak more clearly. One change is to slow down the speaking rate in a nonuniform manner. This was exhibited by all three speakers. A second change that occurred was an increase in the duration of the formant transition rate. Another change was exaggeration of the difference between the VOT of voiced and voiceless consonants. And yet another

difference was that the formant target frequencies were more closely approached.

6. CONCLUDING REMARKS

6.1 CONCLUSIONS

This thesis was an initial study attempting to correlate differences in intelligibilitiy with differences in segmental acoustic parameters of a subset of clear and conversational speech. One important part of this work was the development of a paradigm to elicit the clear speech. It was found that some speakers can improve the intelligiblity of nonsense syllables embedded in a carrier phrase using the paradigm. The one speaker who showed little increase in the intelligibility of his clear speech relative to his conversational speech had a high intelligibility score for his conversational speech. This suggests that trying to increase the intelligibility of already very intelligible conversational speech is likely to be unsuccessful.

To quantify the differences in intelligibility and types of perceptual errors observed in the two styles of speech, intelligibility tests were run. These tests showed that the intelligibility of the feature of voicing is more resistant to degradation by noise than is place, in agreement with Miller and Nicely's (1954) results. Vowels were also observed to be more resistant to noise than

consonants, in agreement with previous findings. The improvements in intelligibility were observed over a wide range of signal-to-noise ratios, which indicate that applications where speech needs to sound clearer is not limited by the signal-to-noise ratio.

Differences in acoustic parameters were observed between the clear and conversational speech. Some of the differences were well defined and seemed to correspond well to the observed intelligibility results. Others, although evident, did not seem to correspond directly to a particular perceptual feature. The strongest examples of correspondence were:

1) The voice onset time of voiceless consonants became significantly larger so that the distribution of VOT's of voiceless consonants became more distinct from that of voiced consonants.
2) In vowels, the formant frequencies were found to cluster more tightly in the clear speech, indicating that the formants reached their target values more closely. The vowel triangle was observed to become larger in the more intelligible speech--that is, the formants moved outwards, emphasizing the

characteristics of the vowel and creating further separation between each vowel.

6.2 SUGGESTIONS FOR FUTURE WORK

There are two basic directions for future work. One involves expansion of the work done thus far, and the other involves testing of the hypotheses on the various acoustic characteristics proposed as being significant to the intelligibility of a speech sound.

In terms of expansion based upon this work, there are several areas. Because of the loose testing of loudness normalization of the tokens, correlation between tokens that were usually perceived incorrectly and the acoustic parameters of those tokens could not be carried out with significant results. Further study to control the normalization procedure better would allow a more detailed analysis of the perceptual errors encountered and of the acoustic differences observed. Statistical analysis of listener responses in the identification tests was not carried out and should be completed. A study on the effect of various types of masking on the intelligibility of the two styles of speech would give additional information as to the frequency range in which certain acoustic parameters are significant. The use of hearing impaired subjects in

measuring the relative intelligibility of the two styles of speech would be a further step towards identifying possible applications involving the development of improved hearing aids.

The results comparing place of articulation with burst frequencies and CV ratios were not strong. Because comparisons were made at the CV level of the speech of each speaker, the number of samples from which trends were observed was small. Use of a larger number of samples and a smaller window that would allow computation of the burst frequency of voiced consonants should be carried out. In some cases, the measured formant transition durations were relatively short. Measurements on the slower formant transitions of liquids and glides would allow more accurate and easier comparison of this acoustic characteristic between clear and conversational speech.

In order to test hypotheses as to which acoustic characteristics are important to intelligibility, signal processing techniques to manipulate one acoustic parameter at a time need to be developed. If this could be done, changing each acoustic parameter of the conversational speech to the value observed in clear speech would allow determination of the contribution of that particular parameter to intelligibility. Later, "overshooting" the

change in an acoustic parameter in an attempt to produce even "clearer" speech could be conducted. It may be found that there is an amount of change which is optimal to produce the clearest speech.

Looking at future work that is not as directly related, one might now use the techniques developed to study changes in acoustic parameters of other phonemes. Hopefully, this would result in a general understanding of the relation between segmental features of phonemes and their intelligibility.

REFERENCES

Chen, F.R., "Acoustic Characteristics of Stop Consonants and Point Vowels in Fast, Normal and Clear Speech," unpublished term paper (1979).

Cooper, F.S., P.C. Delattre, A.M. Liberman, J.M. Borst, L.J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am. 24, 597-606 (1952).

Dreher, J.J. and J.J. O'Neill, "Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases," J. Acoust. Soc. Am. 29, 1320-1323 (1957).

Fairbanks, G. and M.S. Miron, "Effects of Vocal Effort upon the Consonant-Vowel Ratio within the Syllable," J. Acoust. Soc. Am. 29, 621-626 (1957).

Griffiths, J.D., "Rhyming Minimal Contrasts: A Simplified Diagnostic Articulation Test," J. Acoust. Soc. Am. 42, 236-241 (1967).

Hecker, M.H.L., "A Study of the Relationship between Consonant-Vowel Ratios and Speaker Intelligibility," unpublished Ph.D. Thesis, Stanford University (1974).

House, A.S., C.E. Williams, M.H.L. Hecker, and K.D. Kryter, "Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set," J. Acoust. Soc. Am. 37, 158-166 (1965).

Klatt, D.H., "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," J. of Speech and Hearing Research 18, 686-706 (1975).

Kreul, E.J., D.W. Bell, and J.C. Nixon, "Factors Affecting Speech Discrimination Test Difficulty," J. of Speech and Hearing Research 12, 281-287 (1969).

Miller, G.A. and P.E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am. 27, 338-352 (1955).

Peterson, G., I. Lehiste, "Duration of Syllable Nuclei in English," J. Acoust. Soc. Am. 32 693-703 (1960).

Picheny, M.A. and N.I. Durlach, "Speaking Clearly for the Hard of Hearing," Paper presented at the 97th meeting of the Acoustical Society of America, Cambridge, MA (1979).

Salmon, R.D., "Talker Variation as Related to Intelligibility," Tech. Rpt. 31, Communication Sciences Laboratory, Univ. of Florida, Gainesville, Fla. (1970).

Schwartz, R., "Acoustic-Phonetic Experiment Facility for the Study of Continuous Speech," ICASSP 1-4 (1976).

Tolhurst, G.C., "The Effects of an Instruction to be Intelligible upon a Speaker's Intelligibility, Sound Pressure Level, and Message Duration," Joint Proj. NM 011 104 500, Report No. 58, Pensacola, Fla: The Ohio State Univversity Research Foundation and U.S. Naval School of Aviation Medicine (1955).

Tolhurst, G.C., "Effects of Duration and Articulation Changes on Intelligibility, Word Reception and Listener Preference," J. of Speech and Hearing Disorders, 22, 328-333 (1957).

Webster, J.C. and R.G. Klumpp, "Effects of Ambient Noise and Nearby Talkers on a Face-to-Face Communication Task," J. Acoust. Soc. Am. 34, 936-941 (1962).

Woods, W., M. Bates, G. Brown, B. Bruce, C. Cook J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, and V. Zue, "Speech Understanding Systems," Final Technical Progress Report, BBN Report No. 3438 (1976).

Zue, V.W., "Acoustic Characteristics of Stop Consonants: A Controlled Study," unpublished Ph.D. Thesis, Massachusetts Institute of Technology (1976).

APPENDIX 1

Methods of Eliciting Clear and Conversational Speech

The carrier phrase and method of eliciting a particular style of speech can greatly influence how a CV is pronounced. Some carrier phrases and methods encourage one to speak conversationally, others make it easier to speak clearly. In this study, it was desirable to minimize differences in local and global phonetic environments so that the conversational and clear speech could be compared. Hence, the same carrier phrase was to be used in eliciting both conversational and clear speech. This increased the importance of the method for eliciting the two styles of speech, which in turn, increased the need for determining a more precise paradigm for eliciting each style of speech. Therefore, combinations of several types of carrier materials and speech elicitation methods were tried.

In order to investigate the effect of carriers, CV's embedded in paragraphs, CV's embedded in the carrier sentence, "Say /ə 'CVd/ again", and isolated CV's were examined. In addition, clear speech was elicited with each type of carrier material with either the instructions to speak clearly or the instructions to speak clearly with binaurally applied masking noise. The use of masking noise was probed since Dreher and O'Neill (1957) and Webster and

Klump (1962), among others, had suggested that people speak more intelligibily (clearly) when they are in noisy environments.

More specifically, the corpus in the probes consisted of the three types of carrier materials in three different styles recorded by six male speakers. Each type of carrier contained 11 CV combinations, and was repeated four times per CV combination. The first style of speech recorded was one which the speakers considered to be conversational. Then they spoke in a clear style after instructions to enunciate as clearly as possible, and finally they spoke in a clear style after instructions to speak clearly as before, but with masking noise applied to their ears. The carrier material which seemed to produce the most conversational type speech in the various attempted manners of speaking was the paragraph with embedded CV's. The isolated CV's seemed to produce the clearest type of speech in all attempted manners of speaking; and the carrier sentence was intermediate in all cases. Instructing a person to speak clearly resulted in noticeable changes in style of speech, the most striking being a lengthening of duration and an increase in volume. The addition of masking noise produced, in addition to the previous results, distortions in pitch and an even greater increase in volume. This made the use of masking noise applied to the speaker's ears undesirable.

Investigation of a method of eliciting conversational speech which required the subject to repeat a phrase multiple times to eliminate the "unfamiliarity" of the phrase did not prove to be promising. Speakers were asked to repeat in a normal, conversational manner the phrase "Say /ə'pip/ again" 100 times, pausing only momentarily between phrases. After the first few tokens during which the speech seemed to become more conversational, improvements plateaued. As repetition of the phrase continued, speech became singsong in manner. There also were periods of an abrupt switch in style from a more conversational style to the more formal type of the initial tokens, and then a change back to the more conversational styles again. The methods utilized in this thesis are outlined in Section 2.2, and proved to be the most optimal of the various methods investigated.

APPENDIX 2

AUTOMATIC VOLUME CONTROL

An automatic volume control (AVC) was used to normalize a speaker's volume before the listener in the clear speech paradigm heard it. It should be noted that the speech used in the intelligibility tests and acoustic measurements was not processed by the AVC. Tests were performed on the AVC and from them, the AVC was shown to normalize the energy in the speech signal without noticeably distorting formant frequencies. It had a flat frequency response from about 50 to 20,000 Hz, which is more than sufficient to cover the frequency range of speech. Although the circuitry controlling normalization of gain analyzed only up to 5000 Hz, most of the energy of stops and vowels are within this band. The AVC had an attack time of 5 msec, a variable release time, and a variable threshold below which the sound is not normalized, but is instead left at a gain of one to prevent room noise from being amplified during periods of silence. A relatively short release time was used in order to thwart the speaker from learning to make himself more intelligible by speaking the syllables preceeding the CV much more softly than the CV itself, and thereby allow the CV to be processed at the higher gain of the softly spoken syllables. The short release time also helped to minimize effects of carriers spoken at various volumes.