ACOUSTIC MEASUREMENTS FOR SPEAKER RECOGNITION

by

JARED JOHN WOLF

B.E.E., Union College
(1965)

S.M., Massachusetts Institute of Technology
(1967)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September, 1969

Signature of Author Jared John Wolf
Department of Electrical Engineering, August 18, 1969

Certified by Kenneth N. Stevens
Thesis Supervisor

Accepted by _____
Chairman, Departmental Committee on Graduate Students

ACOUSTIC MEASUREMENTS FOR SPEAKER RECOGNITION

by

JARED JOHN WOLF

Submitted to the Department of Electrical Engineering on August 18, 1969 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

This study is directed toward the improvement of speaker recognition techniques by developing an effective set of characterizing measurements to be made on the voice signal. Specific measurements are made on speech events that have been segmented and located in the utterance. In this study, these segments are located manually. The selection of these segments and the aspects of each to be measured are guided by acoustic and phonological theory and relations of vocal tract shapes and gestures to speech. The F-ratio of the analysis of variance is used to evaluate the speaker-separating ability of a measurement, and a technique is developed to evaluate the degree of dependence between pairs of measurements.

The measurements that were found to be useful were fundamental frequency, features of vowel and nasal spectra, estimation of the glottal source spectrum slope, word duration, fricative spectrum shape, and stop consonant prevoicing. The development of these measurements was facilitated by the use of a highly flexible digital computer laboratory facility designed for on-line speech research.

A speaker identification experiment was performed, using 17 measurements. A computationally simple linear classification procedure was used, and the test data were kept independent of the design data. No errors were made in identification of the speaker for 210 test "utterances" by 21 adult male speakers.

Thesis Supervisor: Kenneth N. Stevens

Title: Professor of Electrical Engineering

ACKNOWLEDGEMENTS

It is a pleasure to thank thesis supervisor Ken Stevens for his advice, criticism, and above all, encouragement during the evolution of this research. I would also like to thank my readers, Dennis Klatt and Jon Allen, for their interest and advice, particularly in the preparation of this thesis. Thanks also go to the members of the Speech Communications Group for their willing ears, suggestions, and encouragement. Special thanks go to Bill Henke for the use of his speech synthesis program and to fellow doctoral candidate Mark Medress for his suggestions and his comradeship over this last year.

I would also like to express my gratitude to the National Science Foundation for the Graduate Fellowship which made these last four years possible.

To my wife, Lynne, must go thanks for understanding why many nights had to be spent at M.I.T., for her unending encouragement, and for attempting to unravel my tangled syntax.

This research was performed in the Speech Communication Group of the Research Laboratory of Electronics, M.I.T. The research of that group is supported in part by Air Force Cambridge Research Laboratories under contract F19628-69-C-0044 and in part by National Institutes of Health Grant NB-04332. The major part of the Speech Communication computer facility was acquired under National Institutes of Health Grant GM-14940.

TABLE OF CONTENTS

Page

ILLUSTRATIONS AND TABLES

CHAPTER I

INTRODUCTION

The information contained in human speech includes much more than the words of the language that the speaker intends to transmit. Superimposed on the linguistic component, there is a socio-linguistic component, which can tell the listener about the general background of the speaker, and a personal component, which can give the listener information about the identity of the speaker (Ladefoged and Broadbent, 1957). We may also identify an emotional-expressive component, which reveals the emotional state of the speaker and his feelings about the message.

Recognizing the person from the sound of his voice is a common experience for anyone who uses the telephone or listens to the radio. To be sure, the context of immediate events and the content of what is said often contribute strongly to the identification, but such recognition also occurs in situations where there is no doubt that it was triggered by the acoustic signal alone. This ability of human listeners has been confirmed by experiments (Pollack, et al., 1954; Stevens, et al., 1968).

In this age of information processing, the question of characterizing and recognizing different voices is naturally of interest. It is conceivable that machine (computer) methods can surpass human performance by virtue of their capacity for data storage and rapid, detailed analysis. For

the business world, automatic speaker recognition could open new vistas of convenience services such as voice identification to supplant the credit card or to control access to a facility or to privileged information. It could also find application in security or law enforcement. Furthermore, research on automatic speech recognition shows that differences in speech signals due to different speakers greatly increase the difficulty of the recognition. Better understanding of these speaker differences could make compensation for them possible in such devices.

Differences in voices stem from two broad bases: organic and learned differences (Garvin and Ladefoged, 1963). Organic differences are the result of differences in the sizes and shapes of the components of the vocal tract: larynx, pharynx, tongue, teeth, and the oral and nasal cavities. Since the resonances of the vocal tract and the characteristics of the sound energy sources depend on just these anatomical factors, these differences lead to differences in fundamental frequency, laryngeal source spectrum, and formant frequencies and bandwidths. Learned differences are the result of differences in the patterns of coordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variations in the dynamics of the vocal tract such as the rate of formant transitions and coarticulation effects. Naturally many speaker-dependent properties are affected by both of these factors.

The problem of speaker recognition, like most problems in pattern recognition, may be considered to be divided into two parts: measurement and classification. In the first part, the pattern under test (a voice signal from an unknown speaker, in this case) is subjected to a number of measurements, resulting in a set of numbers which (ideally) characterize the pattern. These values in turn act as inputs to a classification scheme, which compares them with stored information on known reference patterns and makes a decision as to the class membership of the tested pattern. Descriptions of pattern recognition work by various authors place different degrees of emphasis on these two aspects, with the greater emphasis usually placed on classification. Specifically, in speaker recognition, the effort spent on the characterizing measurements does not seem to be consistent with either the effort spent on classification procedures or with the state of our understanding of the way speech is produced.

Well chosen measurements are important to pattern recognition problems in several respects. First of all, they must adequately characterize the patterns under test. No amount of decision-making sophistication can compensate for a basic lack of information. Furthermore, the amount of processing required in the classification phase is primarily determined by the complexity of the distributions underlying the measurement data. For example, optimum classification schemes of the linear type will fail if the classes cannot be described by

disjoint, convex, simply connected regions. It should also be noted that the more sophisticated classification schemes effectively estimate higher order properties of the underlying distributions and consequently require greater quantities of data to achieve statistical significance (Sebestyen, 1962). A well chosen set of measurements should permit the effective use of economical decision making procedures.

In the case of measurements which are adequately representative of, but only generally related to, the differences between speakers, it is effectively left to the classification process to separate the speaker-selective effects from irrelevant variations. More of this burden should be shifted from the classification process to the measurement process. This measurement phase should be selective and efficient rather than merely systematic and sufficient.

The aim of this study was to investigate and specify speaker-characterizing measurements which are both efficient in discriminating speakers and amenable to automatic measurement. These measurements are performed only on selected speech segments, rather than throughout an entire utterance, and each measurement is tailored to its speech segment. The selection of these speech segments and the aspects of each to be measured were motivated by acoustic and phonological theory and the relations of vocal tract shapes and gestures to speech. The development of these measurements was facilitated by the use of a highly flexible digital computer la-

boratory facility designed for on-line speech research.

Chapter 2 contains a review of the literature on automatic or machine methods of speaker recognition, with emphasis on the nature of the measurements used. In Chapter 3 general principles of measurement for speaker recognition and procedures for evaluating speaker-separating ability of individual measurements and intermeasurement dependence are discussed. In Chapter 4 the experimental procedures of this study are described. The specific speech segments and measurements investigated are described in detail in Chapter 5. Quantitative evaluations of these measurements and the results of a limited speaker identification experiment using these measurements are also presented.

CHAPTER 2

REVIEW OF LITERATURE

In order to evaluate the literature on automatic methods of speaker recognition, it is necessary to distinguish the possible tasks a speaker recognition system may perform. Let the term speaker recognition refer to the general problem of relating a voice signal to the person who uttered it. In speaker identification, the task is to classify the unknown voice signal as belonging to one of m speakers (closed set paradigm) or as belonging to one of m speakers or to some person outside that set (open set paradigm). An important special case of the open set paradigm is that in which m equals 1, which is called speaker verification or authentication.

Since past works in automatic speaker recognition differ not only in the form of the task, but also in the size of the speaker ensemble and in the restrictions imposed on the acoustic signal, comparison of the effectiveness of the systems in terms of error rate is possible only in a general way. In the present review, the measurements performed in these past efforts will be emphasized.

In one of the earliest works on automatic speaker recognition, Edie and Sebestyen (1962) proposed the automatic sampling of thirteen measurements during an entire utterance. These measurements were the first four formants, pitch period, envelope amplitude, the time derivatives of all of these, and

a parameter related to length of the current voiced interval. In a verification experiment, using only F_1, F_2, F_3, and F_4 and pitch period sampled at 5 points 0.2 seconds apart, error rates of 7 - 10% with one known and four unknown speakers were obtained. In an identification experiment, F_1, F_2, F_3, and F_4 were measured for 11 speakers and an average error of 43% was obtained. It is not clear whether the training and test data were taken from the same linguistic context.

A different approach to measurements was described by Pruzansky (1963). Intensity-frequency-time patterns were obtained by using a 17 channel filter bank covering 200-7000 Hz. The speech data were single common words extracted from context, spoken by 10 male and female speakers. Reference patterns for each talker and each word were formed by averaging three repetitions, and a separate repetition was used for testing. Classification was performed by correlating the unknown pattern with each reference pattern. Each word was tested separately, and an average error of 11% was obtained. If the patterns were averaged over time so that only frequency information remained, the error was also 11%. Averaging over frequency instead resulted in a much higher error.

Pruzansky and Mathews (1964) used the same data set described above, but considered each time-frequency "cell" as a separate measurement. In an experiment to determine the effect of the size of the cell in the time and frequency dimensions, the "quality" of each such cell for speaker

recognition purposes was evaluated for the data set by means of the F-ratio of the analysis of variance. This statistic is proportional to the ratio of the variance of the speaker means to the average speaker variance; the higher this value, the more distinguishable are the individual speaker distributions, on the average. More will be said about this statistic in Chapter 3. The measurements with the highest F-ratios were used, and performance leveled off drastically after only about 10% of the total number of measurements was used. Furthermore, performance generally increased as the time dimension of the individual cells increased, but decreased when the frequency dimension increased. This confirms the usefulness of the frequency-dependent information found in the previous study.

Becker, et al. (1964) also used this single-word data in evaluating methods of summarization and classification. By summing the frequency information across time and using a non-Euclidean distance metric, they obtained an average identification error of 3%.

In the single word matching schemes just described, time registration was accomplished by aligning the maximum energy points. Differences in total length were resolved simply by truncating to the length of the shortest example. It is thus likely that corresponding segments in other parts of the word were not in exact alignment in many cases. This fact may help to explain the improvements in performance as

the measurements were averaged over successively longer time intervals.

This problem was circumvented in the automatic speaker verification experiment by Carbonell, et al. (1965). The measurements in this experiment were spectra from a 13 channel filter bank covering 250-3000 Hz, sampled at three points in the word "baseball." The measurement points were defined in terms of the phonemes of the word so that they were made at an equivalent point in each word (e.g., 16 msec after the release of the first /b/). The classification scheme measured the Euclidean distance between pairs of normalized spectra in a 13-dimensional space. No numerical results were reported, but this measurement and simple classification scheme operating on the three measurement points was described as "encouraging."

The automatic speaker verification system of Meeker, et al. (1967) also made its measurements during certain specified phones. An automatic speech recognition system selected occurrences of /i, ɛ, u, ʌ/ from continuous speech. The time derivative of each output of a 19 channel filter bank was coarsely measured during each vowel selected, and essentially averaged over the occurrences of that vowel. Classification was accomplished by essentially calculating the Euclidean distance between the four averaged vowels and the reference data of the speaker to be verified. Using 11 male speakers and setting the false dismissal (rejection of the

correct speaker) probability to 1%, an average false alarm (acceptance of an incorrect speaker) probability of 5% was obtained.

The nasal consonant /n/ was the focus of an identification scheme reported by Glenn and Kleiner (1968). Spectra taken over the range 1.0-3.5 kHz during the middle of /n/ in initial, medial, and final positions and averaged over 10 examples were used as design and test measurements, and classification was accomplished by correlation. With a population of 30 male and female speakers, an error rate of 7% was obtained. When the unknown data was averaged over fewer utterances, the average error rate was much higher.

In a recent study, Atal (1968) investigated the significance of fundamental frequency contours for speaker recognition. The fundamental frequency was accurately measured during repetitions of a sentence of about 2 seconds duration by 10 female speakers. The pitch contours were smoothed and time scaled to the same length. Then a reduction of dimensionality using the Karhunen-Loeve transformation and a linear clustering transformation produced the final 10-dimensional data vectors. The classification procedure (identification) used the Euclidean distance between these transformed vectors and obtained an average error of 3%. When sentence duration was added to the data as an additional dimension, the error was reduced to 2%. Therefore, as one might expect, pitch information has also been shown to be a

relevant measurement for speaker recognition.

Over the short history of automatic speaker recognition, a shift can be seen in measurement strategy, from general measurements throughout an utterance to measurements performed on specific speech events. Long-term averages have not been commonly used, probably because there is much in the personal component of the speech signal that is inherently short-term. The use of nasal consonants is an excellent but isolated example of selecting a measurement location especially for its effectiveness in characterizing the speaker. The measurements described here do encompass aspects of the acoustic signal that are dependent on the structural and learned characteristics of individual speakers, but they do so only generally. For example, a 25 component spectrum of /n/ does reflect the size and shape of the speaker's nasal cavity, but the formant around 1 kHz is thought to be closely tied to the length of the nasal tract and would thus characterize the speaker much more directly. Efforts to relate measurements more directly to vocal tract structure and to specific articulatory events should result in increased effectiveness of the measurement phase.

CHAPTER 3

CHOOSING MEASUREMENTS FOR SPEAKER RECOGNITION

This chapter will be a discussion of general principles of choosing and evaluating measurements for speaker recognition, including quantitative measurement evaluations. However, the matter of the selection of the speech events on which the measurements were made will be deferred until Chapter 5.

The function of the measurement phase of a speaker recognition system is to perform a number of characterizing measurements on the voice pattern under test. Simply put, the speech characteristics measured should ideally:

- occur naturally and frequently in normal speech
- vary as much as possible among speakers, but be as consistent as possible for a given speaker
- not change over time or be affected by poor health
- not be affected by reasonable background noise or depend on specific transmission characteristics
- not be modifiable by conscious effort of the speaker, or at least, be unlikely to be affected by attempts to disguise the voice
- be easily measurable

Some of these constraints can be relaxed for most practical systems, but it is good to keep in mind the most generally useful qualities for speaker recognition measurements.

3.1 Segmentation and recognition

One class of measurement schemes that has been used in the past performs a set of measurements at 10-20 msec intervals throughout an entire utterance (Pruzansky, 1963; Pruzansky and Mathews, 1964; Becker, et al., 1964; Li, et al., 1966). There are three outstanding difficulties with this approach.

First, because of the normal detailed differences in timing of each utterance, corresponding articulatory events do not occur at exactly the same times, even if the utterances are registered at a particular point, such as the beginning or the energy peak. Therefore, comparisons of the measurements at points where the utterances are out of alignment are between somewhat different events. It may be argued that these misalignments are reflections of temporal patterns associated with learned characteristics of different speakers. This is indeed so, but in this form the temporal variations interfere with the comparisons of similar events. We need to separate these effects, taking account of the useful temporal patterns while also making comparisons between similar articulatory events.

Secondly, regular and rapid sampling of the voice signal with the characterizing measurements produces sets of data that have a high degree of redundancy. Reducing the sampling rate would only add problems, since it would increase the chances of missing significant speech events.

Finally, a given set of measurements is not optimally

suited to every segment of an utterance. For example, fundamental frequency is meaningless during voiceless intervals; low and mid-frequency formants have no significance during voiceless fricatives.

A selective and efficient approach to measurement is to perform some degree of segmentation and recognition of the linguistic component of the speech signal before the measurements proper. This is done in order to locate certain speech events of interest and then to make appropriate measurements at each of these points. Similar events can then be compared with a minimum of interference due to timing differences. Furthermore, the recognition of events and boundaries in the acoustic representation allows the separate measurement of relevant temporal patterns.

Segmentation of the acoustic signal is one of the knottier problems in speech recognition, but the general question need not concern us. In any application of speaker verification and probably in many instances of speaker identification, the use of a known linguistic context is a valid assumption. In this case, the necessary segmentation would not be difficult. In many instances, the system designer is even free to specify what utterance the speakers must say, so he may tailor the utterance both to contain an advantageous set of phonemes and to be easily segmented.

3.2 General principles

Spinrad (1963) has stated two criteria that a possible characterizing measurement must meet: not only must the measurement characterize the patterns to be recognized, but it must also be able to be performed effectively and correctly. This latter point is not just idle philosophy, but a question of practical significance. For example, the measurement of formant frequencies in cases where the formants are close, as in /a/, is often difficult. It is possible to make provision in the classification procedure for the atypical absence of a measurement, but since this represents a loss in information, in general it is desirable for such absences to occur as infrequently as possible.

If we wish to use these two criteria to evaluate measurements individually, we should add a third one: independence of the measurements. In general, we seek to avoid redundant measurements. If we know the measurements are independent, then we know that all of the measurements are contributing to the classification process. Efficiency of representing the speakers means that the required processing capacity and time is minimized. Furthermore, independence means that in the classification process, the measurements can be validly considered separately rather than jointly. All optimum classification schemes must effectively estimate the joint probability distribution over the measurements for each speaker. If the measurements are independent, the joint distribution is

simply the product of the individual distributions. Similarly, nonoptimal classification schemes must account for dependencies or suffer the loss in performance that results from ignoring them.

In selecting measurements we should be guided by acoustic and phonological theory and by the relations of vocal tract shapes and gestures to speech. There are several such considerations which have direct relevance to speaker recognition.

Measurements which relate mainly to structural differences should do so as directly as possible. The unique vocal tract of each individual is a fundamental basis of speaker recognition. Some people compare the structural basis of speaker identification to that of fingerprint identification (Kersta, 1962). Since the relation of the acoustic signal to anatomy is much less direct than that of the fingerprint, this argument is in general tenuous, but acoustic measurements which do find justification in terms of specific anatomical features should not only be effective ones, but they should have the effect of increasing confidence in the effectiveness and reliability of speaker recognition in general. Furthermore, the sources of variation are minimized if a measurement is related to a specific anatomical feature rather than to anatomy in only a general way.

The vocal tract displays different characteristics during different speech sounds. Rather than make general measurements, such as formant frequencies and their time derivatives,

on every speech segment, we should tailor the measurements to the specific speech event being measured.

There are certain acoustic correlates of the distinctive features in a given language that are significant in the production and perception of those features; they carry the linguistic information (Stevens, in press). Other acoustic attributes of the signal are then extra-linguistic; they do not enter into the process of transmitting the message. It is among the extra-linguistic attributes that we should look for possible speaker-selective characteristics. For example, in the vowel /a/, the essential acoustic feature is the closeness of the first and second formants, resulting in a broad concentration of energy in the neighborhood of 1 kHz. The details of the F_1-F_2 relationship, such as relative amplitude of the spectrum peaks, absolute frequencies, or, within limits, separation of the peaks, are probably not important for the perception of /a/, as evidenced by the fact that the productions of different individuals differ in these respects.

The phonology of the language imposes constraints on the combinations of phonemes which may form words of that language. If the presence of one phoneme constrains the possible phonemes which may follow it, then certain rules for the production of that phoneme may be optional; their function has already been performed by the constraints imposed by the first phoneme. For example, in English, if a stop follows a nasal in a final consonant cluster, it must have the same place of articulation

as the nasal. Words like bump and bunt occur in English, but bumt is not allowed. Hence a speaker does not have to be precise in his articulation of that stop. Variations in that articulation due to different speakers would be extra-linguistic.

Certain articulatory features or feature sets are not used for phonetic distinctions in some languages. These too are attributes which may vary between speakers. For example, in English, voicing during closure of a voiced stop is optional.

In addition to the general requirement that the measurements be efficient, there may also be requirements imposed by the specific implementation of the speaker recognition system. In speaker verification, it may be assumed that the speaker is cooperative, since he wishes to be identified. (The possibility of mimicry has not been extensively studied, but it will eventually have to be considered.) In speaker identification, however, we may not be able to assume that the speaker is not attempting to disguise his voice. This would mainly affect measurements which are derived from learned characteristics, such as dialect and intonation. It is also possible that some structural characteristics would be modified, through a distortion of the vocal tract such as rounding the lips or placing objects in the mouth. A serious effort to thwart voice disguises would have to include a study of the acoustic characteristics that would be affected.

The frequency characteristics of the transmission system that carries the voice signal may affect the choice of measurements. The telephone, for example, has a bandwidth restricted roughly to the range of 250-3000 Hz, so high-frequency spectral characteristics are not present and direct measurement of fundamental frequency is not possible. (Of course it is possible to reconstruct the fundamental by suitable processing.) In general, temporal patterns and measurements with the dimension of frequency, such as fundamental and formant frequencies, would be undistorted, but some sort of compensation for the effect of transmission characteristics would be required for measurements of spectrum amplitudes.

3.3 Evaluating candidate measurements

Given a number of possible measurements which have been selected with as much attention as possible to a priori considerations such as those outlined above, how can we evaluate the suitability of each measurement to the speaker recognition problem, in order to know which ones to keep and which ones to discard? Unfortunately, there is no objective way of evaluating a measurement by itself. The ultimate utility of a measurement depends upon the nature of the classification system that follows it. Only after a classifier has been coupled to the measurement system can such meaningful measures as error rate and distribution of errors be used. However, given the results of a measurement performed on multiple re-

petitions of an utterance by each of a suitably chosen set of speakers, we can evaluate certain general but useful properties related to the capability for separating speakers and to the extent of intermeasurement dependence.

3.3.1 Ability to separate speakers

The measurement data for each individual speaker may be regarded as samples from a distribution associated with that speaker. The individual speaker distributions of an ideally effective measurement would be disjoint. In practice, they are not disjoint, but it is desirable that they be as narrow and as widely separated as possible, in order that the test value of a measurement be associated with as few speaker distributions as possible. An intuitive measure of this condition is the average relative variance, which may be defined as the ratio of the average individual speaker sample variance to the total population sample variance. If this measure is low, then on the average, the individual distributions are narrow with respect to the distribution of the population.

A similar statistic which has been found useful by previous investigators is the F-ratio of the analysis of variance (Pruzansky and Mathews, 1964; Das, 1969). For the case where the number of measurements is the same for each speaker, and equal to n, the F-ratio is given by:

$$F = \frac{n(\text{Variance of speaker means})}{(\text{Average of speaker variances})}$$

(Since this statistic depends on n, a more general statistic would be F/n, but in the present study, n will always be equal to 10, so this normalization need not concern us.) The value of F increases as the individual distributions spread farther apart and as they become narrower. The F-ratio has the desirable property of invariance to translation and scaling. It is shown in Appendix I that ranking measurements according to descending values of F is equivalent to ranking them according to ascending average relative variances. Although the use of F is intuitively appealing, it should be pointed out that it is not optimal in the sense of minimizing any error probability, and it takes no account of possible dependencies between measurements.

Two other statistics have been proposed regarding the capability of a measurement to separate classes. Divergence (Marill and Green, 1963) and mutual information (Lewis, 1962; Kamentsky and Liu, 1963) both require the estimation of the underlying distributions. Since in this study there were only 10 repetitions of each measurement for each speaker, these statistics did not seem readily applicable.

3.3.2 Intermeasurement dependence

Much intermeasurement redundancy can be avoided by intelligent choices of measurements. If a spectrum is measured in the center of a tense vowel, another spectrum 10 or 20 msec later will probably yield little new information. However, if

the measurements concerned are less clearly related (e.g., energies in certain spectrum ranges in certain phonemes), the intuitive approach becomes inadequate. Statistical tests of independence that would apply to this situation do not appear to be easy to find. In the present study, a procedure has been developed which is related to this problem, but not in a strictly quantitative way. This procedure grew out of another inquiry into the separability of classes, which will require explanation.

The results of a single measurement on a set of speakers may be represented pictorially as shown in Fig. 1a. The n repetitions for each speaker are plotted in a horizontal line, and the data for each of the m speakers is plotted on a different line. If the measurement is a good one for speaker recognition, the individual speaker data will be clustered and the clusters will be separated from one another. A datum is termed discriminable from another speaker if it lies outside the range of that speaker as determined by the data of that speaker. For example, in Fig. 1a, the rightmost datum of speaker a is discriminable from speakers b and d, but not from speaker c. Only comparisons with other speakers are considered. If this comparison is performed for each of the mn data against the m-1 other speakers, an average measure of range exclusion discrimination is given by

$$P_{red} = \frac{d}{nm(m-1)} ,$$

where d is the total number of such datum-speaker discrimina-

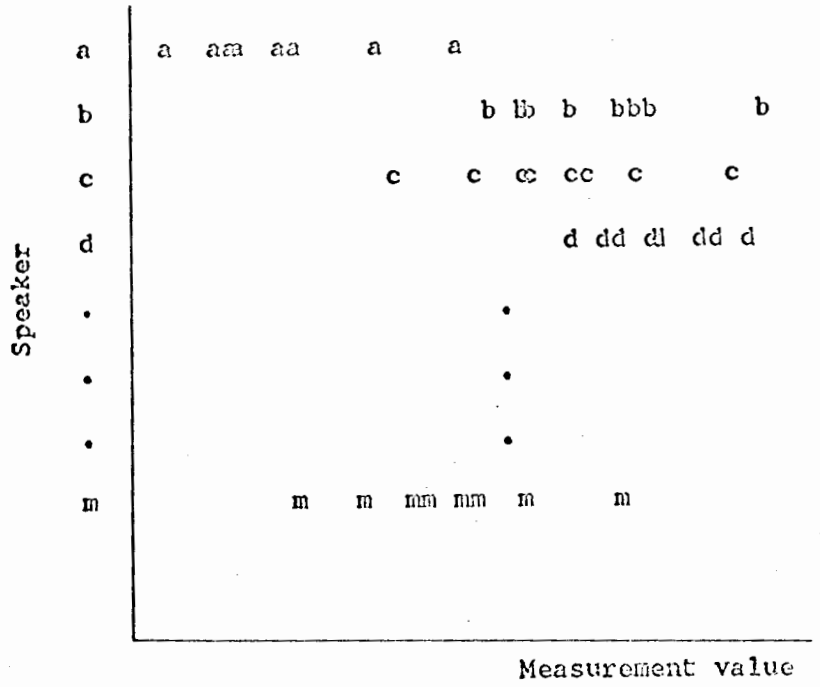

Figure 1a. Hypothetical measurement data on n repetitions by each of m speakers. Each speaker's data is plotted on a separate horizontal line.

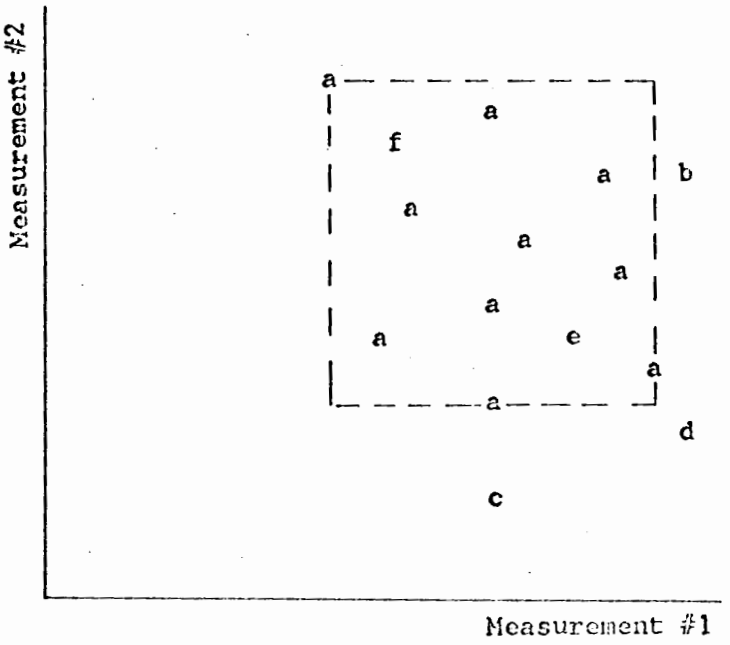

Figure 1b. Range exclusion discrimination using two measurements. The box delineates the ranges of Speaker a data in both dimensions.

tions. This quantity is a relative frequency; therefore it may be regarded as an estimate of the average probability of discrimination of one speaker from all other speakers in the set, using this measurement, and according to the range exclusion criterion stated above. This is admittedly not a good estimate of the discrimination capability of the measurement on the set of speakers, nor is it seriously intended to be. It measures only the extent of overlap of the ranges of the individual speaker data. Roughly speaking, if the measurement is a poor one for speaker recognition, the ranges will overlap; if it is a better one, they will tend to overlap less.

This procedure may be extended to the case of two or more measurements, considered jointly. In this case, a datum (of two or more components) is termed discriminable from another speaker if it lies outside the range of that speaker's data in one or more dimensions. For example, in Fig. 1b, data b, c, and d are discriminable from speaker a, but data e and f are not.

Note that the total proportion of datum-speaker discriminations is the same if the decisions are made considering all measurements jointly, as above, or if the discrimination decisions are made by first applying this procedure using one measurement and then applying it to the datum-speaker pairs not already found discriminable, using the next measurement, and so on. The effect of applying this r.e.d. procedure with

the first measurement may be regarded as removing a proportion of the datum-speaker pairs P_1 from further consideration, since they have been found to be discriminable. The sequential application of the r.e.d. procedure with the second measurement effectively operates on the remaining proportion $(1 - P_1)$. After the second application, a proportion P_{12} of the total datum-speaker pairs has been found to be discriminable. If probabilities, rather than probability estimates, were being used, and the two measurements were independent, then:

$$P_{12} = P_1 + (1 - P_1)\, P_2$$
$$(1 - P_{12}) = (1 - P_1)\,(1 - P_2)$$

We may expect that the estimated probabilities may approximate this relationship, or that the statistic

$$\Delta P = \frac{1 - P_{12}}{(1-P_1)(1-P_2)} - 1$$

will be close to zero for the case of independent measurements. In fact, as will be shown in Chapter 5, this statistic is small for measurement pairs that may be intuitively called independent and large for obviously dependent pairs.

Unfortunately, the distribution of this statistic under the hypothesis of independence is not known, so it is not possible to assign a critical region and significance level to this test. For the purposes of pragmatic pattern recognition, it may not be necessary to have strictly independent measurements. It may suffice to use measurements that are

merely not strongly dependent, for which purpose the ΔP-test with some threshold or the lowest combination of values of ΔP for pairwise comparisons of measurements is a useful technique.

CHAPTER 4

RECORDING AND PROCESSING OF DATA

4.1 Scope of the experiment

The speech data was taken from 21 adult male, American speakers, ranging in age from 22 to 42 years. None had a noticeable speech defect. Regional accent was not closely controlled; two speakers had mild southern accents. All speakers were staff or students at the Massachusetts Institute of Technology. They were apprised of the nature of the experiment and were accordingly asked to speak normally. Ten repetitions of six short sentences were recorded from each speaker.

The text of the speech data was specified by the experimenter. This is the usual condition in any speaker verification paradigm, and it may not be unreasonable in the case of identification. This was necessary because the acoustic measurements were to be performed on specific segments of the utterances.

The speech data was recorded in a single session. Only speakers who were reasonably free from colds or other inflammations were used. The stability of the measurements with respect to time or to the state of health of the speaker was not investigated.

The speech was recorded under low noise conditions with high quality, wide bandwidth equipment. The effects on the

measurements of a reduction in signal to noise ratio, bandwidth, or other condition of fidelity was not investigated.

Since the construction or simulation of a complete automatic speaker recognition system was not the aim of this study, the locations in the utterances where measurements were made were determined manually. They were determined systematically, however, in ways that were felt to be amenable to automatic implementation by simple computer programs.

4.2 Devising test utterances

Devising the test utterances really cannot be separated from selecting the measurements, since the utterances are the vehicles for providing the speech events on which the measurements are made. The measurements that were investigated will be described in detail in Chapter 5. Aside from the matter of the specific speech segments to be included, there are general considerations which must also enter into the process of making up the utterances.

Prior to the main experiment, some informal investigation was done on the sentence, "She remembers me," spoken by 10 people. As a result of this work and some use of the microphone input to the SPADE5 computer configuration (see section 4.4), a number of trial hypotheses were made about speech events that should be included, and something was learned about the rudiments of segmentation. In addition to specific items to be investigated, such as fundamental

frequency in stressed and unstressed positions, nasal consonants, and certain vowels, it was desired to include a wide variety of vowels, diphthongs, fricatives, and stops for possible investigation.

In this exploratory situation and also in practical situations, at least several seconds of speech data are required in order to provide sufficient number and variety of speech events for measurement. One or more sentences are preferable to isolated words or phrases. Grammatical sentences provide the speaker with a standard and hopefully unambiguous interpretation and hence a well defined pronunciation.

Once the specific speech segments to be measured have been selected, the task of incorporating them into a suitable sentence can be a frustrating one. The speech segments should be placed in favorable environments in the sentence, and the sentence should be easy to segment, natural to say, and usually spoken in just one way. It is no wonder that such sentences sometimes end up appearing rather contrived!

In a declarative sentence, the speaker normally lets his pitch and amplitude fall at the end of the sentence. Intra-speaker variability is probably increased at this time, for the pitch periods often become irregular, and there is a tendency to accompany the drop in voice level with less precise articulation. For this reason, the final syllable is generally not suitable for measurement. If a steady-state

vowel measurement is desired, that is, one primarily influenced by vocal tract structure and articulatory position rather than by dynamics, the vowel should be put in a context where the formant targets or steady-state positions are likely to be reached. Vowels should not be put in words in positions where they are reduced. If a vowel is stressed, it is generally lengthened. It has also been found that a vowel is lengthened before a voiced consonant, and that vowel formant targets are more closely approached if the consonantal context is a stop rather than a fricative, even though the duration is shorter (Stevens and House, 1963). Nasal consonants which are inherently low in intensity and are often short, are clearest and loudest when they precede a stressed vowel.

The utterance should be designed with an eye to the segmentation and recognition that will be required in order to locate the measurements. For example, the sentence, "How are you?" would be much harder to segment than "I saw Tom," because of the lack of voiceless segments and stops. Stops and strident fricatives are useful landmarks for cueing the segmentation of the sentence, but they cannot be sprinkled in too liberally, or the sentence becomes difficult and unnatural to say. Since we rely on the speaker using his own, well-established speech gestures, we wish to minimize unnaturalness. The problem of recognizing the beginning of an utterance can be minimized if the utterance begins with a stop or a vowel. Initial fricatives and nasals should be

avoided.

Li, et al. (1966) employed an interesting device to promote naturalness. They used short phrases like, "My name is ( )." They reasoned that the speaker would focus attention on saying his own name, and hence the first two words, which were actually used for measurement, would be free of undue emphasis.

Some words in English have more than one acceptable pronounciation, and individuals are not necessarily consistent in the version they use, particularly if they have lived in different regions and have been influenced by different dialects. Common examples of such words are a, either, and aunt. Stress can be similarly affected, as in downtown. It is desirable to avoid at least the most common of these words, in order that the performance of the system not depend on the speaker's remembering a standard version of the utterance. Naturally, the sentence itself should be unambiguous, since the syntax affects the pronounciation.

For the purposes of this experiment, the six short sentences given below were devised. This task was not an easy one, and some of the considerations mentioned above were occasionally compromised. The linguistic content of these sentences is certainly beside the point. The numbers associated with the sentences will be used to refer to them later.

1. Cool shirts please me.

2. Pay the man first, please.

3. I cannot remember it.

4. Papa needs two singers.

5. A few boys bought them.

6. Cash this bond, please.

Even though no automatic segmentation was contemplated for this experiment, each sentence begins with a stop or a vowel. The first word in sentence 5 was occasionally so brief that utterance initiation logic might have missed it. The last syllable of each sentence is intended as a "filler" because of the voice drop mentioned above. It was found, however, that the use of the word please on the end also kept the voice level high at the end of the sentence.

The sentences were designed to include a variety of speech sounds, not all of which were investigated in this experiment. Some of the principal ones are pointed out below, as are some of the shortcomings.

1. Cool was specifically used to get a good example of /u/. In many words, /u/ is not fully articulated. Sentence 1 also contains the fricatives /š/, /s/, and /z/ and the vowels /ɜ/ and /i/.

2. This sentence contains an example of the diphthongized vowel /eɪ/ and also /ə/ and /ɜ/. The /æ/ in man may be nasalized.

3. This sentence contains the diphthong /aɪ/, and nasals /n/ and /m/ in prestressed positions. It turns out that there are two acceptable stress markings for cannot, and care had to be taken in recording speakers to insure that the second syllable was stressed. The vowels /a/ and /ɛ/ may be influenced by the nasals.

4. This sentence contains /a/, /i/, and an /u/ that is too short to be fully articulated. The /n/ is in prestressed position, but the /ŋ/ is not favorably located for purposes of automatic segmentation and location.

5. This sentence contains examples of diphthongs /iʊ/ and /ɔɪ/ and the vowel /ɔ/.

6. This sentence contains /æ/, /š/, /ɪ/, and /s/. The /a/ may be nasalized. The location of the voiced stop /b/ following a voiceless sound turns out to be useful (see section 5.6).

Obviously, the sentences also contain other speech sounds. The stops and strident fricatives are useful for segmentation, as are the nasals when they occur between vowels.

4.3 Recording procedure

The recordings of speech data were made in the Research Laboratory of Electronics anechoic chamber using an Altec 684A dynamic microphone hung by a cord from the top of the

chamber. The microphone was positioned 10 inches from and 2 inches above the lips of the speaker. The speech signals were recorded on a Presto 800 tape deck at 7.5 inches per second. The recording apparatus was located in a nearby studio. Ten repetitions of the six short sentences listed in the previous section were recorded from each speaker in a single session.

A program tape was played to the subject seated in the anechoic chamber. This tape contained an explanation of the purpose of the experiment, an explanation of the recording procedure, and several practice sentences. The practice sentences also served the purpose of allowing the level control on the tape recorder to be adjusted to the subject's normal voice level.

The 60 utterances the subject was to say were presented by means of the program tape. This was done, instead of having him read from a list, in order to insure uniformity of the stress patterns in each sentence and to pace the subject in order to avoid the intonation pattern of continuation that subjects often use in reading items from a list. The subject was reminded that there was a danger that he might tend to slavishly imitate the exact intonation of the utterances on the program tape, and he was asked to "say the sentences in the same sense as that on the program tape." Most of the subjects were acquaintances of the experimenter, and it was felt that in most cases the utterances were spoken

naturally.

The six sentences were presented to the subject in mixed order, at intervals of 10 seconds, so as not to tire the subject and, again, to avoid "list intonation." A manually operated switch in the recording studio cut off the program material while the subject was speaking. If the subject made a mistake, the program tape was stopped and he was asked to repeat the sentence correctly.

The master data tapes were subsequently dubbed onto submaster data tapes, using two Presto 800 tape decks. The utterances were rearranged in the process so that each submaster tape contained one sentence, with only short pauses between the 10 repetitions by each speaker. A copy of each submaster tape was then made for subsequent analysis, and the submasters themselves were preserved as backups. The double copying increased the tape noise by several dB, but this was noticeable only in the high frequencies on the spectrum analyzer that was used in this study. Since nonstrident fricatives were not going to be studied, this high frequency noise was not important.

4.4 Analysis hardware and software

The computer facility of the Speech Communication group is built around a Digital Equipment Corporation PDP-9 general purpose computer with 24K of core memory. The computer is coupled to peripheral equipment especially designed to

facilitate on-line speech research (Henke, 1968). This highly flexible arrangement makes it possible, through convenient interconnections and proper programming, to create in effect a special purpose on-line laboratory instrument.

A system of programs (SPADE5) was written to enable this facility to be used as a general and special purpose spectral analysis instrument. A block diagram of this configuration is shown in Fig. 2. This diagram depicts logical rather than physical units, for the blocks enclosed by broken lines denote functions which are performed by parts of the computer program rather than hardware. The speech source input was either an Ampex 401A two-channel tape deck, which included provision for control by the computer program, or a microphone and amplifier installed at the operating position.

The principal speech analysis tool in the system is the real-time spectrum analyzer. This consists of a +6 dB/octave stage to emphasize the high frequencies, followed by 36 single-tuned bandpass filters covering the range 150 - 7025 Hz. The filter specifications are given in Table 1. The center frequencies are spaced linearly up to 1650 Hz and logarithmically thereafter. The characteristics of adjacent filters cross at their 3 dB points. Each filter is followed by an electronic rectifier and low pass filter (time constant of 10 msec). A 36 channel multiplexer selects the filter output to be sampled. A logarithmic analog-to-digital converter gives

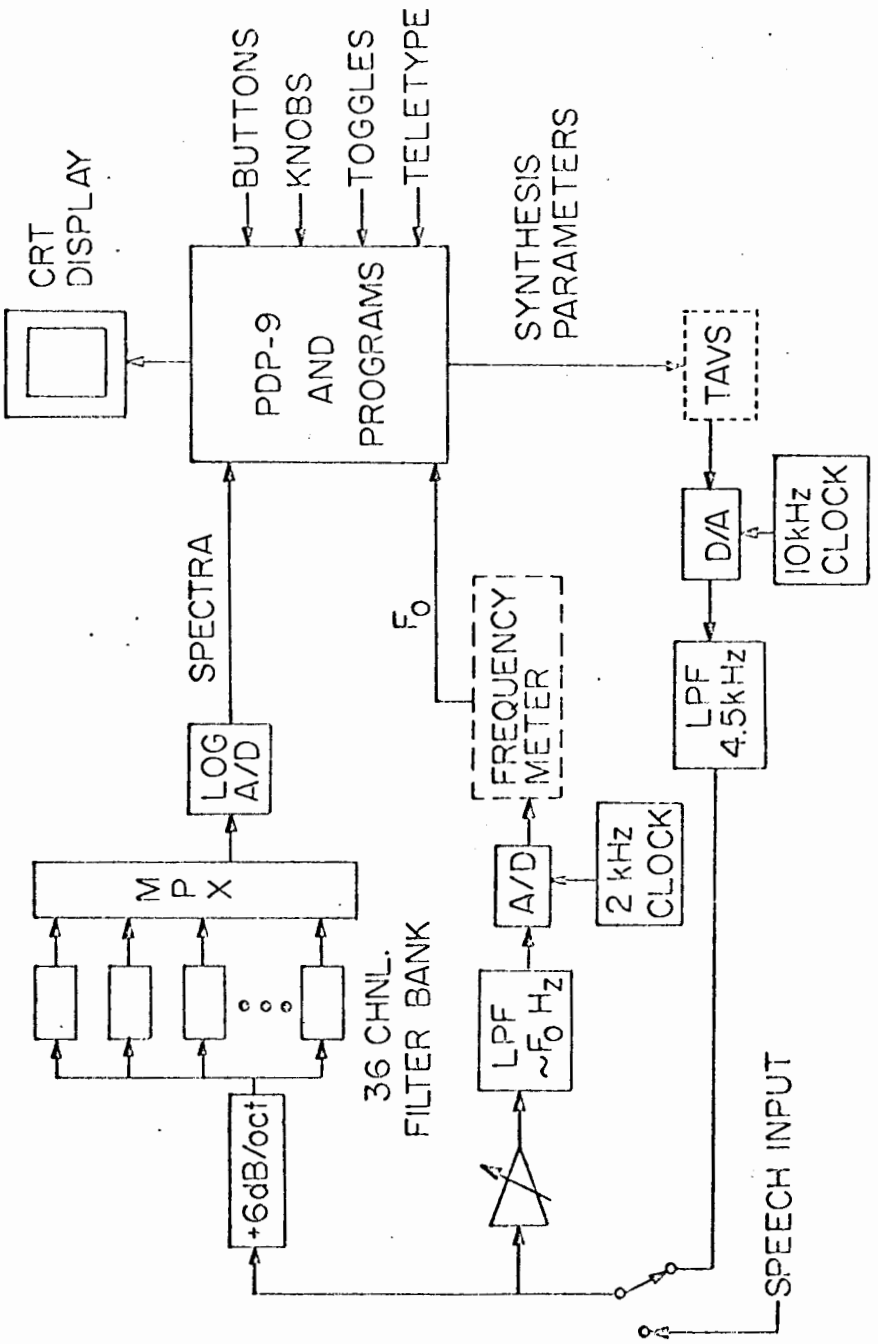

Figure 2. Speech Communication computer facility configuration for SPADES.

the output voltage directly in decibels. A complete scan of 36 channels is performed in 1.3 msec, which is small compared to the averaging time of the smoothing filters.

Since the first harmonic was present in the recorded speech data, fundamental frequency (F_o) could be measured by the rudimentary scheme shown in Fig. 2. A low pass filter with an 18 dB/octave skirt slope was set to a cutoff frequency of about the highest value of F_o expected for the particular speaker (often about 160 Hz). The output of this filter then consisted mainly of the first harmonic. This was sampled and converted to digital form at a 2 kHz rate, and a simple zero-crossing detection algorithm calculated the estimates of F_o. This method sometimes produced spurious values during unvoiced segments (since there was no check to see if voicing was present) and sudden transitions, but for most voiced segments, and for vowels in particular, it gave reliable and repeatable values. More effective and reliable pitch extraction schemes have been described in the literature (Gold, 1962; Noll, 1967). The limitations of this one should not be interpreted as limitations on measuring fundamental frequency in automatic speaker recognition systems.

A variation on the manual analysis-by-synthesis procedure described by Bell et al. (1961) was implemented for vowels on the PDP-9. A vowel spectrum was analyzed by an iterative procedure of postulating a set of pole positions,

TABLE 1. ANALYZING FILTER SET

Filter Channel Number	Center Frequency (Hz)	Bandwidth (Hz)
0	150	100
1	250	100
2	350	100
3	450	100
4	550	100
5	650	100
6	750	100
7	850	100
8	950	100
9	1050	100
10	1150	100
11	1250	100
12	1350	100
13	1450	100
14	1550	100
15	1650	125
16	1775	125
17	1900	150
18	2050	150
19	2200	175
20	2375	175
21	2550	200
22	2750	200
23	2950	225
24	3175	225
25	3400	250
26	3650	275
27	3925	300
28	4225	325
29	4550	350
30	4900	375
31	5275	400
32	5675	425
33	6100	425
34	6550	475
35	7025	475

calculating the filter bank response to a vowel having that pole configuration, comparing the calculated response to the measured spectrum, and revising the pole locations accordingly. In this version, instead of the calculation of the filter bank response for a given pole configuration, a 30 msec segment of vowel is synthesized from these parameters and analyzed by the filter bank itself. The synthesis is performed by TAVS, the vowel portion of a five formant 10 kHz sampled data terminal analog speech synthesis program written by Prof. W. Henke (1969). Not only is this procedure faster than the original implementation, but it is also more accurate, since the synthesized spectrum is derived using the measured value of F_o instead of an assumed value of 100 Hz, and there is no error in calculating the filter bank response. The principal limitation on the accuracy of the present system is the quality of the glottal source spectrum approximation.

The results of most of the sections of the program SPADE5 are displayed on a 16-inch cathode ray tube display. An arrangement of pushbuttons, knobs, and toggle switches provides a convenient interface for the user to control the operation of the program.

The principal features of SPADE5 are illustrated by the typical CRT display shown in Fig. 3. The short-time spectrum and fundamental frequency are measured every 10 msec during an utterance. The data buffers have a capacity of 2.5 seconds

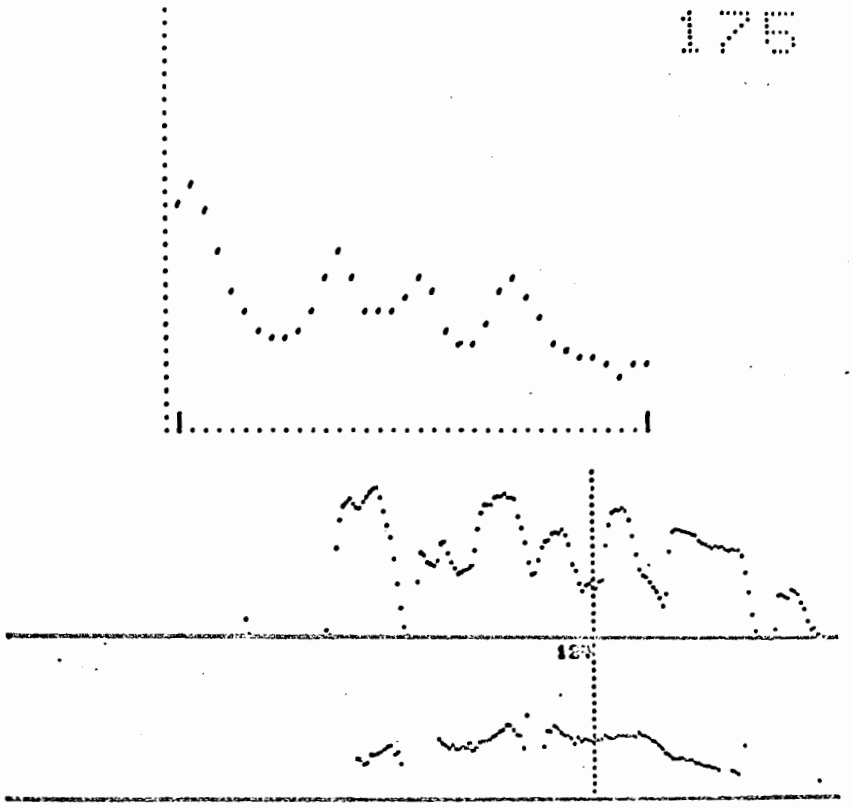

Figure 3. A typical SPADE5 display. The two graphs in the lower half represent functions of time, from 0 to 2.5 sec. The upper one is the sum of the outputs of filters 5-8, and the lower one is F_0. The vertical cursor shows the point in the utterance at which the short-time spectrum displayed above was measured. The horizontal axis of the spectrum represents frequency, from 150 to 7025 hz, and the vertical axis represents amplitude in dB. (The spectrum shown occurs in the first /m/ in I cannot remember it. It is the 175th frame in the buffer, and the value of F_0 at that point is 124 Hz.)

of speech at this rate. The two graphs in the lower part of Fig. 3 represent functions of time, from 0 to 2.5 seconds. The lower graph is fundamental frequency. The upper one, which will be called the "energy function," is formed by summing and averaging the outputs of certain filters, selectable by toggle switches. For nasal consonant measurements (and Fig. 3), filters 5-8 were used; in all other cases, filters 0-5 were found to work well. With these groups of filters, the energy function is a measure of low frequency energy, and it is used as a "syllable map" of the utterance for segmentation purposes. The vertical cursor shows the point in the utterance at which the spectrum displayed above was measured. Other capabilities of this program include the measurement of the amplitude and frequency of any feature on the displayed spectrum, storage of a selected spectrum in one of 16 special buffers for later comparison and measurement, the typing of spectrum data numerically or graphically, and the performance of special measurements, such as second and third central moments.

The speech data was kept in analog form on the tapes, since storage of the spectral data in digital form would have been far too bulky. This meant that the data was not exactly the same each time it was analyzed from the analog tape, since the sampling points were different each time. This was not important, since the type of measurements being studied should not be sensitive to minor variations of this type.

CHAPTER 5

THE MEASUREMENTS INVESTIGATED

In this chapter the matters of the specific speech events to be measured and the form of the measurements performed on each will be discussed. Where appropriate, measurements that were rejected will also be mentioned. Quantitative evaluations, as discussed in Chapter 3, will be presented for each measurement. The collective effectiveness of the measurements was evaluated by means of a simple speaker identification procedure. The algorithm and the results obtained will be discussed in the final section of the chapter.

For each measurement, each example of the appropriate sentence was read into the computer, the measurement location within the sentence was manually located, and the measurement datum was manually or automatically recorded, depending on the specific measurement implementation. Since this study of the effectiveness of the acoustic measurements was essentially exploratory, it was felt that this interaction between the experimenter and the segmentation and measurement processes was preferable to having trial measurements performed automatically, even at the cost of long hours spent taking measurement data.

Once the mechanics of a trial measurement were developed to the point where data would be taken on every speaker's utterances, the measurement locations were determined by simple

rules and procedures, so that the influence of the manual location procedures would be mininized. Figures 4 and 5 contain spectrograms of one example of each of the six sentences contained in the data. The locations marked on them will be referred to in the sections pertaining to the specific measurements.

In the course of the investigation, the various acoustic measurements were given mnemonic names, which will be used in this chapter to refer to them. The pertainent statistics for each measurement are summarized in Appendix II.

5.1 Fundamental frequency

The measurements of fundamental frequency proved to be the most useful single measurements investigated. F_o was measured at specific locations, rather than as an average over the whole utterance, for two reasons. First, pitch measurements at several locations in the utterance would probably be very dependent, but in addition to average pitch, they would contain information about the pitch contour, which had been used by Atal (1968). We wished to find out if such measurements would be useful in the context of a small number of efficient measurements (or if we would do better to measure pitch only once and make other, less dependent measurements). Second, we wished to find out if the increment in F_o due to stress would be useful.

Fundamental frequency was first measured at six loca-

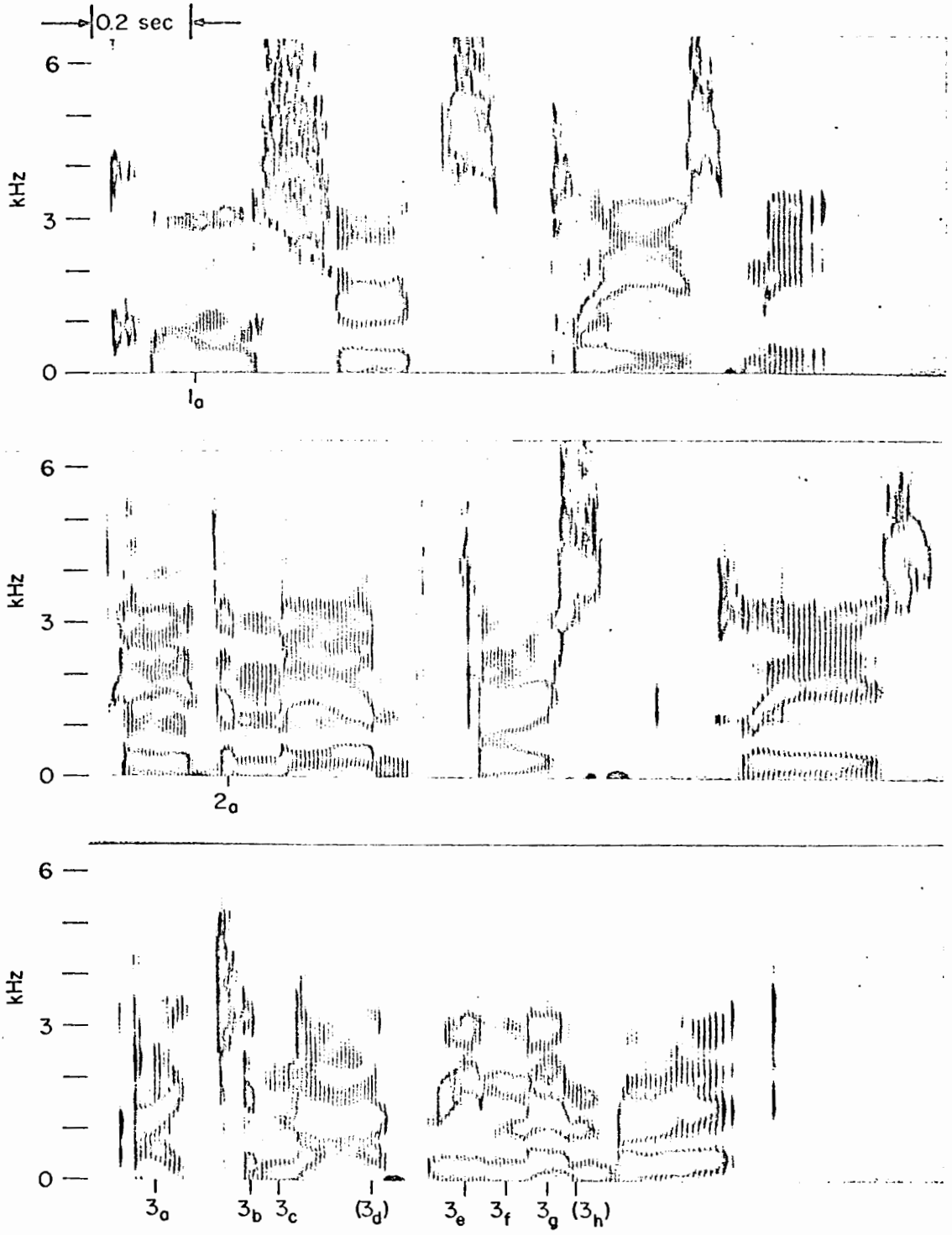

Figure 4. Spectrograms of sentences 1, 2, and 3. Top: Cool shirts please me. Middle: Pay the man first please. Bottom: I cannot remember it.

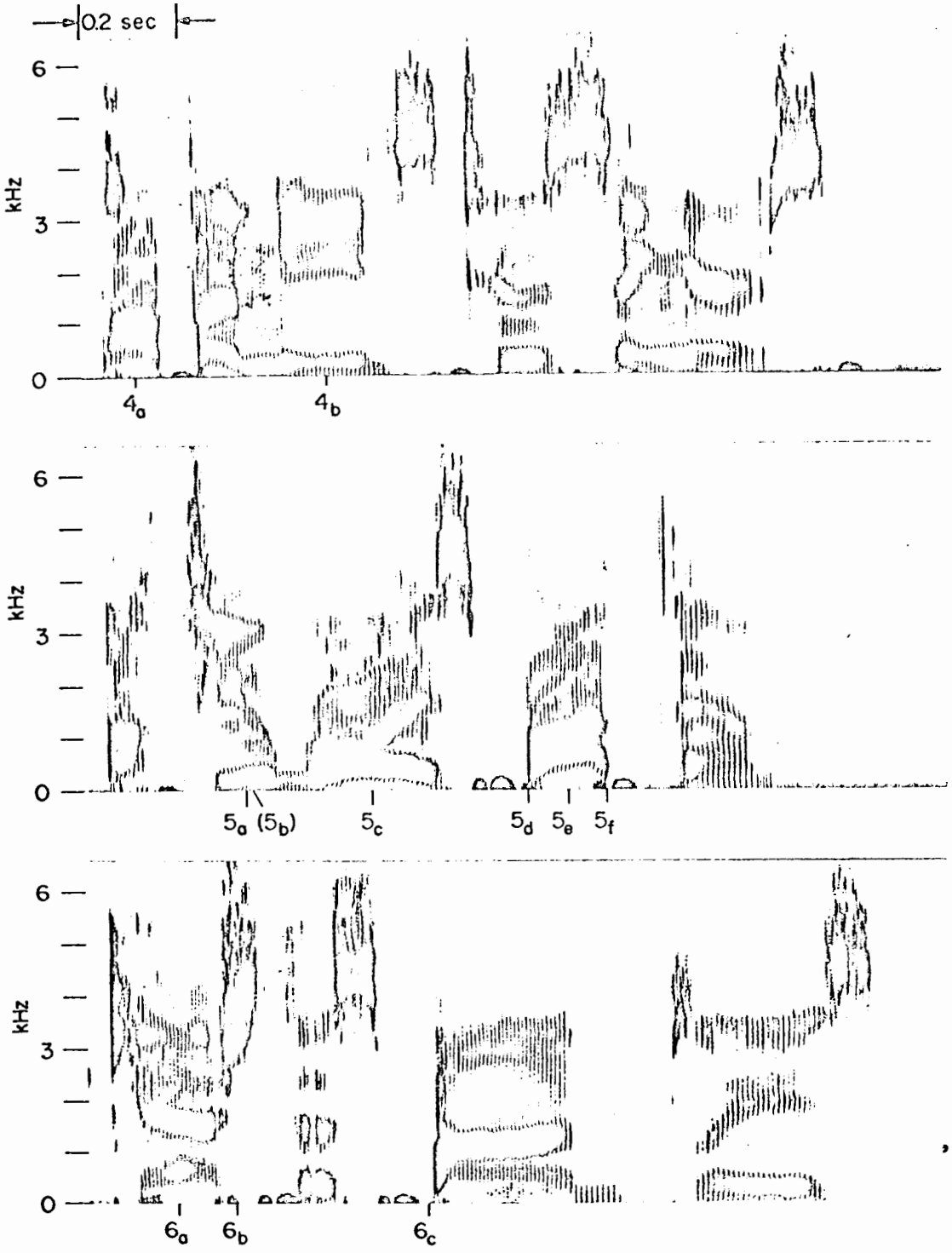

Figure 5. Spectrograms of sentences 4, 5, 6. Top: Papa needs two singers. Middle: A new boat bought them. Bottom: Cash this bond please.

tions in sentence 3. The names and locations of these measurements are listed below:

3F01: during the middle of I as shown by the large peak in the energy function corresponding to to that syllable (point 3a in Fig. 4).

3F02: during the middle of the first vowel in cannot. The energy function drops suddenly at the onset of /n/, so the vowel is delineated between the burst of /k/ and the beginning of /n/ (point 3b in Fig. 4).

3F03: at the peak of F_o during the syllable not (point 3d in Fig. 4). Cannot was stressed on the second syllable.

3F04: during the first vowel in remember, at the peak of the energy function for that syllable (point 3e in Fig. 4).

3F05: during the middle of the second vowel in remember (point 3g in Fig. 4). The nasals on either side clearly delineate the vowel in the energy function.

3F06: at the peak of F_o corresponding to the stress on the second syllable of remember (point 3h in Fig. 4). In those cases where there was a peak in F_o due to the stress, it usually occurred during the second /m/. If there was no rise, 3F06 was given the same value as 3F05.

The increments in F_o at the stressed syllables in cannot and remember were obtained by subtracting 3F03 from 3F02 and 3F06 from 3F04. Inspection of these data showed large variations within most individuals' utterances. Since in many cases these variations were about the same as the total range, the possibility of using these increments as characterizing measurements was abandoned without further analysis. Measurement 3F06 was also discarded, since it was often no different from 3F05.

The F-ratios for the five remaining measurements are given below.

Measurement	F-ratio
3F01	61.8
3F02	71.2
3F03	30.9
3F04	51.8
3F05	52.8

Measurement 3F03 is rated appreciably poorer than the other four. A second look at some examples of sentence 4 easily showed why. The stress in the word cannot was on the second syllable. As a result, F_o began to rise during the /n/ and often did not reach a peak before the vocal tract closure of the /t/ cut off the voicing. As a result of this sudden transition, the F_o measurements often contained several spurious (very high or very low) values at this point. In these cases, the datum recorded for 3F03 was the last value which connected continuously with the previous values. Thus there was a greater time uncertainty in the location of the

measurement due to the sudden articulatory movement at that point and the inability of the F_o measurement technique to handle this case well.

The intra-speaker variability of this pitch measurement is probably also increased by an articulatory phenomenon. There is less need for precise control of the rise of pitch due to stress in this context, since the /t/ which ends the stressed syllable effectively controls it by terminating the voicing. This context may be contrasted to the second syllable in remember, where there is no interruption of the airflow through the larynx, so the F_o contour must be explicitly controlled. The moral of this story is not to place F_o measurements in locations coincident with sudden transitions from a sonorant to a nonsonorant.

Fundamental frequency was also measured in five other locations in sentences 5 and 6.

5F01: in the middle of few, as shown by the energy function (point 5a in Fig. 5). (F = 81.0)

5F02: at the peak of F_o in few, usually very close to 5F01 (point 5b in Fig. 5). (F = 84.9)

5F03: in the middle of the diphthong in boys (point 5c in Fig. 5). (F = 54.3)

5F04: in the middle of bought (point 5e in Fig. 5). (F = 69.5)

AEF0: in the middle of the vowel in cash in sentence 6 (point 6a in Fig. 5). (F = 72.2)

The values of F-ratio for these five measurements confirm that the variability in 3F03 was a special case and not a consequence of the syllable being stressed. Aside from 3F03, every F_o measurement had a higher value of F-ratio than the next best measurement. With the small number of examples at hand, there seems to be no particular advantage to stressed or unstressed syllables. It should be noted that the use of an F_o measurement requires the assumption that the speaker is in some kind of normal, cooperative state. Fundamental frequency is very susceptible to stress on the speaker (Hecker, et al., 1968) and it is perhaps the easiest and most obvious acoustic correlate to modify for the purpose of voice disguise.

5.2 Nasal consonants

The articulatory configuration of the nasal consonants makes them particularly appropriate for speaker recognition measurements. They are formed by closing the mouth cavity at some point and opening the velum, permitting air flow through the nasal cavity. Hence a portion of the acoustic system for nasal consonants is fixed and is not subject to articulatory movement and variation. Glenn and Kleiner (1968) state that the other articulators do not move during the period of oral closure, in contrast to their virtually constant motion during other phones of normal speech. This statement is perhaps an approximation, but spectrograms of

nasal consonants are characterized by largely horizontal formants during the nasal murmur. Furthermore, nasal consonants are not rare events, but comprise 11% of the phonemic content of commonly spoken English (Tobias, 1959).

Since the mouth cavity acts as a shunt, it introduces zeros into the spectrum of nasal consonants. The spectrum of a clear /m/ is shown in Fig. 3 in Chapter 4. The region between the first formant and the second spectral peak contains a pole, but the lowest zero due to the mouth cavity has effectively cancelled its effect in the spectrum. The third and fourth spectral peaks occur around 2 and 3 kHz. The spectrum of a clear /n/ is similar, but the shorter mouth cavity means that the lowest zero occurs higher in frequency, often effectively canceling the pole in the neighborhood of 1.3 kHz, and leaving the pole just below 1 kHz in the clear (Fujimura, 1962).

The interplay between the mouth and nasal cavities can produce considerable variability in the 700-1600 Hz portion of the spectrum, depending on the nature of these cavities, and hence on the individual speaker. The analysis and experiments of Fant (1960) and Fujimura (1962) suggest that certain poles of the transfer functions of the nasal consonants are closely tied to the nasal cavity alone.

The first formant is very low, and it is ascribed to a lumped-circuit resonance between the pharyngeal and the nasal cavities. Fujimura found it to be quite stable. The second formant, usually not visible in /m/, but often visible

in /n/, seems to be approximately a quarter wavelength resonance of the nasal cavity alone when it occurs around 1 kHz in /n/. Fujimura ascribes the formant occurring around 2 kHz in /m/ to the nasal cavity by virtue of its stability and large bandwidth. He also comments on the stability of the formant around 3 kHz in /n/, which Fant assigns to the three-quarter wavelength resonance of the nasal cavity.

These arguments suggest that the locations of these spectral peaks would make good speaker recognition measurements, since they are closely tied to a specific anatomical feature. In actual fact, the nasal spectra, at least as shown by the 36-channel filter bank, do show variations among speakers, but these spectral peaks are often impossible to identify, thus failing the measurability criterion.

Some examples of this variation of the visible features are shown in the computer display photographs in Fig. 6. Each row contains four examples of /m/ by one speaker, and different speakers are represented by different rows. The top row is by the same speaker as in Fig. 3. The second row shows a speaker whose F_2 is not completely cancelled by the zero, resulting in a small peak around 800 Hz. The third row shows a speaker whose F_2 and F_3 are both considerably affected by the zero, resulting in a lack of peaks in that region of the spectrum. The formant damping is generally higher in the nasals than in the vowels, and that may also have contributed to the weakness of the peaks. In the fourth row, the peak around 3 kHz is absent from the spectrum. This is

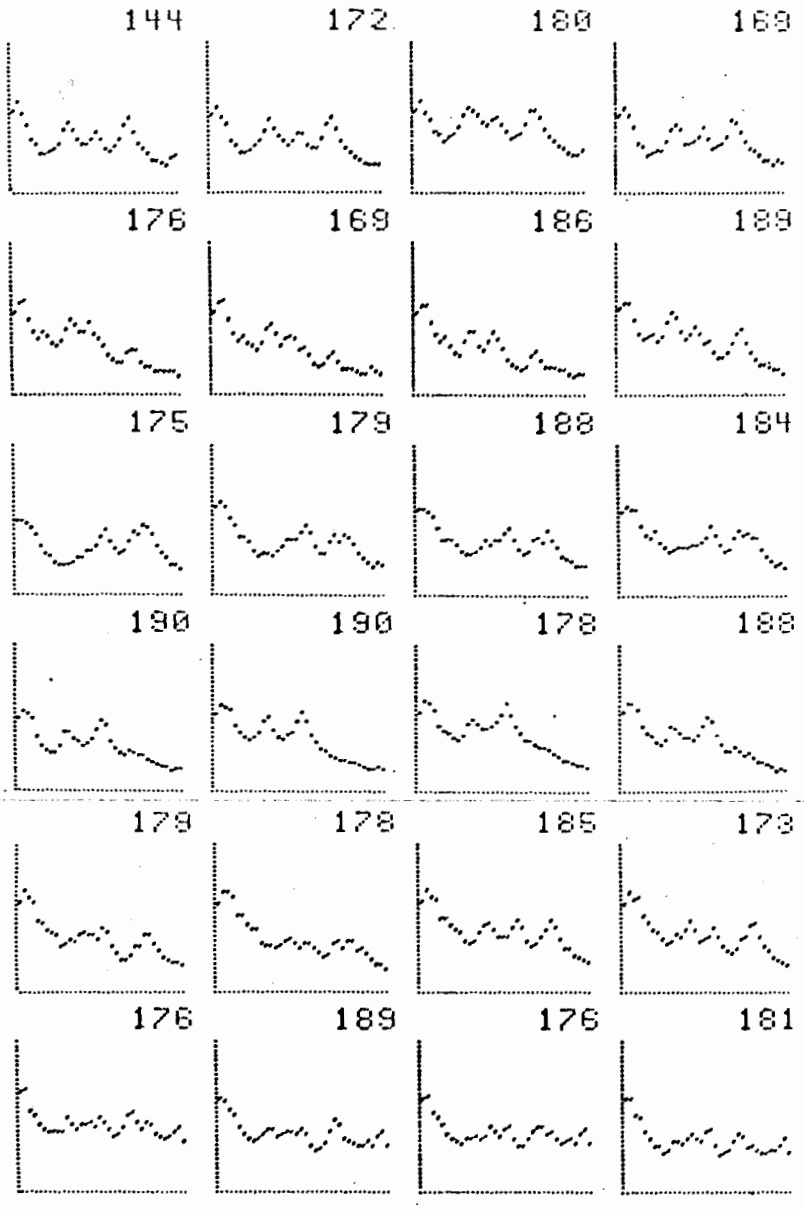

Figure 6. Spectra of /m/. Each row contains 4 examples by one speaker, and different speakers are represented by different rows. (The numbers have no significance.)

probably due in this case to the proximity of that pole with the second zero of the mouth cavity. The bottom two rows show /m/ spectra in which the consistent identification of these peaks is difficult, if not impossible. The spectra of /n/ show similar effects.

The additional variability introduced by the zeros and the reduction in the definition of spectral peaks caused by the higher damping of the nasal consonants make formant measurement a difficult if not impossible technique. Prof. D. Klatt suggested that the individual filter outputs in the neighborhood of these formants be examined to see if they are generally sensitive to changes in formant location. Since such data are subject to variation due to differences in voice level, they must be suitably normalized.

A subprogram was written for SPADE5 which performed the accumulation of data from selected filter outputs, subject to an intensity-normalization term for each utterance. This subprogram was used to make measurements in the middle of the /n/ and first /m/ in sentence 3 (points 3e and 3f in Fig. 4). They were normalized by subtracting the value of the energy function in the following vowel. Filters 5-8 were used for the energy function in this case, since it was found that this frequency region emphasized the lower intensity of the nasals. The nasal consonants then showed up on the energy function as short regions of noticeably lower values (refer back to Fig. 3). These nasal measurements were given

the names 3Mi or 3Ni, where i is the filter number.

In this manner, many different measurements were accumulated for these examples of /m/ and /n/. Figure 7 shows graphs of the F-ratio for each measurement versus filter number (hence versus frequency). The measurements taken from the filters that roughly correspond to the frequencies of the spectrum features described above make broad peaks in the F-ratio curves. The fact that these maxima are not just due to single points having high values supports the contention that they represent these features. These maxima correspond to the region of pole-zero interplay below 1 kHz and to the formants around 0.25, 2, and 3 kHz in /m/ (filters 1, 6, 17, and 23), and to the formants around 1, 2, and 3 kHz in /n/ (filters 8, 18, and 23). This technique falls short of the criterion of tying a measurement directly to a structural feature, but in the absence of the ability to characterize speakers by the location of certain spectral features, it does take cognizance of the locations of the poles and zeros underlying the spectra.

Although the nasal consonants are less subject to movement of the articulators than other sounds, they may be particularly sensitive to the state of health of the speaker. Certainly a bad cold can block the nasal passage completely, with the result that the nasals are transformed into the corresponding voiced stops. No work seems to have been done on the effects of respiratory inflammations on the acoustic

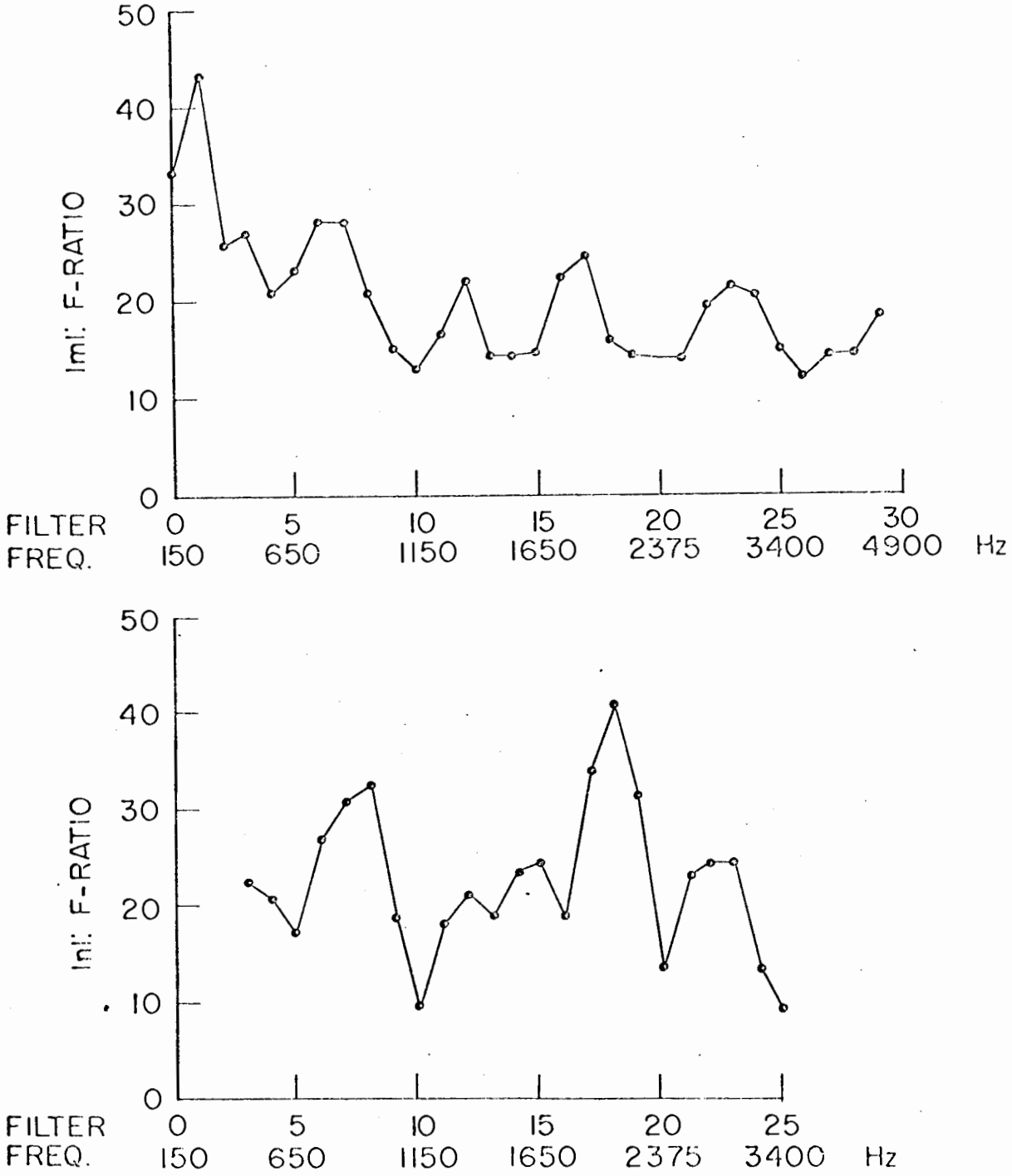

Figure 7. F-ratio vs. filter number for /m/ and /n/.

correlates of speech, so it is not possible to state at this time the conditions under which these or other measurements will deviate significantly from the normal.

5.3 Vowels

The length of the vocal tract and the sizes of its various parts determine the frequency ranges of the formants. Speakers differ somewhat in these ranges, yet listeners can easily perceive the same vowel in spite of large differences in vocal tract size (Peterson and Barney, 1952). Range of formants has been found to be a correlate of voice quality (Ladefoged and Broadbent, 1957; Shearme and Holmes, 1959; Miller, 1964). It has also been shown that the identification of a vowel can be strongly influenced by changing the formant ranges of surrounding vowels (Ladefoged and Broadbent, 1957). This last finding suggests that the listener effectively applies some kind of normalization to the important formants for each speaker. The "calibrating information" for this normalization must come from the first few instants of speech, since we can immediately understand the speech of a stranger. Since this normalization is speaker-specific, its acoustic correlates would be good measurements for speaker recognition.

Hemdal (1967) used the formant frequencies of the schwa vowel (/ə/) as reference data for formant variability compensation in a speech recognition experiment, with moderate success. It was thought that this neutral unstressed vowel would

reflect the vocal tract length of each individual.

Gerstman (1968) has shown that scaling F_1 and F_2 linearly between the extreme values of F_1 and F_2 for each speaker is an effective procedure for reducing vowel formant variability across many speakers. This procedure suggests that the extremes of vowel articulation (/i/, /a/, and /u/) provide some sort of reference points for the formant ranges of the individual. It can perhaps be argued that these articulations are more stable than others since they require control only to the extent of moving the articulators to an extreme position, as opposed to an intermediate position. In fact, it has been found that the first two formants in /i/, /a/, and /u/ are the least sensitive to the effect of context (Stevens and House, 1963). This argument is supported by Stevens' theory of the quantal nature of certain vowel articulations (Stevens, in press).

Four vowels were examined for their use in speaker recognition measurements. They were the schwa in sentence 2 (point 2a in Fig. 4), the /a/ and /i/ in sentence 4 (points 4a and 4b in Fig. 5), and the /æ/ in sentence 6 (point 6a in Fig. 5).

The formants of the schwa vowel are ideally spaced at intervals of approximately 1 kHz, and hence they show up as distinct peaks with the filter bank spectrum analyzer. The formant frequencies were measured using a peak interpolation algorithm in SPADE5, which roughly interpolates the frequency of a spectral peak from the local maximum and the data on

either side. The peak corresponding to F_3 was generally weak and occasionally absent, and F_4 was also frequently unclear, so the first two formants were the only measurements analyzed further. For measurements UHF1 and UHF2, the values of F-ratio were 21.1 and 44.6.

Measuring the frequencies of F_2, F_3, or F_4 in the vowel /i/ is often not possible with the filter bank spectrum analyzer. All three formants combine to form a broad concentration of energy in the 2-4 kHz region, and the analyzing filters are not narrow enough to permit the resolution of the individual peaks in all cases. This condition is illustrated by the examples of /i/ from four speakers shown in Fig. 8. The shape of this broad high frequency peak, which is determined by the frequencies and bandwidths of F_2, F_3, and F_4, seems to be characteristic of the speaker.

A subroutine was written to evaluate the second and third central moments and the skewness of any frequency range of a selected spectrum. These measurements pertain to the shape of the spectrum as displayed, i.e., horizontal coordinate represents filter number, and vertical coordinate represents amplitude in dB. The skewness is defined as

$$\gamma = \frac{\mu_3}{(\mu_2)^{3/2}}$$

where μ_2 and μ_3 are the second and third central moments. It was found that since the variance of μ_3 was much greater than that of μ_2, the skewness measurement behaved similarly to μ_3,

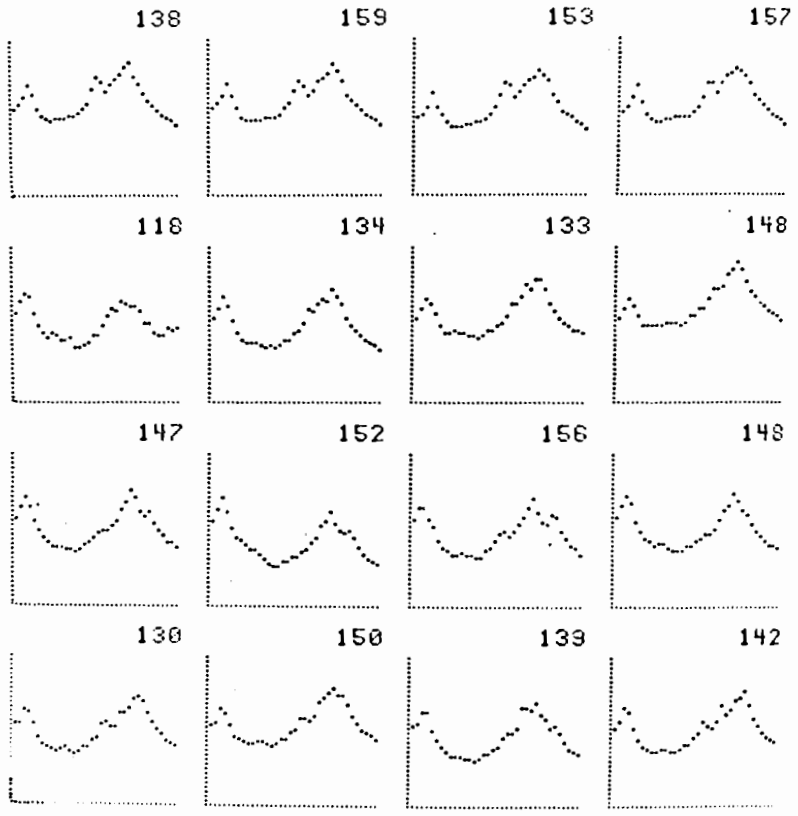

Figure 8. Spectra of /i/. Each row contains 4 examples by a single speaker.

and it was subsequently discarded. It was also realized that these moments depended on the overall height of the curve, and that the portion of the curve below the minimum value made a large contribution to these moments, which remained constant even if the shape changed. The algorithm was modified to calculate the moments, setting the zero coordinate to the minimum value in the range, so as to produce a greater variation in the moments, due to changes of shape.

The frequency range 1.55-4.55 kHz was empirically selected for the vowel /i/, since that included the major portion of the F_2-F_3-F_4 concentration. For measurement IS2, the second central moment, and IU3, the third central moment, F-ratios of 32.7 and 34.4 were obtained.

Measurement of F_1 and F_2 in the vowel /a/ is also difficult for many speakers. These formants are close together, producing a broad peak in the range 500-1500 Hz, as illustrated by the examples of /a/ from four speakers shown in Fig. 9. The second and third central moments for /a/ were measured over the range 350-1550 Hz, but these measurements were not as successful as the ones for /i/. For AS2, the second central moment, and AU3, the third central moment, F-ratios of 11.8 and 10.2 were obtained.

The analysis-by-synthesis program described in Chapter 4 was developed to permit formant measurements in the cases where the spectral peaks are not distinct. The analysis procedure consists of selecting a spectrum to be analyzed and

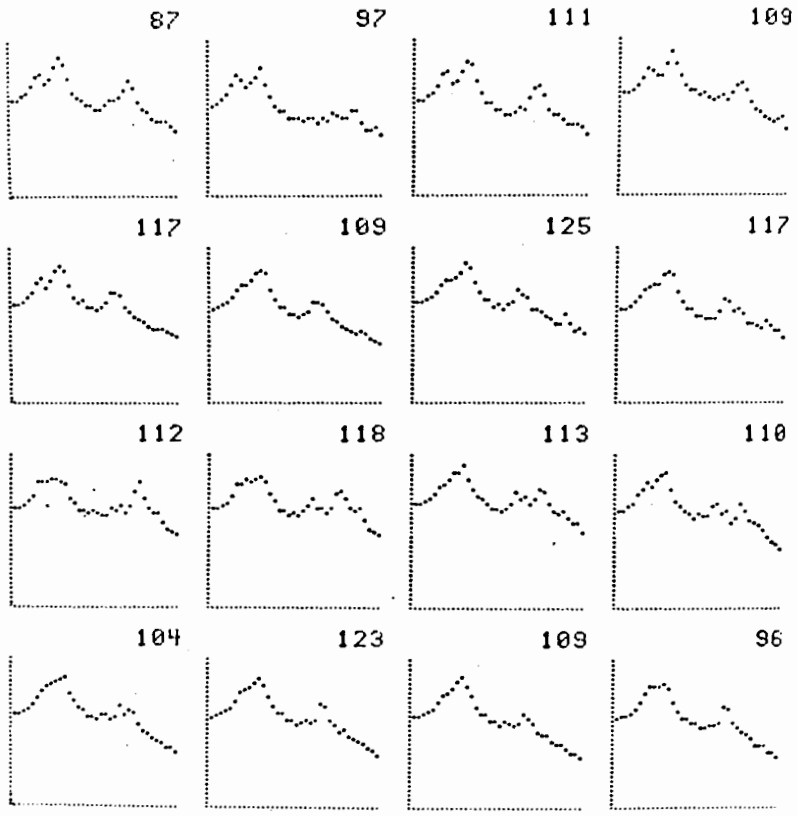

Figure 9. Spectra of /a/. Each row contains 4 examples by a single speaker.

then manually setting the values of frequency and bandwidth of the first four formants by means of pushbuttons and a knob adjustment so that a good match is obtained between the real and synthesized spectra in the range up to 3.2 kHz. The fifth formant was kept constant at 4.5 kHz, for it had little effect in the comparison range. The value of F_o measured at the time of the selected spectrum was used in the synthesis.

This technique was first applied to the analysis of /æ/. This vowel was felt to be an easy one, since like /ə/, the first four formants are generally distinct. With the experience gained on /æ/, the vowel /a/ was also analyzed. In this case, the F_1 and F_2 peaks were usually not distinct, and the task was somewhat harder. For most examples, a successful match could be obtained in a little over a minute, and the job was made easier by the fact that the 10 examples by each speaker were similar. The task became tedious for a large number of analyses. This analysis technique is amenable to automation (Paul, et al., 1964).

The match between the two spectra, as expressed by the squared error (Bell, et al., 1961) is much less sensitive to the formant bandwidths than to the formant frequencies. In addition, the bandwidths sometimes had to be set to extreme values in order to have the formant peaks at the right amplitudes, particularly in the case of F_3. This condition is attributed to inaccuracy of the glottal source spectrum

approximation. Consequently, the bandwidths were not felt to be as accurate as the formant frequencies. The inaccuracy due to the glottal spectrum effect does not necessarily invalidate the use of the bandwidths, but it does mean that their intra-speaker variability comes from two separate sources. For that reason and because of the first effect stated above, they were not tried as speaker-characterizing measurements. The third formant peak in both /æ/ and /a/ is often indistinct, so only the first two formant frequencies were used. The values of F-ratio for the first two formants of /ə/ (repeated), /æ/, and /a/ are given below.

UHF1:	21.1	UHF2:	44.6 (from spectrum peaks)
AEF1:	15.5	AEF2:	46.6 (analysis-by-synthesis)
AF1:	22.9	AF2:	19.0 (analysis-by-synthesis)

In the vowels /ə/ and /æ/, the F_2 measurement is by far the better one. Inspection of the measurement statistics in Appendix II shows that the total distributions of the F_2 measurements are about twice as wide as the F_1 measurements. The higher variability of F_2 can be interpreted as greater opportunity for variation among speakers. In the case of the vowel /a/, the F_1 distribution is the wider, but neither F_1 nor F_2 is as wide as in /ə/ and /æ/.

5.4 Source spectrum slope

The laryngeal excitation exhibits characteristics of individual larynges, as we have seen in the case of fundamental

frequency. The structure of the larynx affects not only the pulse repetition rate, but also the pulse shape, which is reflected in the envelope of the laryngeal source spectrum. Unfortunately, this spectrum is not directly accessible to measurement in the speech signal, since of course it is modified by the transfer function of the upper vocal tract. Mártony (1965), using inverse filtering, has found significant differences among speakers in the high frequency slope of the source spectrum. The inverse filtering technique is complex and probably not amenable to automatic processing.

Using a suggestion by Prof. K. Stevens, a measurement which crudely approximates source spectrum slope from a vowel spectrum was implemented with moderate success. The higher formant peaks in a vowel spectrum fall off in amplitude, due to the source spectrum slope of about -12 to -18 dB/octave and to the increased damping of the higher formants (and the +6 dB/octave radiation characteristic). Amplitude measurements at a low frequency formant peak and at a high frequency formant peak would approximate the extent of this drop, if there were little variation in the sharpness of the peaks, if formants were not so close as to enhance each other's amplitudes, and if there were little variation in the positions of the other formants. Variations in the frequency separation of the peaks could be roughly compensated for by dividing the amplitude difference (in dB) by the frequency difference on a logarithmic scale.

These conditions are approximately satisfied in the vowel /u/. F_1 is low, and F_2 is generally at least an octave higher. F_3 and F_4 are generally separate, and at least one of them is usually visible. That is, if F_3 is very weak, F_4 can serve as a measurement point. The actual algorithm for the measurement is the difference in amplitudes (in dB) between the maximum below 550 Hz (i.e., F_1) and the maximum above 2 kHz (i.e., F_3 or F_4), divided by their frequency difference on a logarithmic scale (i.e., log F_3 - log F_1). This algorithm may not be intuitively pleasing as an approximation to the phenomenon it purports to measure, but it has been used with some success. It is just not certain that this success is not partly due to the combination of other factors which affect it. This measurement, named UM1A, was implemented for the /u/ in sentence 1 (point 1a in Fig. 4), taken at a point one-third of the way through the first syllable, to simplify the problem of segmenting the /u/ from the /l/.

The F-ratio for this measurement was 36.3. Several alternate measurements of this type were also tried, with very little success. One of these was similar to UM1A, except that the second measurement point was the relative minimum between F_2 and F_3. Others omitted the division by the frequency separation term.

5.5 The fricative /š/

The spectrum of the fricative /š/ depends mainly on the

anatomical details of the region around and forward of the alveolar ridge. Hence measurements on /š/ are not influenced by the entire vocal tract, but only by a small portion of it. It was found that the locations of the high frequency spectral peaks are not particularly stable, but that the shape of the high frequency region seems to be characteristic of the speaker.

It is possible to classify examples of /š/ in terms of gross shape. Figure 10 shows examples of /š/ by four speakers. These examples illustrate the four shapes that have been so defined. They are, from the top, single narrow peak, wide or double peak, flat region, and very low major peak. (The asymptotically flat low frequency region shown in the display is due to the fact that the high frequency skirts of those filters are coincident. The amplitude of those filter outputs is an artifact of the filter characteristics, rather than an indication of energy at those points.)

The shape classification algorithm is described by the following ordered set of rules.

1. If there is a major peak (i.e., a dip of at least 3 dB on the high side) lower than the 2.55 kHz filter, call it low major peak.

2. If the maximum drop in amplitude above the highest filter output is 6 dB or less, and

    2a. the highest filter output occurs lower than 5 kHz, or

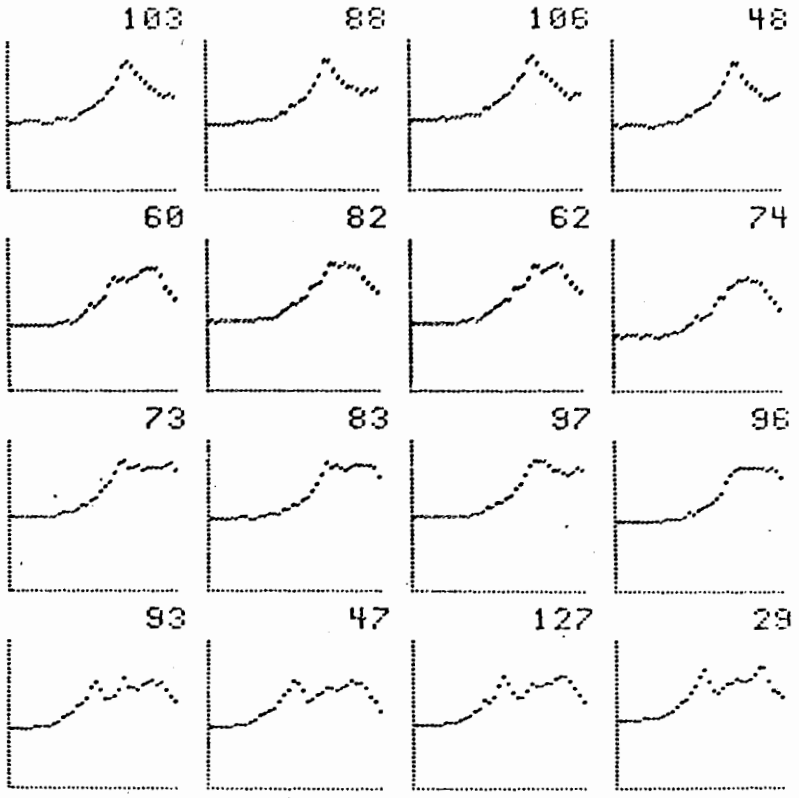

Figure 10. Spectra of /ʃ/, illustrating the 4 shapes. Each row contains 4 examples by a single speaker.

2b. the maximum drop in amplitude from the highest filter output to the 3925 Hz filter is less than or equal to 6 dB,

then call it flat region.

3. Examine the spectrum on either side of the highest filter output. If the 6 dB down points are less than 2 kHz apart, call it single narrow peak. Otherwise call it wide or double peak. This step requires interpolation between data points, since the spectrum may fall steeply. (If the highest filter output is so high in frequency that there is no upper 6dB down point, call it single narrow peak. This would be rare for /š/.)

This shape classification was performed on the example of /š/ in sentence 6 (point 6b in Fig. 5). The /š/ in shirts in sentence 1 was not used, since the lip rounding due to the coarticulation of /ɝ/ strongly modifies the spectrum so that this set of prototype shapes does not hold. Many of the speakers had examples of /š/ falling in two shape-classes, and a few had a single example falling in a third class.

This measurement is a discrete, qualitative measurement, as opposed to the continuous, quantitative measurements discussed previously. The natural way to characterize a speaker is by probability estimates of each shape-class, but this is not directly compatible with quantitative measurements in terms of evaluative measures and classification procedures.

For these purposes, the numerical values 1, 2, 3, and 4 were arbitrarily assigned to the shape-classes narrow peak, wide peak, flat region, and low major peak, respectively. This is probably not an optimum assignment, but on this basis, measurement SH has an F-ratio of 17.5.

Measurements of second and third central moments of the high frequency region of /š/ were also tried, with very little success. The fricative /s/ was found to be similar to /š/, but with features occurring higher in frequency, nearer the upper limit of the spectrum analyzer. It was not formally investigated.

5.6 Voice onset time

In voiced stops following an unvoiced segment, the onset of voicing before the release of the stop is not used for phonetic distinction in English, yet it is not uncommon. (This is the "voicebar" observed in spectrograms.) The speaker-specifity of this phenomenon was pointed out by M. Medress (personal communication). It was examined in the single example of this context in the data, this bond in sentence 6 (point 6c in Fig. 5). A binary distinction (prevoiced or not) seemed appropriate, since the duration of prevoicing showed wide intraspeaker variations, and it is difficult to measure precisely. A stop was termed prevoiced if voicing preceded the burst by 20 msec or more. In 10 examples from each of the 21 speakers, 6 prevoiced more than half the

time, 4 did so only occasionally, and 11 never did. Assigning the value of 1 to prevoiced examples and 0 to unvoiced ones, an F-ratio of 14.5 was obtained for measurement PREV.

This measurement is particularly appealing because it concerns a rapid event which is not likely to be consciously modified, and it is an event of such specificity that it is probably independent of most other measurements. It is, however, dependent on good recording conditions, since low frequency background noise or poor low frequency response would make this measurement impossible.

5.7 Duration of "bought"

As a single example of a measurement of speech timing, the duration of bought in sentence 5 (points 5d to 5f in Fig. 5) was investigated. This measurement, like the last, is dependent on learned rather than organic characteristics of the speaker. The energy function of a word that begins and ends with stop consonants rises and falls sharply during the stopgaps, so the measurement of duration is a simple matter. The measurement BAWT was the number of frames (10 msec intervals) between the half-amplitude points of the energy function. It was found that the range for the set of speakers was not large, so the individual ranges were not narrow with respect to it. In addition, the narrowness of the ranges meant that the 10 msec quantization was too coarse. In spite of these factors, BAWT has some capability for

speaker separation, since an F-ratio of 20.7 was obtained.

5.8 Comparison of measurements

The measurements described above are summarized in Table 2, ranked in order of F-ratio. The presence of nine F_o measurements, with high F-ratios at the top of the list does not necessarily mean they should be implemented first in a speaker recognition system, since they are likely to be heavily dependent.

To give some crude examples of the meaning of these F-ratios, if all the speakers had normal distributions with equal variances of σ^2, and if half of them were centered at $-\sigma$ and half of them at $+\sigma$, the resulting F-ratio would be 10; if they fell in four groups, with adjacent means separated by 2σ, the F-ratio would be 50.

An array of the values of the ΔP statistic for all pairs of these measurements is given in Table 3. As may be expected, the ΔP values for the F_o measurements are generally much greater than zero, with the notable exception of 5F02 and 5F03. Most other pairs, such 3F01 and PREV, have much smaller values. (Regarding the many nasal measurements, it was found that measurements from adjacent channels were highly dependent, but the dependence decreased as the comparison progressed to more distant channels. Compare 3M1 vs. 3M6 and 3M1 vs. 3M17.) There is presently no statistical basis for setting a threshold value on ΔP, but for

TABLE 2. ACOUSTIC MEASUREMENTS RANKED BY F-RATIO

	Name	F-ratio
1.	5F02	84.9
2.	5F01	81.0
3.	AEF0	72.2
4.	3F02	71.2
5.	5F04	69.5
6.	3F01	61.8
7.	5F03	54.3
8.	3F05	52.8
9.	3F04	51.8
10.	AEF2	46.6
11.	UHF2	44.6
12.	3M1	43.4
13.	3N18	41.0
14.	UM1A	36.3
15.	IU3	34.4
16.	IS2	32.7
17.	3N8	32.5
18.	3F03	30.9
19.	3M6	28.4
20.	3M17	24.8
21.	3N23	24.4
22.	AF1	22.9
23.	3M23	21.7
24.	UHF1	21.1
25.	BAVT	20.7
26.	AF2	19.0
27.	SH	17.5
28.	AEF1	15.5
29.	PREV	14.5
30.	AS2	11.8
31.	AU3	10.2

TABLE 3. ΔT FOR PAIRS OF MEASUREMENTS (Decimal point omitted. Read -03 as -0.03.)

	3F01	3F02	3F03	3F04	3F05	5F01	5F02	5F03	5F04	AEF0	3M1	3M6	3M17	3M23	3N8	3N18	3N23	IS2	IU3	AS2	AU3	UHF1	UHF2	AEF1	AEF2	AF1	AF2	UM1A	SH	PREV	BAWT
3F01		(98)	(32)	(79)	(52)	(27)	(36)	(41)	(58)	(44)	-08	-04	-13	-07	03	03	-07	-14	00	-01	-02	-06	-14	05	-01	-04	-01	-01	03	-01	-07
3F02	(98)		(36)	(67)	(69)	(43)	(45)	(47)	(57)	(34)	-16	-08	-14	-10	03	-02	-08	-21	-05	-01	-09	-05	-14	02	08	-03	-06	-04	-04	-01	-11
3F03	(32)	(36)		(35)	(67)	(74)	(73)	(25)	(51)	(65)	-08	-09	-08	02	05	-07	03	-03	-02	00	01	-07	-06	02	16	00	03	-08	-05	00	-07
3F04	(79)	(67)	(35)		(91)	(63)	(70)	(49)	(72)	(59)	-15	06	-04	-07	21	10	03	-23	06	00	-01	02	-03	06	07	-08	05	02	-00	03	-07
3F05	(52)	(69)	(67)	(91)		(63)	(65)	(41)	(92)	(73)	-05	02	-04	03	16	-02	01	-17	06	03	-07	04	-05	05	(29)	01	01	-04	-02	01	-07
5F01	(27)	(43)	(74)	(63)	(63)		(195)	16	(92)	(69)	-07	-04	-02	10	09	-13	01	-00	-02	-03	-05	01	-17	01	(29)	-01	-00	-03	-02	-02	-06
5F02	(36)	(45)	(73)	(70)	(65)	(195)		15	93	(87)	-06	-03	-05	09	07	-15	00	-02	01	-03	-05	-03	-16	-03	21	-02	00	-07	-02	-02	-07
5F03	(41)	(47)	(25)	(49)	(41)	16	15		24	23	-15	00	01	05	02	07	04	-04	04	05	-06	-03	11	01	06	-06	-02	-00	-05	01	-12
5F04	(58)	(57)	(51)	(72)	(92)	(92)	(93)	24		(60)	-02	05	03	05	09	-04	-08	-05	03	-03	-01	07	-11	01	12	06	-03	05	-04	-03	-01
AEF0	(44)	(34)	(65)	(59)	(73)	(69)	(87)	23	(60)		-06	02	00	07	12	-08	02	00	-01	-03	03	-04	-11	-04	10	-05	08	-11	05	-02	-06
3M1	-08	-16	-08	-15	-05	-07	-06	-15	-02	-06		(31)	10	10	01	-00	04	06	00	03	06	04	-03	02	03	02	-01	-05	05	-02	04
3M6	-04	-08	-09	06	02	-04	-03	00	05	02	(31)		10	03	13	00	02	-04	01	-01	-03	04	-01	02	-06	-08	05	-01	11	-02	-01
3M17	-13	-14	-08	-04	-04	-02	-05	01	03	00	10	10		15	09	15	00	05	17	-02	01	10	01	00	-04	06	03	02	03	01	00
3M23	-07	-10	02	-07	03	10	09	05	05	07	10	03	15		01	11	18	19	04	-03	07	12	01	-03	07	-01	-04	01	01	-04	-04
3N8	03	03	05	21	16	09	07	02	09	12	01	13	09	01		11	13	-14	-02	04	06	-04	-01	11	-03	05	03	02	03	-00	-02
3N18	03	-02	-07	10	-02	-13	-15	07	-04	-08	-00	00	15	11	11		19	02	04	02	-01	14	07	-01	06	-10	-08	10	-02	-02	-02
3N23	-07	-08	03	03	01	01	00	04	-08	02	04	02	00	18	13	19		-01	-02	03	06	-00	04	01	11	07	-02	-02	-02	-02	-02
IS2	-14	-21	-03	-23	-17	-00	-02	-04	-05	00	06	-04	05	19	-14	02	-01		09	-01	04	14	-04	-10	-01	21	-02	-07	07	00	-03
IU3	00	-05	-02	-06	06	-02	01	04	03	-01	06	01	17	04	-02	04	-02	09		02	-04	07	-05	-00	04	-01	-04	-04	00	-01	05
AS2	01	-01	00	00	03	-03	-03	05	-03	-03	03	-01	-02	-03	04	03	02	-01	02		04	-00	-01	05	01	10	-03	03	01	03	03
AU3	-02	-09	01	-01	-07	-05	-05	-06	-01	03	06	-03	01	07	06	-01	06	04	-04	04		-04	-01	-01	07	09	-01	01	03	01	02
UHF1	-06	-05	-07	02	04	01	-03	-03	07	-04	04	04	10	12	-04	14	-00	14	07	-00	-04		-01	07	03	00	-06	-03	-02	-03	03
UHF2	-14	-14	-06	-03	-05	-17	-16	11	-11	-11	-03	-01	01	01	-01	07	04	-04	-05	-01	-01	-01		-04	23	-05	-03	-01	02	05	06
AEF1	05	02	02	06	05	01	-03	01	01	-04	02	02	00	-03	11	-01	01	-10	-00	05	-01	07	-04		-05	09	-06	05	-04	-01	01
AEF2	-01	08	16	07	(29)	(29)	21	06	12	10	03	-06	-04	07	-03	06	11	-01	04	01	07	03	23	-05		03	02	-07	-06	02	-03
AF1	-04	-03	-00	-08	01	-01	-02	-06	06	-05	02	-08	06	-01	05	-10	-07	21	-01	10	09	00	-05	09	03		-02	07	03	07	-01
AF2	-01	-06	-03	05	01	-00	00	-02	-03	08	-01	05	03	-04	03	-08	-02	-02	-04	-03	-01	-06	-03	-06	02	-02		-03	01	02	-04
UM1A	-01	-04	-08	02	-04	-03	-07	-00	05	-11	-05	-01	02	01	02	10	-02	-07	-04	03	01	-03	-01	05	-07	07	-03		-03	-03	-00
SH	03	-04	-05	-00	-02	-02	-02	-04	-04	05	05	11	03	01	03	-02	-02	07	00	01	03	-02	02	-04	-06	03	01	-03		-00	04
PREV	-01	-01	00	03	01	-02	-02	01	-03	-02	-02	-02	-01	-04	-00	-02	-02	00	-01	03	01	-03	05	-01	02	07	02	-03	-00		03
BAWT	-07	-11	-07	-07	-07	-06	-07	-12	-01	-06	04	-01	00	-04	-02	02	-02	-03	05	03	02	03	06	01	-03	-01	-04	-00	04	03	

illustrative purposes, all values greater than or equal to 0.25 have been circled. On that basis, the only measurement pairs which are somewhat dependent are the F_o measurements, AEF2 and two of the F_o measurements, and 3M1 and 3M6. Pairwise independence (in the loose sense we have been using the term) does not guarantee total mutual independence, but again, it will probably suffice for the purpose of avoiding heavily redundant measurements.

In order to determine whether the set of speakers used in this experiment was biased by the inclusion of speech researchers, who might tend to speak in a particularly consistent way that would render them easy to identify, the speaker set was divided into two groups. One group, of 10 members, were those who were substantially concerned with speech research; the other group of 11 were not. For each of the 20 measurements listed in Table 4, the relative variance (i.e. individual variance divided by the total variance) for each speaker was tabulated and averaged for that speaker. Then the averages of each group were computed and compared. There was no significant difference in the identifiability of the two groups as given by the relative variance.

5.9 Identification results

Although the actual construction of a speaker recognition system was not a primary aim of this study, the temptation of finding out whether these measurements actually "work"

was too great to resist. Accordingly an elementary speaker identification algorithm was programmed in FORTRAN IV on the PDP-9. Twenty measurements were selected from Table 2, choosing those with the highest F-ratio, but rejecting those with significant dependence (arbitrarily, $\Delta P \geq 0.25$) on the measurements already selected. This set of measurements is listed in Table 4.

The data, which consisted of ten repetitions by each speaker, was partitioned into design and test sets. The design set was used to form references for each speaker, by calculating the mean and variance for each measurement for each speaker; the test data was used to test the effectiveness of these references in characterizing the individual speakers. This testing was done with data that had no role in the determination of the references. In order to make full use of the available data, each of the ten repetitions was used in turn as the test set, while the remaining nine were used to form the references.

The classification algorithm was a minimum distance procedure using a weighted Euclidean distance metric similar to that used by Pruzansky and Mathews (1964). If r measurements are used, each datum is represented by a point in an r-dimensional space. The average of the nine repetitions for each speaker in the design set is the centroid of those nine points. The square of the distance between a datum $\overline{x} = (x_1, x_2, \ldots, x_r)$ and the centroid of the j-th class $\overline{\mu_j} = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jr})$

TABLE 4. MEASUREMENTS SELECTED FOR IDENTIFICATION EXPERIMENTS

	Name	F-ratio
1.	5FO2	84.9
2.	5FO3	54.3
3.	AEF2	46.6
4.	3M1	43.4
5.	3N18	41.0
6.	UM1A	36.3
7.	IU3	34.4
8.	IS2	32.7
9.	3N8	32.5
10.	3M17	24.8
11.	3N23	24.4
12.	AF1	22.9
13.	3M23	21.7
14.	UHF1	21.1
15.	BAWT	20.7
16.	AF2	19.0
17.	SH	17.5
18.	AEF1	15.5
19.	PREV	14.5
20.	AS2	11.8

is given by

$$d^2(\bar{x},\bar{w}_j) = \sum_{k=1}^{r} \frac{(x_k - \mu_{jk})^2}{\langle \sigma^2_{jk} \rangle_j}$$

where $\langle \sigma^2_{jk} \rangle_j$ is the average speaker variance for the k-th measurement of the reference data. Dividing the squared distance in each dimension by the average speaker variance weights it according to the average narrowness of the individual speaker distributions. The distance to the centroid of each speaker is computed, and the test datum is associated with the speaker whose centroid is closest. Although this algorithm is non-probabilistically motivated, it is in fact the optimum classification procedure for speakers that are a priori equally likely and measurements that are independent Gaussian random variables with equal variances for each speaker (Nilsson, 1965).

When the first 17 measurements in Table 4 were used, no identification errors were made in the classification of 10 repetitions by each of 21 speakers. If the measurements are selected by trial and error rather than systematically by a priori evaluation as was done here, perfect recognition can be achieved with fewer measurements, since the F-ratio is not an optimizing statistic. However, its usefulness, along with that of the ΔP statistic, is demonstrated by the success achieved here with a computationally simple classifier and a small number of effective measurements.

CHAPTER 6

CONCLUSION

This study has been directed toward the improvement of speaker recognition techniques by means of improving the characterizing measurements made on the voice signal. The approach adopted here makes specific measurements on speech events which have been segmented and located in the utterance. The choices of the phonetic segments and the measurements made on them are guided by considerations of vocal tract structure and the ways in which the various speech sounds are produced. The final selection of measurements is aided by techniques of evaluating the speaker separating ability and the interdependence of the measurements.

For the conditions of this experiment, measurements of fundamental frequency proved to be the most useful single measurements investigated. They were generally interdependent, so other, independent measurements were usually preferable to multiple F_o measurements. Most of the other measurements were not heavily dependent on each other. Nasals were characterized by certain individual filter outputs. Formant measurements of the vowels that were studied were useful, as were spectrum shape parameters in cases where formant locations were difficult to measure. The wider interspeaker variation of the second formant made that one generally better than the first formant. A rough estimation of the glottal source spectrum slope was also effective. The information

conveyed by the measurements of duration of the word bought, shape of /ʒ/ spectrum, and prevoicing was limited by their coarseness of quantization, but they also proved to be useful.

The validity of this selective and efficient approach to acoustic measurements for speaker recognition is demonstrated by the success achieved in speaker identification with a small number of such measurements and a simple linear classification procedure. A direct comparison of results by different workers is usually not possible due to different sets of constraints placed on the problem. Subjectively, however, the result achieved here compares very favorably with the reports in the current literature.

The set of measurements developed here cannot be called optimum, since only a relatively small number of possible measurements were investigated. Extended research will probably produce more independent acoustic measurements with equivalent F-ratios at least in the 40's and 50's. There are several specific areas that should be good candidates for such extensions:

1. The spectra of vowels should yield more useful data. The improvement and automation of the analysis-by-synthesis technique would be a great aid for providing fast, reliable formant measurements. The recently introduced chirp z-transform algorithm (Rabiner, et al., 1969) should also be useful.

2. The nasal consonants should bear a closer analysis so that a more satisfactory means of characterizing them can be devised. As indicated in Chapter 5, the pole and zero locations are probably the significant factors. Analysis-by-synthesis, as used by Fujimura (1962), if amenable to automation in this more complex case, may prove useful in this respect.

3. Further investigation of the laryngeal excitation characteristics should be done. Acceptable automatic inverse filtering might be accomplished by means of parameters derived by automated analysis-by-synthesis. Perturbations in pitch period may also be characteristic of individual larynges.

4. A largely untapped area is that of temporal patterns in the speech signal. This area includes effects such as rate and extent of formant transitions, the coordination of different articulators, and durations of certain segments. An interesting problem is that of normalizing temporal patterns for the rate of speech. Temporal patterns are admittedly more difficult to characterize than some spectral patterns, but they must contain much information about learned characteristics.

Both present and future measurements must be subjected to close scrutiny in terms of their stability with respect

to time and the state of health of the speaker. The influence of the emotional state of a speaker and the effect on a person's learned characteristics of moving to a region where a different dialect is spoken are not known. The susceptibility of speaker recognition measurements to voice mimicry and disguise should also be investigated. There is also a need for simultaneous investigation of vocal tract anatomy and acoustic characteristics of different speakers.

Another speaker recognition paradigm which has already arisen in law enforcement situations may be called uncontrolled speaker verification. The only acoustic evidence available is two speech samples, and the only question is whether they were uttered by the same speaker. In this case, it may not be possible to form a set of reference patterns from many repetitions of utterances. Different techniques may be required.

The measurements described here were done on phones in single, fixed contexts. This method can be eventually generalized to any context, or at least to a subset of contexts. Then it is conceivable that future automatic speaker recognizers with advanced speech recognition capability will be able to extract the necessary measurements from arbitrary context.

APPENDIX I

EQUIVALENCE OF AVERAGE RELATIVE VARIANCE AND F-RATIO

Let x_{ij} denote the measurement datum of the i-th repetition by the j-th speaker, i=1,2,...,n, j=1,2,...,m. Let $\langle\ \rangle_k$ denote the average over the subscript k. Let μ_j and σ_j^2 denote the mean and variance of the data of the j-th speaker. Let $\bar{\mu}$ and σ_{tot}^2 denote the mean and variance of the data pooled over all the speakers.

Let α denote the average relative variance, and let F denote the F-ratio.

$$\alpha = \frac{\langle \sigma_j^2 \rangle_j}{\sigma_{tot}^2}$$

$$F = \frac{n[\ \mathrm{Var}(\mu_j)\ ]}{\langle \sigma_j^2 \rangle_j}$$

$$\mathrm{Var}(\mu_j) = \langle \mu_j^2 \rangle_j - \bar{\mu}^2$$

$$= \langle\langle x_{ij}^2 \rangle_i - \sigma_j^2 \rangle_j - \bar{\mu}^2$$

$$= \langle x_{ij}^2 \rangle_{i,j} - \langle \sigma_j^2 \rangle_j - \bar{\mu}^2$$

$$= \langle x_{ij}^2 \rangle_{i,j} - \bar{\mu}^2 - \langle \sigma_j^2 \rangle_j$$

$$= \sigma_{tot}^2 - \langle \sigma_j^2 \rangle_j$$

Hence

$$F = \frac{n[\ \sigma_{tot}^2 - \langle \sigma_j^2 \rangle_j]}{\langle \sigma_j^2 \rangle_j} = n\left(\frac{1}{\alpha} - 1\right)$$

Since the F-ratio is a monotonic function of the average relative variance, ranking measurements in decreasing order of F-ratios is equivalent to ranking them in increasing order of average relative variances.

APPENDIX II

SUMMARIES OF MEASUREMENTS

This appendix contains summaries of the individual measurements described in Chapter 5. Each summary represents 10 repetitions of the utterance by each of 21 speakers. "Total sigma" is the estimated standard deviation of the data pooled over all speakers. "Rsigma" is the ratio of the individual estimated standard deviation to the pooled standard deviation, or the square root of the relative variance.

For the purposes of format or representation as integer variables, certain measurements were subjected to translation or scale change. The following measurements had 1000 Hz subtracted from them: UHF2, AEF2, and AF2. The following measurements were multiplied by 10: IS2, IU3, AS2, AU3, and UM1A.

MEASUREMENT 3F01

TOTAL RANGE= 65
TOTAL SIGMA= 13.43
AVG. SIGMA= 4.83
AVG. RSIGMA=2.362
F-RATIO= 61.8

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	10	123.6	2.95	0.223
2	7	118.3	2.26	0.168
3	13	121.2	3.94	0.293
4	14	110.8	5.43	0.403
5	38	114.9	11.84	0.822
6	11	91.6	3.57	0.265
7	17	129.1	5.22	0.387
8	13	102.8	3.77	0.280
9	10	130.2	3.22	0.242
10	15	109.9	5.34	0.398
11	15	112.1	4.61	0.343
12	21	137.4	5.56	0.414
13	14	113.0	4.37	0.325
14	13	115.5	4.43	0.330
15	21	101.9	5.88	0.438
16	13	143.6	4.20	0.312
17	11	94.7	3.59	0.267
18	13	117.4	4.26	0.302
19	26	112.7	7.01	0.522
20	17	102.0	5.16	0.384
21	19	122.4	5.82	0.433

MEASUREMENT 3F02

TOTAL RANGE= 76
TOTAL SIGMA= 13.88
AVG. SIGMA= 4.70
AVG. RSIGMA=2.338
F-RATIO= 71.2

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	8	115.3	2.71	0.195
2	8	118.5	2.82	0.202
3	19	120.3	5.89	0.424
4	15	112.8	4.29	0.309
5	25	118.7	7.96	0.573
6	17	91.7	5.21	0.361
7	9	107.8	3.55	0.256
8	8	102.7	2.71	0.195
9	11	129.3	3.68	0.265
10	7	117.2	2.20	0.159
11	17	112.5	4.62	0.333
12	18	146.3	6.96	0.501
13	19	116.5	5.42	0.390
14	15	118.7	5.19	0.374
15	24	103.2	8.70	0.627
16	12	145.8	3.68	0.265
17	19	100.3	5.81	0.419
18	17	113.8	5.03	0.362
19	9	117.0	3.71	0.267
20	10	104.9	3.14	0.226
21	20	127.7	5.62	0.405

MEASUREMENT 3F03

TOTAL RANGE= 99
TOTAL SIGMA= 17.33
AVG. SIGMA= 8.17
AVG. RSIGMA=3.471
F-RATIO= 32.9

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	21	135.7	6.77	0.398
2	19	143.2	6.41	0.372
3	19	154.3	7.47	0.431
4	29	128.8	8.42	0.485
5	61	158.8	18.24	1.053
6	24	123.3	7.10	0.410
7	14	132.4	3.72	0.215
8	32	133.5	8.77	0.505
9	28	155.2	9.10	0.525
10	18	145.3	6.31	0.364
11	21	131.6	7.29	0.421
12	53	179.8	15.05	0.868
13	36	142.2	12.73	0.735
14	22	135.6	6.31	0.364
15	25	134.2	8.12	0.469
16	17	169.8	5.73	0.331
17	24	130.1	7.65	0.441
18	28	133.9	8.80	0.508
19	16	141.3	4.67	0.269
20	19	112.8	5.09	0.294
21	22	143.2	7.79	0.449

MEASUREMENT 3F04

TOTAL RANGE= 74
TOTAL SIGMA= 15.81
AVG. SIGMA= 5.74
AVG. RSIGMA=2.363
F-RATIO= 51.8

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	14	117.3	4.60	0.291
2	7	137.8	2.48	0.157
3	25	134.4	8.41	0.532
4	12	124.6	4.01	0.253
5	14	131.2	5.67	0.359
6	8	102.2	2.52	0.155
7	30	123.5	10.66	0.674
8	7	109.3	2.83	0.179
9	39	150.5	13.83	0.875
10	16	129.1	4.20	0.266
11	15	113.7	4.35	0.275
12	18	152.2	6.00	0.379
13	24	127.5	7.28	0.460
14	6	115.6	2.01	0.127
15	32	116.4	12.12	0.767
16	22	149.8	6.23	2.394
17	16	102.3	4.64	0.294
18	12	126.2	3.79	0.240
19	22	120.9	6.69	0.423
20	13	103.9	3.63	0.230
21	12	127.6	4.45	0.282

MEASUREMENT 3F05

TOTAL RANGE= 71
TOTAL SIGMA= 14.84
AVG. SIGMA= 5.42
AVG. RSIGMA=0.365
F-RATIO= 52.8

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	17	124.1	5.07	0.341
2	8	132.0	2.94	0.198
3	18	131.9	5.57	0.375
4	6	113.3	1.83	0.123
5	25	139.1	8.20	0.552
6	17	111.8	5.75	0.387
7	11	117.1	3.41	0.230
8	13	112.0	4.00	0.270
9	39	150.7	12.81	0.863
10	15	132.6	4.45	0.300
11	12	124.0	4.16	0.281
12	21	158.2	8.08	0.544
13	21	128.5	5.88	0.464
14	10	120.2	3.68	0.248
15	32	120.1	11.70	0.788
16	8	141.8	3.08	0.208
17	13	106.9	4.12	0.278
18	16	121.6	5.23	0.353
19	17	130.0	6.11	0.412
20	8	104.4	2.41	0.163
21	13	132.6	4.27	0.288

MEASUREMENT 5F01

TOTAL RANGE= 125
TOTAL SIGMA= 23.74
AVG. SIGMA= 7.41
AVG. RSIGMA=0.312
F-RATIO= 81.0

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	19	137.2	5.72	0.241
2	10	149.9	3.51	0.148
3	45	167.5	13.49	0.568
4	15	137.0	5.52	0.232
5	29	192.9	8.35	0.351
6	23	131.0	7.07	0.298
7	15	134.0	5.33	0.225
8	19	141.2	5.73	0.241
9	27	174.5	9.20	0.388
10	28	162.7	8.35	0.352
11	18	137.8	6.07	0.256
12	36	214.0	12.05	0.508
13	17	144.2	5.81	0.245
14	9	145.8	3.46	0.146
15	33	135.3	10.11	0.426
16	34	182.4	10.53	0.444
17	26	128.1	7.99	0.337
18	43	140.5	14.34	0.604
19	17	156.9	5.63	0.237
20	7	127.5	2.12	0.089
21	15	150.0	5.25	0.221

MEASUREMENT 5F02

TOTAL RANGE= 126
TOTAL SIGMA= 24.06
AVG. SIGMA= 7.44
AVG. RSIGMA=0.309
F-RATIO= 84.9

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	17	144.8	4.73	0.197
2	10	152.5	3.87	0.161
3	45	170.7	12.37	0.514
4	16	133.5	5.70	0.237
5	37	198.0	10.23	0.427
6	23	134.1	5.79	0.242
7	14	137.7	4.76	0.198
8	21	142.9	6.54	0.272
9	28	177.1	9.50	0.395
10	26	165.3	7.98	0.332
11	18	143.2	7.25	0.301
12	37	217.3	12.88	0.535
13	19	144.9	6.33	0.263
14	18	152.9	5.22	0.217
15	27	148.5	8.77	0.365
16	29	187.6	8.95	0.372
17	27	128.9	7.77	0.323
18	40	143.0	13.83	0.575
19	15	160.1	5.38	0.224
20	8	129.8	2.45	0.102
21	13	154.5	4.79	0.199

MEASUREMENT 5F03

TOTAL RANGE= 60
TOTAL SIGMA= 12.67
AVG. SIGMA= 4.66
AVG. RSIGMA=0.367
F-RATIO= 54.3

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	10	109.1	3.21	0.253
2	13	121.9	3.95	0.312
3	12	110.3	4.22	0.333
4	6	98.5	1.98	0.156
5	16	103.2	4.99	0.394
6	5	83.2	1.40	0.110
7	11	104.6	3.50	0.276
8	21	94.6	6.05	0.478
9	21	118.3	5.86	0.473
10	13	98.2	3.74	0.295
11	15	107.4	4.65	0.367
12	33	120.5	9.97	0.787
13	21	110.5	6.36	0.502
14	9	110.4	2.91	0.230
15	25	102.8	8.05	0.635
16	13	134.9	4.72	0.373
17	26	92.4	8.72	0.688
18	9	105.9	2.69	0.212
19	9	107.8	3.08	0.243
20	8	91.6	3.20	0.253
21	12	118.4	4.40	0.347

MEASUREMENT 5FM4

TOTAL RANGE= 72
TOTAL SIGMA= 14.84
AVG. SIGMA= 4.93
AVG. RSIGMA=0.332
F-RATIO= 59.5

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	8	116.0	2.98	0.281
2	9	131.3	2.26	0.152
3	13	128.0	4.78	0.322
4	8	112.9	2.92	0.197
5	34	146.4	10.35	0.697
6	19	105.9	5.59	0.376
7	14	109.4	4.62	0.312
8	9	106.9	2.77	0.186
9	15	133.8	4.18	0.282
10	27	124.0	8.41	0.566
11	16	111.0	4.22	0.284
12	27	154.5	8.64	0.582
13	12	122.8	3.91	0.263
14	11	115.3	3.55	0.240
15	23	111.2	5.45	0.434
16	20	136.7	5.70	0.384
17	15	102.1	5.47	0.368
18	23	124.7	7.50	0.505
19	9	119.7	2.75	0.185
20	9	101.5	2.32	0.156
21	12	126.9	4.20	0.283

MEASUREMENT AEF0

TOTAL RANGE= 82
TOTAL SIGMA= 18.12
AVG. SIGMA= 6.20
AVG. RSIGMA=0.342
F-RATIO= 72.2

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	17	121.5	5.85	0.323
2	19	138.0	5.37	0.297
3	12	148.8	4.13	0.228
4	11	125.7	4.22	0.233
5	23	166.1	6.17	0.341
6	18	114.0	6.04	0.333
7	16	120.1	4.95	0.273
8	27	124.1	7.92	0.437
9	39	159.4	11.15	0.615
10	25	135.2	6.09	0.336
11	13	124.8	3.91	0.216
12	27	158.6	9.17	0.506
13	19	127.9	6.37	0.351
14	26	122.7	7.29	0.402
15	27	127.5	8.48	0.468
16	16	167.0	5.64	0.311
17	12	118.7	3.33	0.184
18	12	121.2	3.85	0.213
19	17	132.7	6.86	0.379
20	21	110.2	6.49	0.358
21	16	133.9	6.08	0.336

MEASUREMENT 3M1

TOTAL RANGE= 18
TOTAL SIGMA= 3.64
AVG. SIGMA= 1.54
AVG. RSIGMA=0.424
F-RATIO= 43.4

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	3	16.5	0.97	0.267
2	6	21.2	1.94	0.534
3	3	20.4	1.17	0.323
4	3	22.8	1.23	0.284
5	2	16.8	0.63	0.174
6	8	19.4	2.46	0.576
7	4	21.1	1.29	0.354
8	4	21.2	1.55	0.426
9	3	13.4	1.27	0.295
10	6	18.7	1.89	0.519
11	5	18.1	1.60	0.438
12	9	14.3	2.45	0.674
13	5	15.1	1.85	0.529
14	5	12.2	1.40	0.384
15	6	13.1	1.66	0.457
16	6	19.3	1.89	0.519
17	4	13.1	1.10	0.322
18	3	14.3	1.34	0.368
19	3	20.8	1.14	0.312
20	7	21.7	2.31	0.635
21	6	19.7	1.64	0.450

MEASUREMENT 3N6

TOTAL RANGE= 22
TOTAL SIGMA= 3.93
AVG. SIGMA= 1.96
AVG. RSIGMA=0.493
F-RATIO= 28.4

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	7	3.1	1.85	0.471
2	3	6.1	0.99	0.253
3	5	2.8	1.87	0.476
4	3	7.3	0.82	0.209
5	3	9.0	1.15	0.294
6	7	4.5	2.42	0.614
7	9	4.7	2.71	0.689
8	7	7.6	2.07	0.525
9	7	1.7	2.00	0.509
10	6	-0.7	2.36	0.600
11	6	3.2	2.15	0.547
12	11	-2.5	3.17	0.806
13	9	1.0	2.71	0.688
14	3	2.7	1.16	0.295
15	9	-1.2	2.97	0.756
16	7	3.3	1.95	0.495
17	3	-4.6	1.07	0.273
18	5	-1.4	1.58	0.401
19	4	5.0	1.41	0.359
20	8	4.1	2.56	0.650
21	7	0.9	2.13	0.542

MEASUREMENT 3N17

TOTAL RANGE= 30
TOTAL SIGMA= 5.52
AVG. SIGMA= 2.24
AVG. RSIGMA=2.515
F-RATIO= 24.8

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	5	-2.6	2.27	0.412
2	6	1.0	2.07	0.363
3	3	8.1	1.10	0.199
4	7	13.5	2.27	0.412
5	8	18.7	2.71	0.491
6	8	4.5	2.72	0.493
7	5	1.9	2.85	0.516
8	6	9.7	2.54	0.461
9	6	9.9	2.13	0.386
10	10	6.0	3.06	0.554
11	12	11.5	3.60	0.652
12	17	7.1	4.82	0.873
13	7	9.0	2.62	0.476
14	5	6.4	1.58	0.286
15	20	8.2	6.27	1.136
16	11	4.6	3.20	0.581
17	6	6.0	2.05	0.372
18	8	9.1	2.56	0.464
19	11	5.4	3.78	0.685
20	10	9.6	3.41	0.617
21	6	4.2	2.10	0.380

MEASUREMENT 3N23

TOTAL RANGE= 32
TOTAL SIGMA= 5.87
AVG. SIGMA= 3.11
AVG. RSIGMA=2.530
F-RATIO= 21.7

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	5	-6.9	1.52	0.259
2	12	-2.7	3.59	0.611
3	4	11.7	1.34	0.228
4	8	9.9	2.64	0.450
5	5	4.1	1.97	0.335
6	10	1.3	3.06	0.520
7	11	3.0	3.74	0.637
8	12	0.2	3.52	2.599
9	11	-2.4	3.69	0.628
10	11	3.6	3.57	0.607
11	14	2.4	4.35	0.741
12	16	-4.7	4.62	0.786
13	8	-2.2	2.74	0.467
14	9	-1.0	2.67	0.454
15	26	1.3	7.83	1.333
16	9	2.9	3.11	0.529
17	5	-0.2	1.75	0.298
18	9	-4.6	2.46	0.419
19	8	8.8	2.62	0.445
20	9	8.5	2.80	0.476
21	6	-2.7	1.77	0.301

MEASUREMENT 3N8

TOTAL RANGE= 22
TOTAL SIGMA= 4.56
AVG. SIGMA= 2.15
AVG. RSIGMA=0.471
F-RATIO= 32.5

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	5	-4.3	1.42	0.311
2	6	-1.7	1.72	0.373
3	4	-3.1	1.20	0.262
4	10	-1.3	2.71	0.594
5	4	6.9	1.52	0.334
6	12	3.5	3.57	0.782
7	6	-4.1	1.97	0.432
8	7	4.2	2.57	0.564
9	5	-4.7	1.49	0.328
10	7	-4.6	2.07	0.453
11	8	3.0	2.42	0.527
12	4	-5.3	1.34	0.293
13	7	-3.9	2.13	0.467
14	3	1.1	1.10	0.241
15	16	-2.4	4.09	0.896
16	5	-6.2	1.55	0.340
17	8	-8.2	2.74	0.601
18	7	3.8	2.44	0.535
19	9	-2.9	2.96	0.649
20	7	1.5	2.27	0.498
21	5	-5.9	1.85	0.406

MEASUREMENT 3N18

TOTAL RANGE= 28
TOTAL SIGMA= 5.60
AVG. SIGMA= 2.45
AVG. RSIGMA=2.438
F-RATIO= 41.0

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	5	-9.0	1.75	0.315
2	6	-4.1	1.85	0.331
3	5	1.4	1.71	0.306
4	11	7.6	3.06	0.547
5	7	4.3	2.06	2.368
6	4	-4.6	1.26	0.226
7	9	-0.6	3.13	2.560
8	9	-4.5	2.45	2.440
9	8	3.1	2.33	0.416
10	8	-3.4	3.37	0.603
11	10	3.3	2.67	0.477
12	10	0.7	2.79	0.499
13	9	4.9	2.42	0.433
14	4	-9.6	1.35	0.241
15	12	-0.1	3.57	0.638
16	5	-2.8	1.62	0.289
17	7	-3.2	2.30	0.411
18	12	10.2	3.52	0.629
19	10	5.6	3.24	0.579
20	7	3.4	2.37	0.423
21	10	1.2	2.66	0.475

MEASUREMENT 3N23

TOTAL RANGE= 29
TOTAL SIGMA= 6.14
AVG. SIGMA= 3.19
AVG. RSIGMA=2.519
F-RATIO= 24.4

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	8	-8.4	3.03	0.492
2	9	-7.0	2.36	0.384
3	11	2.8	4.16	0.677
4	10	-2.1	3.28	0.534
5	9	-0.2	2.66	0.433
6	7	0.0	2.21	0.360
7	14	4.4	5.04	0.820
8	13	-6.6	3.72	0.605
9	10	-5.9	3.13	0.517
10	12	-8.2	3.85	0.627
11	4	-7.1	1.45	0.236
12	10	-7.8	2.90	0.472
13	8	-5.6	2.72	0.442
14	4	-8.9	1.45	0.236
15	15	-1.0	5.16	0.842
16	4	-6.2	1.69	0.274
17	6	-9.0	2.26	0.368
18	12	9.8	4.42	0.719
19	19	3.5	6.08	0.989
20	9	2.1	2.77	0.450
21	10	-5.5	2.64	0.429

MEASUREMENT IS2

TOTAL RANGE= 92
TOTAL SIGMA= 14.27
AVG. SIGMA= 6.51
AVG. RSIGMA=0.456
F-RATIO= 32.7

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	18	134.2	4.83	0.338
2	36	138.6	11.23	0.787
3	24	166.1	7.43	0.521
4	27	104.2	7.69	0.539
5	11	133.3	3.33	0.233
6	18	133.5	5.89	0.413
7	22	134.2	6.58	0.461
8	22	138.1	7.98	0.559
9	15	123.7	5.46	0.383
10	24	136.1	5.62	0.464
11	30	146.9	11.08	0.777
12	42	143.8	12.26	0.860
13	23	148.8	6.58	0.461
14	13	133.1	4.61	0.323
15	22	159.9	6.59	0.462
16	19	137.6	5.10	0.358
17	10	134.5	3.44	0.241
18	35	132.5	11.12	0.779
19	11	119.8	3.29	0.231
20	7	138.9	2.64	0.185
21	10	142.2	2.90	0.203

MEASUREMENT I03

TOTAL RANGE= 356
TOTAL SIGMA= 77.96
AVG. SIGMA=35.68
AVG. RSIGMA=0.458
F-RATIO= 34.4

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	93	-65.7	29.12	0.374
2	138	-223.3	42.94	0.551
3	160	-186.1	49.92	0.642
4	70	13.2	22.22	0.285
5	93	-75.3	35.77	0.459
6	102	-243.5	33.26	0.427
7	116	-239.3	36.47	0.468
8	215	-72.4	62.62	0.803
9	45	-62.8	14.01	0.180
10	145	-123.3	41.07	0.527
11	218	-125.8	67.51	0.866
12	98	-190.3	35.40	0.454
13	120	-210.3	36.06	0.463
14	57	-205.3	14.87	0.191
15	124	-227.3	33.73	0.433
16	78	-114.7	27.29	0.350
17	48	-146.9	15.53	0.199
18	126	-99.6	39.96	0.513
19	133	-133.1	40.62	0.521
20	132	-87.3	35.43	0.454
21	135	-173.5	35.53	0.456

MEASUREMENT AS2

TOTAL RANGE= 124
TOTAL SIGMA= 15.43
AVG. SIGMA= 8.79
AVG. RSIGMA=2.570
F-RATIO= 11.8

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	23	84.2	6.86	0.445
2	40	73.7	11.74	0.761
3	23	86.9	7.49	0.486
4	19	91.6	6.36	0.412
5	26	80.7	9.08	0.589
6	41	84.9	11.81	0.765
7	16	84.3	5.06	0.328
8	26	101.0	7.82	0.507
9	15	83.6	4.74	0.307
10	28	76.3	9.17	0.594
11	13	80.5	4.30	0.279
12	36	84.9	9.93	0.647
13	17	85.1	5.59	0.362
14	7	91.0	2.31	0.150
15	25	94.1	6.94	0.450
16	30	96.5	11.49	0.745
17	14	80.7	4.30	0.278
18	21	78.6	8.17	0.529
19	22	83.7	6.52	0.422
20	94	130.2	34.41	2.230
21	34	86.7	10.49	0.680

MEASUREMENT AU3

TOTAL RANGE= 138
TOTAL SIGMA= 26.46
AVG. SIGMA=18.35
AVG. RSIGMA=?.693
F-RATIO= 10.2

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	31	-39.2	9.43	0.356
2	89	-47.7	23.02	0.878
3	40	5.3	12.74	0.481
4	67	-35.2	18.72	0.708
5	85	-13.6	24.98	0.940
6	83	-64.1	29.79	1.126
7	73	-58.5	22.19	0.838
8	74	-17.7	26.04	0.984
9	40	-49.5	12.36	0.467
10	54	-35.9	18.57	0.702
11	52	-59.7	17.07	0.645
12	79	-39.5	19.55	0.739
13	43	-27.5	16.33	0.617
14	22	-49.9	6.35	0.240
15	52	-39.2	17.94	0.678
16	101	-47.1	29.99	1.133
17	34	-57.8	11.85	0.448
18	43	-14.4	15.23	0.575
19	56	-55.1	17.99	0.680
20	44	-76.7	15.51	0.586
21	58	-37.7	19.82	0.749

MEASUREMENT UHF1

TOTAL RANGE= 250
TOTAL SIGMA= 50.19
AVG. SIGMA=25.50
AVG. RSIGMA=0.528
F-RATIO= 21.1

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	30	527.0	9.49	0.189
2	20	432.0	6.32	0.126
3	100	481.0	32.13	0.640
4	50	381.0	15.24	0.304
5	80	465.0	24.61	0.490
6	80	471.0	27.67	0.551
7	40	496.0	12.65	0.252
8	40	485.0	14.34	0.286
9	80	483.0	30.57	0.609
10	140	459.0	50.43	1.005
11	120	442.0	42.64	0.849
12	120	522.0	43.41	0.865
13	60	421.0	26.21	0.518
14	60	505.0	15.81	0.315
15	122	442.0	35.53	0.708
16	80	494.0	32.39	0.645
17	20	482.0	10.33	0.206
18	80	412.0	28.98	0.577
19	50	567.0	17.67	0.352
20	110	444.0	33.73	0.672
21	140	452.0	46.62	0.929

MEASUREMENT UHF2

TOTAL RANGE= 590
TOTAL SIGMA=117.62
AVG. SIGMA=49.24
AVG. RSIGMA=?.419
F-RATIO= 44.6

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	150	542.0	45.66	0.388
2	165	673.5	52.97	0.450
3	120	496.?	37.77	0.321
4	120	433.0	31.99	?.272
5	120	602.0	32.93	0.280
6	100	658.0	34.25	?.291
7	205	639.0	60.77	?.517
8	180	498.?	55.14	0.469
9	265	614.5	69.94	0.595
10	120	510.?	33.67	0.286
11	82	386.0	25.91	?.220
12	220	572.0	71.82	?.610
13	120	431.0	35.42	?.301
14	140	341.?	44.58	?.379
15	200	553.0	58.51	?.497
16	165	735.?	59.35	?.505
17	230	409.0	74.90	?.637
18	82	615.0	23.69	?.201
19	210	712.5	69.05	?.587
20	140	539.0	49.09	?.417
21	210	606.0	66.70	?.567

MEASUREMENT AEF1

TOTAL RANGE= 267
TOTAL SIGMA= 52.75
AVG. SIGMA=32.20
AVG. RSIGMA=?.610
F-RATIO= 15.5

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	84	698.8	27.41	?.523
2	38	584.8	12.97	?.246
3	128	626.6	29.77	?.564
4	90	658.?	26.93	?.511
5	126	652.3	39.42	?.747
6	132	656.9	43.75	?.829
7	86	692.8	30.75	?.583
8	64	609.5	17.61	?.334
9	126	701.7	35.20	?.667
10	112	705.4	48.73	?.772
11	118	611.5	38.53	?.731
12	139	735.6	46.80	?.887
13	120	679.1	32.33	?.622
14	59	647.1	18.92	?.359
15	103	612.1	37.57	?.712
16	56	622.5	17.84	?.338
17	91	673.6	25.98	?.491
18	72	604.9	23.55	?.447
19	165	682.3	44.25	?.839
20	81	585.1	25.22	?.478
21	204	648.1	60.25	1.142

MEASUREMENT AEF2

TOTAL RANGE= 488
TOTAL SIGMA=134.02
AVG. SIGMA=42.35
AVG. RSIGMA=0.437
F-RATIO= 46.6

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	85	591.8	29.08	0.280
2	160	697.8	48.54	0.467
3	149	697.9	42.38	0.407
4	77	579.7	25.95	0.249
5	93	698.7	27.64	0.266
6	112	870.9	40.95	0.394
7	116	825.2	32.65	0.314
8	127	594.4	49.93	0.480
9	143	785.5	55.12	0.530
10	192	707.6	57.23	0.550
11	153	537.3	47.32	0.455
12	223	551.4	82.98	0.798
13	158	577.2	45.08	0.433
14	63	589.1	20.86	0.200
15	142	719.6	49.16	0.473
16	195	773.5	68.82	0.662
17	82	579.8	23.45	0.225
18	107	602.2	36.08	0.347
19	127	763.3	37.52	0.361
20	102	743.1	29.37	0.282
21	131	666.3	39.42	0.379

MEASUREMENT AF1

TOTAL RANGE= 388
TOTAL SIGMA= 70.23
AVG. SIGMA=36.98
AVG. RSIGMA=0.527
F-RATIO= 22.9

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	169	809.3	45.92	0.654
2	109	774.1	37.34	0.532
3	59	731.3	20.82	0.296
4	95	726.8	32.58	0.464
5	104	828.0	34.45	0.491
6	203	792.5	59.46	0.847
7	113	817.5	37.70	0.537
8	62	666.8	17.25	0.245
9	142	859.8	40.64	0.579
10	151	756.2	41.39	0.589
11	137	684.4	45.72	0.651
12	242	828.8	77.66	1.106
13	82	748.2	24.17	0.344
14	62	736.0	19.60	0.279
15	70	713.3	21.70	0.309
16	182	724.0	53.29	0.759
17	121	809.0	35.06	0.499
18	185	709.7	53.74	0.765
19	110	836.8	30.69	0.437
20	89	644.2	27.21	0.388
21	57	804.8	20.23	0.288

MEASUREMENT AF2

TOTAL RANGE= 254
TOTAL SIGMA= 45.52
AVG. SIGMA=26.19
AVG. RSIGMA=2.575
F-RATIO= 19.0

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	117	184.2	34.82	0.765
2	73	129.9	24.12	0.530
3	149	126.9	44.12	0.969
4	33	141.6	12.55	0.276
5	84	255.7	29.73	0.653
6	103	235.8	31.71	0.697
7	123	165.4	39.47	0.867
8	94	154.3	32.59	0.674
9	80	180.6	23.21	0.510
10	56	141.8	20.98	0.461
11	31	135.2	10.37	0.228
12	63	186.6	17.53	0.385
13	82	235.5	26.98	0.593
14	149	143.3	38.44	0.845
15	81	200.6	25.07	0.551
16	89	203.5	25.88	2.569
17	49	188.5	16.03	0.352
18	57	154.3	17.91	0.394
19	91	176.5	29.08	0.639
20	94	191.9	30.87	0.678
21	68	231.8	20.48	0.450

MEASUREMENT UM1A

TOTAL RANGE= 39
TOTAL SIGMA= 8.36
AVG. SIGMA= 3.82
AVG. RSIGMA=0.457
F-RATIO= 36.3

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	12	8.2	4.39	2.526
2	16	27.9	5.40	0.647
3	7	2.1	2.23	0.267
4	8	12.2	2.70	0.323
5	13	14.1	3.84	0.462
6	16	17.5	4.74	0.568
7	10	19.3	3.13	0.374
8	8	29.8	2.49	0.297
9	12	13.3	4.03	0.482
10	16	22.3	5.19	0.621
11	22	18.3	6.15	0.736
12	15	18.9	4.58	0.548
13	8	8.9	3.21	0.384
14	4	1.5	1.51	0.181
15	14	12.9	4.58	0.548
16	10	6.3	3.56	0.426
17	12	7.7	3.68	0.441
18	14	5.2	5.35	0.640
19	12	11.4	3.81	0.455
20	11	10.5	3.27	0.392
21	7	6.5	2.32	0.278

MEASUREMENT SH

TOTAL RANGE= 3
TOTAL SIGMA= 0.90
AVG. SIGMA= 0.42
AVG. RSIGMA=0.461
F-RATIO= 17.5

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	2	1.5	0.71	0.782
2	2	1.9	0.99	1.099
3	0	2.0	0.00	0.000
4	1	2.8	0.42	0.466
5	2	2.5	0.85	0.940
6	2	1.4	0.70	0.773
7	1	1.6	0.52	0.571
8	0	1.0	0.00	0.000
9	0	1.0	0.00	0.000
10	0	2.0	0.00	0.000
11	0	1.0	0.00	0.000
12	1	1.3	0.48	0.534
13	1	1.3	0.48	0.534
14	0	1.0	0.00	0.000
15	0	1.0	0.00	0.000
16	2	2.1	0.99	1.099
17	0	4.0	0.00	0.000
18	2	1.4	0.84	0.932
19	2	1.8	1.03	1.142
20	1	1.2	0.42	0.466
21	1	1.1	0.32	0.350

MEASUREMENT PREV

TOTAL RANGE= 1
TOTAL SIGMA= 0.42
AVG. SIGMA= 0.18
AVG. RSIGMA=0.424
F-RATIO= 14.5

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	0	0.0	0.00	0.000
2	0	0.0	0.00	0.000
3	1	0.5	0.53	1.252
4	0	0.0	0.00	0.000
5	0	0.0	0.00	0.000
6	0	0.0	0.00	0.000
7	1	0.1	0.32	0.751
8	1	0.8	0.42	1.002
9	0	0.0	0.00	0.000
10	1	0.1	0.32	0.751
11	0	1.0	0.00	0.000
12	1	0.1	0.32	0.751
13	0	0.0	0.00	0.000
14	0	0.0	0.00	0.000
15	0	0.0	0.00	0.000
16	1	0.8	0.42	1.002
17	1	0.7	0.48	1.148
18	1	0.2	0.42	1.002
19	0	0.0	0.00	0.000
20	1	0.5	0.53	1.252
21	0	0.0	0.00	0.000

MEASUREMENT BAWT

TOTAL RANGE= 9
TOTAL SIGMA= 1.50
AVG. SIGMA= 0.84
AVG. RSIGMA=0.563
F-RATIO= 22.7

SPKR	RANGE	MEAN	SIGMA	RSIGMA
1	3	19.2	0.92	0.614
2	2	18.9	0.74	0.493
3	3	15.9	1.10	0.736
4	3	18.3	0.82	0.550
5	3	18.4	0.97	0.646
6	5	17.5	1.27	0.849
7	2	16.2	0.79	0.527
8	2	16.9	0.74	0.493
9	2	17.6	0.70	0.467
10	2	17.8	0.63	0.423
11	3	15.6	0.97	0.646
12	2	16.1	0.74	0.493
13	2	15.5	0.71	0.473
14	1	19.8	0.42	0.282
15	5	19.4	1.51	1.006
16	3	17.7	0.82	0.550
17	1	16.4	0.52	0.345
18	2	17.8	0.79	0.527
19	2	16.6	0.70	0.467
20	4	16.8	1.23	0.822
21	2	17.8	0.63	0.423

REFERENCES

Atal, B.S. (1968), Automatic Speaker Recognition Based on Pitch Contours, Ph.D. Thesis, Department of Electrical Engineering, Polytechnic Institute of Brooklyn.

Becker, M.E., R. Gnanadesikan, M.V. Mathews, R.S. Pinkham, S. Pruzansky, and M.B. Wilk (1964), "Comparisons of Some Statistical Distance Measures for Talker Identication," J. Acoust. Soc. Am., 36, 1988 (Abstract).

Bell, C.G., H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques, J. Acoust. Soc. Am., 33, 1725-1736.

Carbonell, J.R., M.C. Grignetti, K.N. Stevens, C.E. Williams, and B. Woods (1965), "Speaker Identification Techniques," AD-468993.

Clarke, F.R., R.W. Becker, and J.C. Nixon (1966), "Characteristics that Determine Speaker Recognition," ESD-TR-66-636.

Das, S.K. (1969), "A Method of Decision Making in Pattern Recognition," IEEE Trans. on Computers, C-18, 329-333.

Edie, J., and G. Sebestyen (1962), "Voice Identification General Criteria," RADC-TDR-62-278.

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

Floyd, W. (1964), "Voice Identification Techniques," RADC-TDR-64-312.

Fujimura, O. (1962), "Analysis of Nasal Consonants," J. Acoust. Soc. Am., 34, 1865-1875.

Garvin, P., and P. Ladefoged (1963), "Speaker Identification and Message Identification in Speech Recognition," Phonetica, 9, 193-199.

Gerstman, L.J. (1968), "Classification of Self-Normalized Vowels," IEEE Trans. on Audio and Electroacoustics, AU-16, 73-77.

Glenn, J.W., and N. Kleiner (1968), "Speaker Identification Based on Nasal Phonation," J. Acoust. Soc. Am., 43, 368-372.

Gold, B. (1962), "Computer Program for Pitch Extraction," J. Acoust. Soc. Am., 34, 916-921.

Hecker, M.H.L., K.N. Stevens, G. von Bismarck and C.E. Williams (1968), "Manifestation of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Am., 44, 993-1001.

Hemdal, J. (1967), "Some Results from the Normalization of Speaker Differences in a Mechanical Vowel Recognizer," J. Acoust. Soc. Am., 41, 1594 (Abstract).

Henke, W.L. (1968), "Speech Computer Facility," Quarterly Progress Report No. 90, Research Laboratory of Electronics, M.I.T., 217-219.

Henke, W.L. (1969), "TASS-Terminal Analog Speech Synthesis," R.L.E. Speech Communication Computer Facility Memo No. 6, M.I.T., July 15, 1969.

Holmgren, G.L. (1966), "Speaker Recognition, Speech Characteristics, Speech Evaluation, and Modification of Speech Signal - A Selected Bibliography," IEEE Trans. on Audio and Electroacoustics, AU-14, 32-39.

Kamentsky, L.A., and C.N. Liu (1963), "Computer-Automated Design of Multifont Print Recognition Logic," IBM J. Res. Develop., 7, 2-13.

Kersta, L.G. (1962), "Voiceprint Identification," Nature, 196, 1253-1257.

Ladefoged, P., and D.E. Broadbent (1957), "Information Conveyed by Vowels," J. Acoust. Soc. Am., 29, 98-104.

Lewis, P.M. (1962), "The Characteristic Selection Problem in Recognition Systems," IRE Trans. on Information Theory, IT-8, 171-178.

Li, K.P., J.E. Dammann, and W.D. Chapman (1966), "Experimental Studies in Speaker Verification, Using an Adaptive System," J. Acoust. Soc. Am., 40, 966-978.

Marill, T., and D.M. Green (1963), "On the Effectiveness of Receptors in Recognition Systems," IEEE Trans. on Information Theory, IT-9, 11-17.

Mártony, J. (1965), "Studies of the Voice Source," Quarterly Progress and Status Report 1/65, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 4-9.

Meeker, W.F., T.B. Martin, M.B. Herscher, D. Phyfe, and M. Weinstock (1967), "Automatic Speaker Recognition Using Speaker Recognition Techniques," J. Acoust. Soc. Am., 42, 1182 (Abstract).

Miller, J.E. (1964), "Decapitation and Recapitation, a Study in Voice Quality," J. Acoust. Soc. Am., 36, 2002 (Abstract).

Nagy, G. (1968), "State of the Art in Pattern Recognition," Proc. IEEE, 56, 836-862.

Nilsson, N.J. (1965), Learning Machines (McGraw-Hill, New York).

Noll, A.M. (1967), "Cepstrum Pitch Detection," J. Acoust. Soc. Am., 41, 293-309.

Paul, A.P., A.S. House, and K.N. Stevens (1964), "Automatic Reduction of Vowel Spectra: An Analysis-by-Synthesis Method and Its Evaluation," J. Acoust. Soc. Am., 36, 303-308.

Peterson, G.E., and H.L. Barney (1952), "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am., 24, 175-184.

Pollack, I., J.M. Pickett, and W.H. Sumby (1954), "On the Identification of Speakers by Voice," J. Acoust. Soc. Am., 26, 403-406.

Pruzansky, S. (1963), "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Am., 35, 354-358.

Pruzansky, S., and M.V. Mathews (1964), "Talker Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Am., 36, 2041-2047.

Rabiner, L.R., R.W. Shafer, and C.M. Rader (1969), "The Chirp z-Transform Algorithm and Its Application," Bell System Tech. J., 48, 1249-1292.

Sebestyen, G.S. (1962), Decision-Making Processes in Pattern Recognition (Macmillan, New York).

Shearme, J.N., and J.N. Holmes (1959), "An Experiment Concerning the Recognition of Voices," Language and Speech, 2, 123-131.

Spinrad, R.J. (1963), Machine Recognition of Hand Printed Block Letters, Ph.D. Thesis, Department of Electrical Engineering, M.I.T.

Stevens, K.N. (in press), "The Quantal Nature of Speech: Evidence From Articulatory-Acoustic Data," in E.E. David, Jr. and P.B. Denes (Eds.), Human Communication: A Unified View (McGraw-Hill, New York).

Stevens, K.N., and A.S. House (1963), "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study," J. Speech Hearing Res., 6, 111-123.

Stevens, K.N., C.E. Williams, J.R. Carbonell, and B. Woods (1968), "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material," J. Acoust. Soc. Am., 44, 1596-1607.

Tobias, J.V. (1959), "Relative Occurrence of Phonemes in American English," J. Acoust. Soc. Am., 31, 631.

BIOGRAPHICAL NOTE

Jared John Wolf was born in Wilmington, Delaware on June 17, 1942. After graduating from Westtown School in 1960, he entered Union College in Schenectady, New York. He spent the year 1962-63 studying in the Faculty of Arts and Sciences at the University of St. Andrews in Scotland as the Union College exchange student. In 1965, he received the degree of Bachelor of Electrical Engineering, summa cum laude, from Union College. In 1967, he received the degree of Master of Science in Electrical Engineering from the Massachusetts Institute of Technology. He was a National Science Foundation Fellow during his four years at M.I.T.

He was a Teaching Assistant in the Department of Electrical Engineering in 1966-67. During the summers of 1967 and 1968, he was on the DSR staff at the Research Laboratory of Electronics, M.I.T., where he participated in the design of much of the peripheral equipment and interfaces of the Speech Communication computer facility. He has worked for the Computer Division of the Philco Corporation and the Philadelphia Electric Company. He was also an IAESTE trainee at Hartmann und Braun, A.G., Frankfurt am Main, Germany.

He is a member of Sigma Xi, Tau Beta Pi, the Institute of Electrical and Electronics Engineers, and the Acoustical Society of America.

He married the former Eveline Rose of Albany, N. Y. in 1967.