INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

VOWEL NORMALIZATION:
THE ROLE OF FUNDAMENTAL FREQUENCY AND UPPER FORMANTS

by

BENJAMIN HALBERSTAM

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

1998

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

August 22, 1998
Date

Lawrence J. Raphael
Chair of Examining Committee

August 22, 1998
Date

Executive officer

Dr. Katherine Harris

Dr. Arthur Boothroyd

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

VOWEL NORMALIZATION:
THE ROLE OF FUNDAMENTAL FREQUENCY AND UPPER FORMANTS

by

BENJAMIN HALBERSTAM

Advisor: Professor Lawrence J. Raphael

Some vowel normalization schemes attempt to account for successful vowel classification by listeners despite the interspeaker overlap between acoustic vowel categories in the F1 x F2 space, by assuming perceptual exploitation of F0 and F3 information.

This study tested an implicit prediction made by these vowel normalization schemes: availability of F0 and F3 information should improve listeners' classification of naturally produced vowels with typical interspeaker formant variability presented in a mixed-speaker condition, relative to their ability to classify the same stimuli in a blocked-speaker condition.

Eight subjects classified phonated and whispered vowels from the set /i I ɛ æ ɑ ɔ ʌ U u/ produced by speakers varying in age and sex. Stimuli were presented with and without F3 and upper formants filtered out, in blocked-speaker and mixed-speaker conditions.

Vowels were classified with greater accuracy for blocked-speaker than for mixed-speaker condition, for phonated vowels than for whispered vowels and for unfiltered than for filtered vowels.

The interaction between presentation condition and phonatory type was just short of significance at the 0.05 level [p. = 0.068]. Phonated vowels were classified with similar accuracy in blocked-speaker and mixed speaker conditions. Whispered vowels were classified with significantly lower accuracy in mixed-speaker condition than in blocked-speaker condition. This finding indicates that F0 is likely to be perceptually useful in vowel normalization.

The interaction between phonatory type and presentation condition was of low magnitude, and was not significant at the 5% level [p. = 0.189]. Thus, the results did not provide perceptual support for including F3 as a parameter in vowel normalization.

The interaction between phonatory type and availability of upper formants was extremely robust [p=0.001]. The implication of this unexpected finding is that third formant information is relatively inconsequential for phonated vowel perception, but of considerable importance for whispered vowel perception.

Acknowledgments

First and foremost I would like to express my appreciation to my wife and to my parents for their dedicated support throughout my graduate studies. I could not have made it without them.

I gratefully acknowledge Distinguished Professors Katherine Harris and Arthur Boothroyd for their comments, criticism and encouragement. The quality of this work has been immeasurably enhanced by their contributions.

I would like to thank Dr. Irving Hochberg for his support and encouragement throughout my stay at the Graduate Center. His direct and indirect impact on my experience in the program has helped shape my professional orientation.

I would like to thank the faculty and staff of the department for all of their contributions. In particular, my thanks to Professor Mark Weiss for many fruitful discussions and help with some computer programs, and my thanks to Gary Chant for patient assistance whenever it was needed.

Lastly, I must express my indebtedness to Dr. Lawrence Raphael. He guided me through this dissertation, helping me to develop my own vision, without ever imposing his own.

The extent to which he made himself available to help me was truly remarkable.

It is self-understood that any errors or shortcomings of omission or commission in my research and dissertation are entirely my own.

Table of Contents

List of Tables

List of Tables (continued)

List of Figures

CHAPTER I

INTRODUCTION

RECOGNITION OF INADEQUACY OF A SIMPLE FORMANT MODEL

Over the course of more than a century of vowel experimentation, it was established that vowels are produced by a complex harmonic source (vocal fold vibration) with some of the harmonics reinforced as a consequence of falling within highly resonant frequency bands relating to the shape of the vocal tract. It was also established that typically more than one such peak of energy could exist for a given vowel. It was assumed that these peaks of acoustic energy (known as formants) were the primary cues to vowel identity. (For an historical review, see appendix A.)

Subsequently it was seen that similar formant patterns produced by different speakers could result in different vowel percepts (Peterson & Barney, 1952). Ultimately this led to the realization that a simple acoustic target model of vowel perception, with the first two formants as the sole cues to vowel perception, was inadequate to explain vowel perception (Jenkins, 1987).

FIXED PITCH OR VARIABLE PITCHES?

Before scientists could begin to deal with the problems caused by inter-speaker formant frequency variability within vowels, it was necessary to establish that such variability existed. The history of this endeavor is confusing because the issue has often been discussed under the rubric of whether vowels as linguistic entities are acoustically characterized by "fixed or variable pitches," a term which meant different things to different researchers.

McKendrick (1898) discusses the fixed vs. variable pitch issue as it relates to the harmonic and inharmonic theories. He interpreted the fixed pitch theory as being inconsistent with the idea that vowels are composed of harmonics. This is because the frequency of the harmonic of the greatest amplitude must vary with the fundamental frequency, and so this harmonic cannot have a fixed pitch for vowels produced with different fundamental frequencies. Because it is now understood that voiced vowels are composed of harmonics, McKendrick's (1898) discussion is of limited interest; clearly, the frequency of the most intense harmonic does vary with the fundamental frequency.

Scripture (1904) treats the fixed vs. variable pitch issue in terms of whether or not cavity tones (or formants)

tend to rise as the fundamental frequency rises within the same vowel produced by the same speaker. Similarly, the issue was sometimes treated in terms of whether the loudest harmonic in a given vowel is a particular multiple of the fundamental or one of a particular frequency range (Rayleigh, 1896). Donders (1864) (cited by Scripture, 1904) and Helmholtz (1877/1954) are among those who assume that a very narrow resonance peak is characteristic of each vowel.

Although no data have ever shown that a specific harmonic of the fundamental is characteristic of each vowel, Lloyd is cited by McKendrick (1898) as believing that "as we ascend a scale in singing a vowel, the pitch of the oral cavity slightly changes." Joos (1948), on the other hand, suggests that speakers adjust formants to "habitual or average pitch, NOT to the pitch used at the moment: spectrographic evidence is emphatically clear on this point." Joos' position is assumed to be correct today, and textbooks teach that speakers do not deliberately alter the shape of their vocal tracts in order to produce a peak of resonance more suited to the fundamental frequency which they happen to be using. There has, however, been limited research in this area.

The related but more serious issue implied by the terms fixed and variable pitches is whether there is variability in

the frequencies of the formants for a given vowel between speakers. The existence of this variability is the source of the most serious problem relating to vowel perception, but was not immediately recognized by those who studied vowel acoustics.

HISTORIC METHODOLOGY FOR DETERMINATION OF INTERSPEAKER VARIABILITY FOR VOWEL FORMANT FREQUENCIES

Historically, the determination of whether or not formant frequencies were fixed or variable for a given vowel was primarily studied in two ways. The first was by estimating or measuring formant frequencies for various speakers and comparing them. The second method was more theoretically based and involved changes in playback speed of recorded vowels. These two methods will be reviewed sequentially.

Helmholtz researched the issue by estimating vowel resonances for various speakers (Helmholtz, 1877/1954). He certainly knew enough about resonance to know that children and adult males would be expected to have very different peaks of resonance for similarly-shaped vocal tracts. He was most likely attempting to avoid the difficulties involved in explaining perceptual constancy that is the natural outcome of this variability when he made the following error and took the position that vowels of all speakers can be characterized by a fixed pitch: "I have in general found the same resonances in men women and children. The want of space in the oral cavity of women and children can be easily replaced by a great closure of its opening, which will make the

resonance as deep as in the larger oral cavities of men." Clearly, this theory could not have been based on personal observation.

As late as 1937, Miller (1937) argues that "...all the different vowels have distinctly different characteristic regions of resonance which remain the same for all voices." An examination of Miller's (1937) hand drawn "energy curves" reveals a large degree of arbitrariness. He often draws curves around individual harmonics in a manner that supports, rather than undermines, his argument.

With the improvement of research techniques through the development of ever-improving recording technology and culminating in the development of the sound spectrograph (Potter, Kopp and Green, 1947), the debate became more and more one-sided and gradually faded into history.

Lloyd (1890, 1898), for instance, recognizes the existence of formant frequency variability for different speakers for the same vowels both from a theoretical basis as well as phonographic evidence.

Paget (1930) cites a study by Crandall and Sacia using a condenser transmitter finding that "female voices show a tendency toward higher resonances than for male voices." Chiba and Kajiyama (1941), Joos (1948), Potter and Steinberg

(1950) Peterson and Barney (1952) and others all report finding systematic variability of formant frequencies within individual vowels across speakers. Female speakers had formants of higher frequency than male speakers and children's formants were higher still. In the face of such overwhelming evidence, the issue was no longer whether or not formant frequencies were variable across speakers, but rather why human vowel perception is seemingly undisturbed by this variability.

An additional method of evaluating whether vowels are characterized by fixed or relative pitches, or, more appropriately, fixed or relative formants, has been the alteration of playback speed. This method was also used to test the perceptual validity of the formant ratio-theory.

Chiba and Kajiyama (1941) explain the rationale of the playback speed alteration experiments as follows: "If the vowel has two formants, their pitch relation is maintained on account of their simultaneous rising or falling. From this it would appear that the relative formant theory would hold if a vowel remained the same in spite of the change in the velocity of rotation, and otherwise the fixed formant theory would be legitimate." They list several researchers who have used this method.

McKendrick (1898) cites Hermann as pointing out that "the quality of a vowel-tone varies considerably, according to the rate at which the cylinder was rotated. This should obviously be not the case were the relative pitch correct." However, in a telling remark he adds: "It is curious that, even with competent observers, there should be such difficulty in deciding this apparently simple question of fact, some asserting that there is no difference in quality, and others as positively stating that there is a difference when the cylinder is caused to move slowly." This is an excellent recommendation for the use of unbiased subjects in perceptual experimentation, yet more than forty years later, Chiba and Kajiyama (1941) were still using their own perception as the basis to determine the perceptual effects of similar alterations in playback speed.

McKendrick (1898) himself maintains that in his own analysis, changes in playback speed do "undoubtedly" alter vowel quality, but that "the sound of the vowel never passes...into the sound of another vowel."

Scripture (1904) cites both Hermann's experiments and Rousselot's experiments of the same kind, as well as his own, and argues that alterations in vowel quality can always be demonstrated by changing playback speed to a sufficient

degree. He writes that "these phonograph experiments show not only that an important essential of the vowel character does not lie in the relation of the cavity tone to the cord tone, but also that it does lie in the presence of a tone of limited range of pitch."

LIMITATIONS OF PLAYBACK SPEED ALTERATION EXPERIMENTS

Faced with a large body of observations (however unobjective) that playback speed does alter vowel quality, an explanation must be provided for the empirical fact that variable formants for different speakers are the rule for the same vowels (Chiba and Kajiyama, 1941; Joos ,1948; Potter and Steinberg, 1950; Peterson and Barney, 1952). There must be a flaw in the assumptions made by many of the researchers who used this method. As cited above, Chiba and Kajiyama (1941) explain this rationale by explaining "that it would appear that the relative formant theory would hold if a vowel remained the same in spite of the change in the velocity of rotation, and otherwise the fixed formant theory would be legitimate."

There are in fact numerous flaws in the assumption that a finding of perceptual changes in vowel identity as a result of changes in playback speed proves that formants must have

fixed values. One of these is the resolution implied by Chiba and Kajiyama (1941). They argue that while fixed frequencies are not required for the maintenance of vowel identity, a limited range of frequencies is required. If their argument is accepted, it is only if variations in playback speed move formant frequencies out of their normal range that changes in vowel identity should be expected to result.

A second flaw in the use of changes in playback speed for perceptual experimentation is that such experimentation fails to take into account a truly auditory frequency scale. When a recording is speeded up or slowed down, each frequency is altered by a constant multiple. Although on the surface this would indicate that the method results in all elements being identically treated, this is not the case. Several auditory frequency scales have been developed that reflect frequency as it is perceived, rather than as a function of physical values. These include the Koenig scale (Koenig, 1949), the mel scale (Stevens and Volkman, 1940) and the Bark scale (Zwicker, 1961). With an understanding of the concepts underlying these auditory frequency scales, it becomes clear that while a physically-based frequency ratio may be maintained when multiplying the frequencies of each component

of a sound by a constant, an auditory-based pitch ratio will not.

A third flaw in this methodology relates to a concept that will be important in this dissertation: Other elements of the acoustic waveform that may normally contribute to the perception of formants are altered by changes in playback speed. Primarily, the use of fundamental frequency in a perceptual normalization process becomes less possible when the relationship between the fundamental frequency and formant frequencies is unnaturally altered by changes in playback speed.

Finally, as in so much of vowel perception research, duration was not controlled for in the typical playback speed methodology. Although this problem is avoidable using modern instrumentation, simply slowing and speeding playback speed increases and reduces duration. Vowel duration has been shown to affect categorization of vowels (Tiffany, 1952; Strange et al., 1983).

For these reasons, this very common, old experimental method, is not seen as being as crucial as it was once viewed. Formant frequencies are indeed variable among speakers; this fact is not inconsistent with the finding that

vowel quality may be altered by changes in playback speed of recorded vowels of a single speaker.

In light of the recognition of the existence of formant frequency variability, an important research issue became to explain how listeners can perceive variable cues as equivalent. This issue became more critical once it was recognized that vowels produced with similar F1 and F2 frequencies produced by different speakers could result in different vowel percepts (Peterson & Barney, 1952; Potter and Steinberg, 1950).

LLOYD'S SOLUTION

Lloyd's (1890, 1898) proposed solution to the perceptual problems that arise from the variance found in interspeaker vowel formant frequencies was to suggest that it is the relationship among frequencies of the formants, rather than absolute frequencies of the formants that is characteristic of vowels. He writes (Lloyd, 1898): "Note how much less variable is the ratio between the resonances than are the resonances themselves," and goes on to state that his conclusion "has derived further support from every set of Fourierian analyses yet published."

Lloyd's recognition of the problem of acoustic variability and the necessity of developing a theory to explain perception in the face of this variability, was not lost on his contemporaries. Rayleigh (1896) describes Lloyd's view and comments that "In this way he explains the difficulty arising from the fact that the articulation for a given vowel appears to be the same for an infant and a grown man, although on account of the great difference in the size of the resonant cavities, the absolute pitch must vary." McKendrick (1898) writes that Lloyd's approach accounts for the fact "that the articulation of a vowel seems to be the same for a child as for an adult."

Chiba and Kajiyama (1941) describe their own theory as an extension of Lloyd's (1890) theory. Their theory, called the "space pattern theory," states that: "A vowel is characterized by its relative formants, providing the centres of the formants are situated within certain frequency regions fixed for a given vowel." This region is limited by "the ratio of the frequency of the upper limit to that of the lower limit being about 1.7."

They describe playback speed variation experiments that confirm their view. They report that with an adult speaker, "even a considerable increase in speed does not affect the

vowels, which however undergo some change in the event of the speed being reduced even slightly. But exactly the opposite phenomenon is observed with the voice of a child, increase in speed brings about a quicker change in the vowels than does a diminution."

In fact, Chiba and Kajiyama's (1941) theory differs substantially from Lloyd's (1890, 1898) theory. Lloyd, as we have seen, was insistent that all vowels had at least two characteristic resonances, and it was the ratio of these resonances that was seen to be invariant. Chiba and Kajiyama (1941) applied their theory to cases of "so-called single formant vowels." Rather than attaching a numerical ratio to vowels, they simply argue that by limiting the variability to a specific range, the vowel acoustics can all be shown to have recognizably different patterns. In a sense, their theory bears a relationship to the whole spectrum approaches of Bladon, Hendon, and Pickering (1984) and Suomi, (1984).

It seems unlikely, however, that their theory could adequately explain successful vowel perception in some of the richer vowel systems found in other languages such as English or Dutch. Different vowels in American English can have F1 and F2 frequencies that are similar or identical when different speakers are taken into account (Peterson and

Barney, 1952). Arguably, Chiba and Kajiyama's (1941) system seems more defensible when it is applied (as they mostly do) to the relatively limited Japanese vowel system. In any event, their theory is difficult to evaluate on its own terms. They provide no specific algorithm to allow for the placement of vowels into appropriate categories.

Miller (1989) lists several articles other than Chiba and Kajiyama's (1941) article that present some form of a formant ratio theory. These articles, like Chiba and Kajiyama's article, generally do not evaluate whether the suggested approach could successfully eliminate overlapping cues for individual vowels across speakers, while maintaining differences between different vowels. Such evaluations, for more specifically designed vowel classification models have been done more recently (Disner 1980; Gerstman, 1968; Hillenbrand and Gayvert, 1993; Lobanov, 1971; Miller, 1989; Syrdal and Gopal, 1984 and others).

MODERN APPROACHES TO FORMANT VARIABILITY

The following will provide a brief review of the status in the 1950s of the "simple target model," (Strange, 1989) described by Jenkins (1987) as the "received theory" of vowel perception. The development of technology that could

accurately locate the first two peaks of acoustic energy or formants, referred to as formants one and two (F1 and F2), revealed that different speakers produce formants with very different frequencies within the same vowel category. In particular, formants for adult males tend to be lower than formants for adult females, which in turn are lower than formants for children.

Conversely, and perhaps more importantly for understanding vowel perception, it was seen that similar formant patterns produced by different speakers could result in different vowel percepts (Peterson & Barney, 1952). Thus it became evident that a simple acoustic target model with the first two formants as the sole cues for vowel perception, was inadequate to explain vowel perception (Jenkins, 1987).

As a result of these findings, theories of vowel perception moved in two general directions. One type of theory that developed is referred to by Strange (1989) as "dynamic specification models," and a second is referred to as "elaborated target models." Dynamic specification models propose that dynamic acoustic information, such as transitional information partly conditioned by neighboring consonants, and not the steady-state frequencies of the first two formants, is the primary cue to vowel perception.

Elaborated target models use additional acoustic information contained in steady-state vowels as parameters additional to F1 and F2 frequencies to disambiguate vowels that have similar frequency values for the first and second formants.

REVIEW OF DYNAMIC SPECIFICATION MODELS

Although this dissertation focuses on issues relating to "elaborated target models," I will briefly describe the approach of those advocating a "dynamic specification model."

The approach of the dynamic specification models is to suggest that formant frequency overlap for different vowels does not pose a true perceptual problem because steady-state formants are not the primary cue for vowel perception at all. A key claim is that vowels "that overlap in the F1/F2 acoustic space...may nevertheless be distinguished with respect to intrinsic duration and trajectory shape" (Strange, 1989).

Another key element of this approach is the argument that in light of formant frequency variability, the complexities involved in explaining formant-based perception are "too high a price to pay for the simple structure of the vowel target theory" (Jenkins, 1987).

There are two types of research findings that are used to support the claims of those who support dynamic specification models. One is the finding that vowels are better identified in a consonantal environment than in an isolated environment. The other is research that shows that transitional elements of vowels with the vocalic nuclei removed are better identified than are the vocalic nuclei themselves.

With regard to the issue of the perception of vowels in isolated and consonantal context, an early study by Strange Verbrugge, Shankweiler, & Edman, (1976) found that error rates in mixed speaker condition vowel identification were 17% for vowels in a (CVC) environment but were as high as 42.6% for isolated vowels. Strange et al. (1976) viewed the results as support for the idea that steady-state formants are insufficient for accurate vowel identification, and that the acoustic consequences of consonantal environments provide the primary cues for vowel identification.

Subsequent research has shown that some of the conclusions of the Strange et al. (1976) study were premature. That is to say, that although transitional cues can contribute to vowel perception, highly accurate vowel perception is possible without these transitional cues.

Assman, Nearey, & Hogan, (1982) found that difficulties in relating vowel stimuli to written responses were more severe for isolated vowels than for vowels in a consonantal context: Spoken responses transcribed by phonetically-trained experimenters resulted in error rates as low as 5.4%. They suggest that a failure to control for task-related issues, having little to do with actual perception, could explain earlier findings of a consonantal context advantage for vowel perception.

Several other studies have found that for vowel stimuli elicited under carefully designed elicitation procedures, listeners have error rates between 2% and 8% (Kahn, 1978; Macchi, 1980). Importantly, neither the Macchi (1980), nor the Assman et al. (1982) study was able to find significant advantages for vowel identification in consonantal versus isolated context.

At a minimum, the fact that listeners have been shown to perceive isolated quasi-steady state vowels with high levels of accuracy demands some explanation. Even if dynamic cues are involved in normal vowel perception (and the evidence as reviewed above would suggest that they are), listeners seem to have the ability to perceive vowels very accurately without the benefit of such cues. Apparently

there exists some perceptual mechanism to deal with the interspeaker formant frequency variability that does not rely on dynamic cues.

In addition, Papçun (1980) performed a correlational study comparing errors made in consonantal and isolated contexts. The consonantal context and isolated context stimuli which were perceived as other than the intended vowel showed fairly high correlations. Papçun argues that his data show "evidence for an essentially unitary process" of perception of these two types of stimuli. He argues that this would indicate that "the parameters derived from steady-state vowels are good candidates for input parameters to the study of the dynamic aspects of speech."

There are, however, some interesting data suggesting that transitional elements are more crucial to perception than the centers of vowels themselves (Jenkins, Strange and Edman, 1983; Strange, Jenkins and Johnson, 1983). Nonetheless, it is likely that these transitional elements may provide information about formants frequencies which are never reached. Certainly a very similar concept is the basis for the standard theory of recognition of place of articulation in voiceless stops. In that case it is believed that transitional information can point to certain acoustic

loci that are the characteristic for each place of articulation, but which are not actually present in the signal (Delattre, Liberman and Cooper, 1955).

In addition, more recent articles from the dynamic specification viewpoint (Strange, 1989) recognize a role for steady-state formants, as well as formant transitions into and out of these steady-state formants. Clearly, such transitions cannot be viewed as being completely unrelated to the target formant frequencies. Other factors, such as duration, can be incorporated into a formant-based model of vowel perception, without accepting some of the more radical implications of the dynamic specification models. In fact, both Jenkins (1987) and Strange (1989) have called for a more comprehensive view which recognizes the importance of both dynamic and "steady state" information.

REVIEW OF ELABORATED TARGET MODELS

The second type of theory, referred to by Strange (1989) as "elaborated target models," can also be described as "vowel normalization models." This term suggests that seemingly variable formant frequency values can be normalized through the use of additional acoustic information. Vowel normalization approaches suggest that the resultant

normalized formant frequency values are invariant for each vowel across subjects.

Several articles have reported on normalization schemes, designed to reduce the problems of vowel overlap in the F1 x F2 space (Fant, 1975; Gerstman, 1968; Lobanov, 1971; Miller, 1989; Syrdal, 1985; Wakita, 1977). These can be divided into two broad categories: extrinsic and intrinsic normalization. In extrinsic normalization schemes, information about the entire vowel system (or point vowels, which provide information about extreme F1 and F2 values) is used to provide the basis for calibrating the perceptual system and disambiguating vowels with overlapping formants. In intrinsic normalization schemes, acoustic information available within a given vowel is used to disambiguate the vowel.

EXTRINSIC NORMALIZATION

The originator of the extrinsic normalization theory is Joos (1948). He suggested that "upon first meeting a person, the listener hears a few vowel phones, and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete vowel pattern to serve as

background (coordinate system) upon which he correctly locates new phones as fast as he hears them."

Joos proceeds to hypothesize that the first few vowels heard are used to fix the end-points of the extremes of the F1 x F2 parameters, which reflect the extremes of the speaker's articulation. Vowels are then perceived in reference to these points. His approach is essentially one that depends on calibration.

Other, more sophisticated models of extrinsic normalization include Fant (1975), Gerstman (1968) and Lobanov (1971). One of these models (Gerstman, 1968) was not intended as a perceptual strategy, but rather as an algorithm that might be useful for machine recognition of vowels. This distinction has been ignored at times by critics of this approach.

Extrinsic normalization schemes greatly reduce the problems of vowel overlap (Disner, 1980; Fant, 1975; Gerstman, 1968; Lobanov, 1971), and as extrinsic normalization models would predict, varying the F1 and F2 values of precursor vowels was shown to have perceptual effects on test vowel perception (Ladefoged and Broadbent, 1957; Ainsworth, 1975; Nearey, 1989; but see Strange et al., 1976 and Verbrugge et al. 1976). In addition, vowels

presented in a blocked-speaker condition were found to be better identified than the same vowels presented in a mixed-speaker condition (Strange et al., 1976; Verbrugge et al. 1976). Thus, it is understandable that "it became the dominant theory in the field" (Jenkins, 1987). Nevertheless, extrinsic normalization is not taken as seriously as the key perceptual strategy for the disambiguation of vowels as it once was. This is in large part due to the finding that listeners perceive vowels fairly well without the benefit of precursor vowels (Nearey, 1989). In fact, vowel perception experiments using a mixed-speaker condition sometimes show very low error rates (Assman et. al., 1982; Kahn, 1978; Macchi, 1980). Clearly, prior experience with a speaker is not as crucial for vowel perception as extrinsic normalization models would suggest.

INTRINSIC NORMALIZATION

Models within the second broad category of elaborated target models are called intrinsic normalization models. Intrinsic normalization schemes exploit additional within-vowel acoustic information. One type uses the fourth formant, possibly in conjunction with other formants and bandwidth information, as the basis for the estimation of

vocal tract length, which in turn is used to derive a scale factor for F1 and F2 (Wakita, 1977). Another type uses FO and F3 as additional cues (Miller, 1989; Syrdal, 1985; Syrdal and Gopal, 1986).

Wakita's (1977) proposal was to scale formants by the same scale factor found for the relationship of the speaker's vocal tract to a reference vocal tract length. He suggested several methods for estimating vocal tract length from acoustic information within the vowel, and applied his formula to a set of vowels produced by male, female and child speakers. The normalized values showed less variability for male, female and child speakers than did the unnormalized values, and reduced the problems of vowel overlap in the F1 x F2 space.

To test his model, Wakita (1977) used a second data set based on another set of vowels produced by other speakers. After computing his normalized formant frequencies for this second data set, a formula was applied to determine identification of the input sounds. Thus, Wakita's (1977) study tests the viability of his proposal data-analytically, rather than perceptually. Although he reported between 5.5% and 17.3% improvement in correct vowel classification rates, depending "on the particular definition of what constitutes

an error," there are serious weaknesses in his approach. In particular, in all of his suggested methods for vocal tract estimation, there is some reliance on F4. In reality, however, vowels can be identified quite well without F4 information (Lehiste and Peterson, 1959; Peterson, 1954). In fact, there has been no published research seen by this author that would suggest that the presence or absence of F4 has any effect on vowel perception.

Although Miller's (1989) article deals with speech perception in a much more comprehensive way than Syrdal's (1985) article does, their approaches are very similar in terms of the normalization issue per se. Both use values obtained by calculating differences between F3 and F2, F2 and F1, and F1 and F0 as their variables.

Syrdal calculates these differences on a Bark scale which is based on critical band research and tends to be linear in low frequencies and logarithmic in the middle and high frequencies. Using the Bark scale has the effect of equating small shifts in lower frequencies and larger shifts in higher frequencies. According to the linear discriminant classifier that Syrdal uses, error rates are about 14% for this Bark difference method when applied to the Peterson and Barney (1952) data.

Miller (1989) retains the use of a log frequency scale, but uses a "sensory reference" instead of F0. The sensory reference is derived from a standard formula applied to the mean of the speaker's fundamental frequency calculated on a log frequency scale. The use of the sensory reference instead of F0, has the effect of reducing the magnitude of inter-speaker F0 differences to 1/3 of their actual magnitudes (Nearey, 1989). The use of the sensory reference seems to reflect the fact that on a log frequency scale, F0 inter-speaker differences are typically greater than formant frequency differences. Miller draws "vowel slabs" within the three dimensional space derived from the differences between F3, F2, F1 and the "sensory reference" for each vowel. Miller (1989) reports that 403 of 413 vowel formant values found in the literature would fall within these hand drawn vowel slabs, and thus be correctly identified.

It is interesting to note that an analysis by Hillenbrand and Gayvert (1993) reported on the benefit of various parameter sets for vowel categorization using a quadratic discriminant analysis. They found that the addition of F0 values to F1 and F2 provides nearly identical benefit for vowel categorization as the addition of both F0 and F3. Despite this finding, no normalization scheme that

uses just FO in addition to F1 and F2 has received the attention of the previously mentioned schemes that use both FO and F3.

THE PERCEPTUAL FORMANT

Although not directly related to the experiment that will be described in this dissertation, there are two important perceptual issues, key to vowel normalization, that have not been consistently addressed in the literature. They are (1) what the perceptual definition of a formant should be, and (2) what scale is appropriate for the necessary transformation of frequency values for the purposes of normalization.

Vowel formant frequencies are generally identified based on acoustic measurement of the external waveform. They are therefore estimates of the resonance peaks based on the harmonics of the greatest amplitude. As discussed earlier, it was originally assumed that it was single harmonics or tones that were perceptually crucial for vowel identity. With the recognition of the fact that more than one harmonic could fall within the bandwidth of the peaks of resonance in the vocal tract, it became clear that more than one harmonic might also be involved in the perception of vowels.

Helmholtz (1877/1955) was among the first to recognize this fact. He writes that the centers of resonance of formants are "situated within sufficiently narrow intervals from the upper partials of the speaking tone to create

sensible resonance of one or more of these partials and thus characterize the vowel."

Scientists toward the end of the last century began theoretical discussion of the correct approach to mathematical derivation of the formant from the amplitudes of individual harmonics. Lloyd (1898) discusses three proposed "centre-of-gravity calculations" using as many "heavy" (high amplitude) harmonics as are seen. Essentially, Hermann's method [cited by Lloyd (1898)] uses a linear scale to determine the weighted "mean partial;" Pipping's method [cited by Lloyd (1898)] is identical, but used a logarithmic scale. Lloyd's (1898) own method multiplies amplitude by frequency instead of using amplitude alone before calculating the center of gravity on a logarithmic scale. Aside from defending his method on theoretical grounds, Lloyd (1898) considered his method to be a "correction" for the fact that, in terms of amplitude, a neutral voice is composed of a "rapidly declining series" of harmonics.

The question of how several harmonics are used to derive a perceptual formant is an issue that needs to be empirically, rather than theoretically, defined. There has been some experimentation attempting to answer this question. Two important issues are how many harmonics are involved in

the estimation, and how much weight should be given to each harmonic.

Carlson, Fant and Grandstrom (1975) conducted an experiment designed to use perceptual data to decide between four hypotheses for calculating the perceptual formant. The best fit to the data was one that used the two most prominent harmonics in the loudness (sone) space.

Javkin, Hermansky and Wakita (1987) used both perceptual experiments and a re-analysis of difference-limen experiments to study the issue. They show that harmonics close to formant centers tend to attract judgments of formant location. Their research indicates that although the two most prominent harmonics are key to the derivation of a perceptual formant it is necessary to use a scale that effectively expands the differences in amplitude between these two harmonics.

Additional research by Assman and Nearey (1987) shows that "upper-edge harmonics" are more critical than lower-edge harmonics in formant perception. This presumably relates to the tendency for amplitudes of successive harmonics to be progressively lower. In other words, since amplitudes of lower frequency harmonics generally tend to be greater for voiced vowels produced by any vocal tract configuration,

harmonics with amplitudes that go against this trend are more indicative of vocal tract resonances than are harmonics that follow this trend.

It is worth noting that the Assman and Nearey (1987) study also found that LPC analysis fared as well as Carlson, Fant and Grandstrom's (1975) formant estimation based on the two most prominent harmonics. This finding is particularly important today, as computer estimation of formant frequency using LPC analysis has become commonplace in vowel research.

PERCEPTUAL FREQUENCY SCALES

An additional important issue for vowel normalization is the frequency scale appropriate for performing mathematical transformations of formant and fundamental frequency values. Since frequency is perceived as pitch, when we, for example, raise a 500 Hz tone, to 1000 Hz and we wish to treat a 2000 Hz tone in the same way, it is arguably inappropriate to do so by raising the tone to 4000 Hz. Doubling frequency values may not be equivalent to doubling perceived pitch.

The importance of the choice of frequency scale in vowel perception was recognized by researchers before the turn of the twentieth century. Lloyd (1898), as cited above,

agrees with Pipping's argument that a logarithmic scale is more suited to vowel perceptual theories than is a linear scale.

In this century, it has been recognized that although the logarithmic scale has perceptual importance, it is not linearly related to a true perceptual scale of pitch. As mentioned above, several auditory frequency scales have been developed that relate to pitch, both directly (Stevens and Volkman, 1940), and indirectly (Koenig, 1949; Zwicker, 1961). These scales would seem to be, a priori, more likely reflective of the perceptual processes that underlie theories of vowel normalization. Some studies have directly addressed this issue.

Fant (1975) justified his use of the mel scale over the logarithmic frequency scale on the basis of improved visual separability of vowels in the F1 x F2 space. The mel scale has particular appeal because it has been directly "constructed on the basis of subjective pitch evaluations" (Fant, 1975). Conversely, unlike the Bark scale, the research supporting the mel scale relates to simple, rather than complex sounds.

Bladon, Hendon, and Pickering (1984) use an "auditory transformation" whose first step is conversion of frequency

into "the more perceptually motivated 'tonality' (= perceived pitch) scale of Bark units...This conversion reflects the finding that the transformation of incoming frequencies into places of stimulation on the basilar membrane of the cochlea takes place in terms of auditory 'critical bands.'"

Syrdal and Gopal (1986) justify their use of critical band or Bark scale on similar grounds. They describe the Bark scale as "probably most appropriate for the representation of the complex speech spectrum." They cite a "wide variety of psychophysical experiments," and the critical band's "physiological correlates."

A few articles have attempted to justify their choice of frequency scale based on specific studies relating directly to vowel normalization. Miller (1989) is one of these. Miller's article first shows that the mel, Koenig and Bark scales are very similar, and, in fact "over most of the range of values of the center frequencies of F1, F2 and F3," these scales are also nearly equivalent to a linear frequency scale in Hertz.

More importantly, Miller (1989) has a complex analysis of the effectiveness of clustering the same vowels for different speakers using differences, ratios and log ratios of the first three formants calculated in scales based on

Hertz, mels, Koenigs and Barks. Miller found that "when the vowels are specified by the logs of the ratios of formant center frequencies measured in Hertz or mels, the best clustering is observed." Miller (1989) uses this finding as justification for using the log frequency scale for his perceptual model.

Hillenbrand and Gayvert (1993) reported on the benefit of parameter sets including FO and the first three formants for vowel categorization using a quadratic discriminant analysis. They performed the analysis for linear frequency, log frequency, Bark, mel and Koenig scales. Their results were similar to Miller's (1989) results, but they found very similar results for all but the linear frequency scale, which fared substantially worse for vowel classification than the other frequency scales.

Analysis of this kind may be considered to be a reasonable basis for selection of a frequency scale for a vowel normalization procedure that hypothesizes perceptual processes. This is because the decision of which frequency scale to use for vowel perception can logically be made based on which scale can best separate the data. In fact, Miller (1989) uses his analysis as one of his bases for selecting the log frequency scale.

However, Miller is on much less secure ground when he tries to justify the use of the log frequency scale on general auditory grounds. In particular, there is a large body of literature showing that non-logarithmic frequency scales are not inconsistent with the importance of the log frequency scale in musical scales, which is clearly a perceptual scale. (A full discussion is beyond the scope of this review, but for a particularly elegant solution, see Shepard, 1982.)

One limitation of these data-analytic approaches to the selection of an appropriate frequency scale is that "these methods can suggest logically possible perceptual strategies, but other information is required to determine whether listeners actually adopt a proposed strategy" (Hillenbrand and Gayvert, 1993). This qualification is true for all aspects of theoretical attempts to explain vowel perception, and is often ignored.

PERCEPTUAL REALITY OF INTRINSIC VOWEL NORMALIZATION

It has been demonstrated that the inclusion of FO and F3 frequency data can disambiguate vowels which overlap in plots of formant frequency in the Fl x F2 space (Disner, 1980; Hillenbrand and Gayvert, 1993; Miller, 1989; Syrdal, 1985). The issue of whether these variables can be shown to perform such a role in actual perception has not been well-explored. This study employs a method for clarifying this point.

A key issue in evaluating the potential strength of intrinsic vowel normalization proposals is to determine what the perceptual effects of varying FO and/or F3 are on vowel and formant perception. This question has been studied for close to 30 years. Some of the studies that examined this issue varied FO and/or F3 for given Fl and F2 values in synthetic vowels and examined the effect of these changes on perception.

A well-known study of this type was done by Fujisaki and Kawashima (1968). This study measured the effect on perceptual boundaries of varying FO and F3 along continua between /o/ and /a/ and between /e/ and /u/. The study found that there were perceptual boundary shifts caused by

such variation, and that the greatest boundary shifts were found when FO and F3 were both varied. A similar experiment by Hirahara and Kato (1992) also found boundary shifts caused by variation in FO.

Studies by Ainsworth (1975) and Nearey (1989) examined the perceptual effects of changes in FO and/or F3 by calculating the shifts in the means of F1 and F2 for specific vowels as a result of the FO or F3 changes. Ainsworth (1975) examined increases in perceived F1 and F2 values by observing the vowel labels given by twenty subjects to synthetic stimuli with varied F1 and F2 frequencies. They found that, in conjunction with raising the F1 and F2 frequencies of precursor vowels, doubling FO resulted in 8% increases in perceived F1 and F2 values. An elaborate post hoc explanation estimated the true shift of perceived F1 and F2 values at approximately 16%. Doubling FO without covarying the F1 and F2 frequencies resulted in increases in perceived F1 and F2 values of lower magnitudes.

Nearey (1989) found that doubling FO resulted in an 8% shift and that F3 changes had a smaller effect. However, Traunmuller (1990) argues that Nearey's experiment was improperly designed. He suggests that a more sophisticated

interpretation of Nearey's results would be consistent with shifts that are double those reported by Nearey.

Two additional studies, Slawson (1968) and Traunmuller (1981) examined the issue of the effects of FO and/or F3 shifts from different perspectives. Slawson (1968) examined within-category judgments of vowel quality. Subjects were asked to compare the vowel quality of synthetic vowel standards with test vowels that had a doubled F0 and/or raised upper formants concurrent with incremental shifts in the values of the first two formants. For the doubled fundamental condition, test vowels were judged most similar to standard vowels when F1 and F2 were raised by a factor of 1.10. For the raised upper formant condition, test vowels were judged most similar to standard vowels when there was no shift of F1 and F2. For the raised fundamental and upper formants condition, test vowels were judged most similar to standard vowels when F1 and F2 were raised by a factor of 1.15.

Traunmüller (1981) examined the issue of whether the perceptual dimension of "openness" in vowels is determined by absolute values of F1 or by the distance between F1 and FO. He studied perception of synthetic vowels of a Central Bavarian dialect which has five degrees of openness.

Subjects were asked to identify one-formant synthetic sounds with various FO and Fl values. Results showed that distance between Fl and FO, and not the absolute value of Fl, was related to perception of openness or vowel height for FO values below 350 Hz. Alternatively, the results can be viewed as evidence that changes in FO are required in order to maintain vowel quality as Fl is shifted. However, two subsequent studies by Hoemeke and Diehl (1994) and by Fahey, Diehl and Traunmüller (1996) suggest a much less simple relationship among Fl and FO and vowel height.

Although the research cited above does demonstrate that vowel perception is affected by FO or F3 values, there has been little research demonstrating that they are actually used to resolve the difficulties which vowel normalization schemes were designed to resolve. Specifically, it is necessary to investigate whether the presence of appropriate FO or F3 information enhances perceptual classification of vowel stimuli with the formant variability typically found across speakers.

Nusbaum and Morin (1992) explored this issue using synthetic vowels. They synthesized simulated "phonated" and "whispered" vowels with and without the energy above F2 filtered out. In a blocked-speaker condition (in which

vowels were presented in blocks simulating one "speaker", followed by a block of vowels simulating a second "speaker" etc.), approximately 95% of the stimuli of all types were correctly identified. In a mixed-speaker condition (in which successive vowels were produced by different speakers), unfiltered phonated vowels were identified best, followed closely by phonated filtered vowels, followed more distantly by whispered unfiltered vowels and whispered filtered vowels.

The implication of this research is that in a mixed-speaker condition, in which listeners must contend with vowel overlap problems (without the benefit of speaker context), FO performs a primary role in disambiguating vowels, while the upper formants seem to perform a secondary role. Nusbaum and Morin's (1992) study is important because it provides perceptual evidence for the roles of FO and upper formants for improving vowel identification, which were, until their study, likely but theoretical constructs.

There should, however, be serious concerns about generalizing from these results to actual vowel perception. First, it is possible that the large advantage of phonated vowels over whispered vowels found by Nusbaum and Morin (1992) was related to the fact that the formant frequencies for the synthesized whispered vowels were matched to the

formant frequencies of the phonated vowel. In naturally produced vowels, formant frequencies for whispered vowels would be expected to be higher than for phonated vowels (Peterson, 1961; Kallail and Emanuel, 1984; Eklund and Traunmüller, 1997). Perhaps the phonated vowels would not have shown greater compensation for mixed-speaker condition than did the whispered vowels, if more appropriate formant frequencies had been used in the synthesis.

Another difficulty with Nusbaum and Morin's (1992) study is the composition of their mixed-speaker condition. Only two synthesized "male" and two synthesized "female" voices were used. It is possible that this condition was too constrained to create the potential for the ambiguity in formant values that normalization schemes are designed to address.

The present study was designed to avoid these objections. The study is similar to Nusbaum and Morin's (1992) study, but uses naturally spoken rather than synthesized vowels. Stimuli were again whispered and phonated vowels with and without low-pass filtering. The use of naturally spoken stimuli may demonstrate the importance of F0 and/or F3 to vowel perception, rather than demonstrate an effect that is an artifact of decisions concerning the

selection of synthesis parameters. Fifteen speakers varying in age and sex were used to create the potential for real ambiguity of vowel identity based solely on F1 and F2 frequencies.

An intrinsic difficulty in the present study is the fact that formant values and other acoustic parameters can vary significantly between phonated and whispered vowels produced by the same speaker (Peterson, 1961; Kallail and Emanuel, 1984; Eklund and Traunmüller, 1997). In fact, as will be reported, a clear pattern of higher whispered formant frequencies for the first three formants was noted in the vowel stimuli used in this study.

There are, however, two reasons to expect the results to be meaningful, despite the fact that formant frequencies are not identical for the two types of stimuli. Firstly, the Nusbaum and Morin (1992) study has already found results in the predicted direction with synthetic vowels whose formants were identical in the phonated and whispered vowels. Secondly, it can be assumed that the best possible formant pattern for perception is found in whispered as well as phonated vowels. Since formant frequencies can be modified by changes in tract shape, speakers can make alterations that can enhance the likelihood that the intended vowel will be

correctly identified by the listeners. In addition, the perceptual system for whispered vowels presumably has the capability of compensating for the differences in formant values inherent in phonated and whispered vowels. Therefore, both phonated and whispered vowels can be considered to have the formant frequencies optimal for vowel perception for their particular phonatory type, while only the phonated vowels contain information about fundamental frequency.

The purpose of the study was to investigate to what extent, if any, perceptual classification of vowel stimuli produced by multiple speakers is enhanced by the presence of appropriate F0 or F3 information. This question is crucial in evaluating leading vowel normalization theories.

HYPOTHESIS

As described above, the perceptual exploitation of F0 and F3, specifically for the purposes of vowel classification, is assumed by several researchers (Syrdal, 1985; Miller, 1989). They assume that F0 and F3 frequency values influence classification of vowels of given F1 and F2 frequency values, by providing information that enhances the likelihood of classifying vowels "correctly" (as intended by the speaker). Although these researchers have supported this theoretical orientation by conducting data-analytic, rather than perceptual experimentation, there is a good deal of perceptual research that suggests that F0 and F3 frequency information might be used in the manner that Syrdal (1985) and Miller (1989) suggest.

As previously mentioned, research by Fujisaki and Kawashima (1968), Nearey (1989) and Slawson (1968) has shown that vowel perception is influenced by F0 and F3 frequency changes for given F1 and F2 frequency values. Additional research by Ainsworth (1975), Fahey, Diehl and Traunmüller (1996), Hirahara and Kato (1992), Hoemeke and Diehl (1994) and Traunmüller (1981) have all shown that vowel perception is influenced in similar ways by F0 changes for given F1 and F2 frequency values.

This influence has been found to be unidirectional. Vowel "centers of gravity" or category boundaries move toward higher F1 and F2 values when F0 or F3 frequencies are raised. Their findings represent clear perceptual effects indicating that vowel perception is sensitive to F0 frequency values and, to a lesser degree, to F3 frequency values.

This sensitivity is potentially useful in the perceptual classification of vowels produced by many different speakers. Specifically, listeners may classify vowels of given F1 and F2 frequencies differently based on the F0 or F3 frequencies. This would make them more likely to correctly classify vowels produced by multiple speakers who tend to have F1 and F2 frequency values that have correlations to their F0 and F3 frequencies for given vowels.

It was therefore hypothesized, based on the above research, that the assumptions made by Syrdal (1985) and Miller (1989) would be demonstrable in a perceptual experiment, in which vowels from multiple speakers would be presented with and without F0 and F3 frequency information. Namely, it was predicted that the availability of F0 and F3 frequency values would increase listeners' ability to classify vowels as intended by the speaker when vowels are presented in a mixed-speaker condition relative to their

ability to classify the same stimuli in a blocked-speaker condition.

The only published research that specifically examines whether F0 and F3 information can be shown to perceptually influence vowel classification is the previously mentioned Nusbaum and Morin (1992) study. Although they report results that suggest that the presence of F0, and, to a lesser degree, F3, do improve correct vowel classification rates in a mixed-speaker condition, their study has limitations that make additional research necessary. As discussed above these limitations include the use of synthetic rather than naturally produced vowels, inappropriately selected formant frequencies for the synthesized whispered vowels, inadequate numbers and types of "speakers" to simulate a truly mixed-speaker condition and sparse reportage of statistical findings. The present study attempted to avoid these limitations.

CHAPTER II

METHODOLOGY

STIMULI:

Speakers:

Speakers were five adult males, 19 to 29 years of age, five adult females, 18 to 49 years of age, and five children, 7 to 14 years of age. All of the speakers were native New Yorkers, whose first language is English. An attempt was made to select speakers with matched dialect.

Vowels:

The speakers were recorded producing each of nine phonated and nine whispered vowels in a /hV/ syllable. The unfamiliarity of final position short vowels was mitigated through the use of an elicitation technique described below. The vowels in the set were /i ɪ ɛ æ ɑ ɔ ʌ ʊ u/.

Elicitation technique:

Each vowel was spoken by the experimenter to the speaker, who then said twice "I just heard the /hV/ again." The speakers were provided with a printed English word, corresponding to /hVd/ (e.g. "head") and were told to leave off the /d/. Thus, speakers attempted to pronounce all

vowels, including the lax vowels not found in word-final position in English, as they are pronounced when produced in a word ending in a /d/. The /hV/ context follows Kahn (1978) and allows the use of vowels in a familiar context (i.e. the /hVd/ "words"), without the coarticulation cues found in other consonantal environments. Additionally, speakers paused before the word "again" in order to minimize the coarticulatory effects of the following /ə/ on the test vowel. The experimenter repeated the process for all vowels that were judged by the speaker or by the experimenter to be mispronounced. Recordings were made with an Electrovoice omnidirectional microphone, model 635A, and a Marantz tape recorder, model PMD420.

Measurement and filtering of stimuli:

The vowels and their carrier phrases were digitized at a 16 bit, 20 kHz sampling rate on a Gateway 2000 computer. The target vowels were excised from the carrier phrase. The rest of the operations described in this paragraph were done using programs written by Professor Mark Weiss at the CUNY Graduate Center. The average amplitude levels measured over the entire vowel segments were made equal for all stimuli using the WAVEXAM program. Wide-band spectrograms and

averaged amplitude spectra for all of the vowel segments were made using the SPECTO program. F2 and F3 frequencies were measured for each vowel using judgments made by eye from wide-band spectrograms and from the averaged amplitude spectra for the entire phonated segment.

Formant frequencies of a sample of 10 vowels were estimated using the same procedure by an independent observer, whose reported frequencies for F2 never differed from the experimenter's measured frequencies by more than 40 Hertz (Since the second formant frequency was used to determine the cutoff frequency for the low-pass filtered vowels, this measurement was critical).

Two copies of each of the excised vowels were made. One copy of each vowel was low-pass filtered using a bank of two Stamford Research Model SR 650 Programmable Filters. The two filters were used to create an extremely steep cut-off, of approximately 120 dB per octave. The cut-off frequencies were set to 100 Hz above the center frequency of F2. This effectively removed F3 and higher formant information. However, the bandwidths were sufficiently narrow so that the second formant energy was always clearly present after filtering. The vowel stimuli were then converted back to

analog signals and recorded on a digital tape using a Sony portable DAT, model TCD-D7.

Four stimulus types were generated: phonated-filtered, whispered-filtered, phonated-unfiltered, and whispered-unfiltered. The total number of stimuli generated was 540 (9 vowels x 15 speakers x 4 types).

METHOD:

Subjects:

Listeners were four males and four females, 18 to 21 years of age. All were native New Yorkers whose first language is English. They were drawn from the same community as the speakers, and the experimenter judged them to have a dialect that is the same as that of the speakers. None of the subjects had previously participated in perceptual research, but all received training in the required task (see the procedure section, below). Subjects received $20 or $25 for their participation, depending on the length of time required to complete the training and testing.

Procedure:

Stimuli were presented over Telephonics TDH50 headphones with an impedance of 60 Ohms, using a Sony portable DAT, model TCD-D7 with an output impedance of 27 Ohms. The headphone response was tested and found to be essentially flat up to 8000 Hertz.

Training:

All listeners went through a training period immediately before testing. This training was provided for

two reasons which were unrelated to the perceptual issues being investigated. One reason was to provide practice with the task. The second reason was the possibility that the subject might have initially associated a printed keyword with a vowel other than the one that it was intended to exemplify.

The training was done as follows: After listening to each of the key words read by the experimenter, the subject was asked to read the words without further cueing from the experimenter. When the subject read all of the keywords with what the experimenter judged to be the desired pronunciation, training for the listening task was initiated.

A training tape was constructed with each vowel prepared in the manner described above for the preparation of the actual listening tape. The training tape had two blocks of stimuli. One block consisted of phonated-filtered vowels and the other of whispered-unfiltered vowels. Each block contained two sets of nine vowels in a mixed-speaker condition. Nine vowels were spoken by an adult female speaker, and nine vowels were spoken by a child speaker. The speakers were different from those used for the actual listening task, but they were selected using the criteria described above. Both sets were randomized together as a

group. Although this training task was more limited in its conditions than the actual listening task, it was sufficiently representative of the actual task, while having the advantage of being relatively brief (2 speakers x 2 stimulus types x 9 vowels = 36 stimuli).

The subjects listened to each stimulus and circled a keyword on a response sheet. Feedback was provided immediately following each response, and "misperceived" vowels were replayed. All subjects met an 80% correct criterion, before proceeding to the actual testing. This criterion is similar to levels found by Tartter (1991) for whispered vowels and higher than Kallail and Emanuel's (1984) levels [but see Eklund and Traunmüller (1997)]. In fact, even for phonated vowels, researchers have rarely found identification rates for mixed isolated vowels at rates above 90% without special conditions [i.e. the use by Assman, Nearey, & Hogan, (1982) of spoken responses transcribed by phonetically-trained experimenters].

Testing:

The stimuli were presented in blocks of 45, at 4.5-second intervals, with an interblock interval of 30 seconds.

Each of the four stimulus types (phonated-filtered, whispered-filtered, phonated-unfiltered, and whispered-unfiltered) was presented separately in a blocked-speaker condition and in a mixed-speaker condition. In the blocked-speaker condition, a different randomization was used for each speaker's set of nine vowels. In the mixed-speaker condition, all fifteen speakers' vowels were randomized as a group.

The randomization was done for the phonated and whispered vowels separately in the following manner: Each set of each subject's nine vowels was divided into three subsets composed /I æ ʌ/, /I ɑ U/ and /ɛ ɔ u/. Each subset of each speaker's vowels was combined into one of three blocks so that five of each of the three subsets appeared in each block. Each block was then randomized. Although no attempt was made to prevent the same speaker's vowels from being presented consecutively, this occurred infrequently, since each speaker had only three vowels per block. No attempt was made to prevent the same vowel from being presented consecutively by different speakers.

All subjects were presented with all eight groups of stimuli (4 types x 2 conditions), for a total of 1080 stimuli (4 types x 2 conditions x 15 speakers x 9 vowels). The order

of presentation of these groups was counterbalanced across subjects.

Following Assman et al. (1982), subjects were provided with answer sheets containing lines of the following keywords: "heed, hid, head, had, hod, hawed, hud, hood, who'd." Each answer sheet contained 45 lines of the key words, so that each stimulus type required three separate answer sheets. The 45 lines on the answer sheets were divided into blocks of 9 lines to enable subjects to match each speaker's set of 9 vowels in the blocked-speaker condition to a block of 9 lines on the response sheet. Subjects were asked to listen to each stimulus and to circle the word on the response sheet that contained the same vowel. Subjects were told that in both the mixed-speaker and blocked-speaker conditions they might hear the same vowel twice within the same block, and that they might not hear every vowel even once within a given block.

ANALYSIS:

After the data were collected, it became clear that, even in the blocked, unfiltered condition, there were certain vowels that had been consistently identified as other than the vowel intended by the speaker. In a subsidiary

examination, the experimenter and an experienced listener agreed that three phonated and 15 whispered vowels had been "mispronounced" and these were omitted from the data pool and thus from the analysis of results. (See Appendix B for a detailed description of how vowels were selected for elimination, and see the discussion section for some implications of the necessity of eliminating these vowels.)

Mean percent correct scores were calculated for each subject for each stimulus type in each condition. No subjects scored 100% for any of the stimulus types in either condition, and a visual analysis of the data indicated what appeared to be close to standard distributions for the phonated and whispered vowels examined separately. In addition an examination of the standard deviations for the various conditions indicated that there was homogeneity of variance. Therefore, analysis of variance was used to analyze the data. Confusion matrices were prepared for the speakers as a group for each condition in order to analyze the nature of the identification errors.

SPECIFIC PREDICTIONS

Most published research that has examined vowel classification for mixed-speaker vs. blocked-speaker conditions has shown higher vowel classification error rates for mixed-speaker condition than for blocked-speaker conditions (see Nearey, 1989 for a review). However, some studies have found very low error rates for both conditions (Assman et. al., 1982; Kahn, 1978; Macchi, 1980).

Pilot work by this author indicated that stimuli presented in mixed-speaker condition were likely to be identified with less accuracy than the same stimuli presented in blocked-speaker condition for some conditions. This relationship was imperative for this research, since the usefulness of F0 and F3 as parameters in perceptual vowel normalization was judged based on the degree to which these parameters improved classification rates in the mixed-speaker condition to the levels found in the blocked-speaker condition.

In addition, published research by Eklund and Traunmüller (1997), Kallail and Emanuel (1984) and Tartter (1991) has shown that whispered vowels are significantly less well identified than phonated vowels. Pilot work by this

author indicated that a similar finding was the likely outcome of this research.

The effect of the removal through filtering of information carried in frequencies above the F2 frequency was not considered to be highly predictable prior to this study. Lehiste and Peterson (1959) and Peterson (1954) reported high levels of identification for low-pass vowels of this nature. However, it was considered unlikely that filtered vowels, lacking upper formant information, would be identified more accurately than unfiltered vowels.

POSSIBLE OUTCOMES AND THEIR IMPLICATIONS

This study depends on a trend (in at least some of the stimulus types) for lower correct vowel classification rates in mixed-speaker condition than in blocked-speaker condition. A mixed-speaker condition is one that creates the greatest potential for the perceptual ambiguity based solely on F1 and F2 frequency values. Therefore, it is likely that a mixed-speaker condition would result in lower classification rates than a blocked-speaker condition. If this effect is less strongly seen when F0 information is available, such as for the phonated vowels, or when F3 information is available, such as for the unfiltered vowel, the results would indicate

that F0 or F3 is likely to be used in perceptual normalization. This finding is implicitly predicted by both Syrdal (1985) and Miller (1989), for both F0 and F3.

For example, it was possible that blocked-speaker condition stimuli would be classified with ten percentage points greater accuracy than mixed-condition stimuli for whispered (no F0 information) vowels, but equally well for phonated vowels. Similarly, it was possible that blocked-speaker condition stimuli would be classified with ten percentage points greater accuracy than mixed-condition stimuli for filtered (no F3 or higher formant information) vowels, but equally well for unfiltered vowels. Respectively, these findings would strongly indicate that F0 information or F3 information is critically useful in perceptual normalization of vowels, as implied by Syrdal (1985) and Miller (1989). To the extent that either of these interactions is less strongly seen, the evidence for a perceptual normalization of steady-state vowels relying on these parameters would be decreased.

CHAPTER III

RESULTS

GROUP MEANS AND STANDARD DEVIATIONS

Each subject's data were scored for percent correct identification of the target vowels. The percent correct identification scores for each of the eight subjects for each of the eight conditions are reported in appendix C.

The group means and standard deviations were calculated for each of the eight conditions and are presented in Table III.i.

Table III.i. Percent correct identification of phonated and whispered vowels with and without filtering to remove F3 and upper formants, under two presentation conditions. The data displayed are group means and standard deviations for eight listeners.

	BLOCKED-SPEAKER	MIXED-SPEAKER
Phonated Unfiltered	86.8 (7.9)	87.4 (5.1)
Phonated Filtered	87.2 (7.2)	85.2 (7.7)
Whispered Unfiltered	79.7 (7.7)	74.5 (5.4)
Whispered Filtered	74.1 (5.8)	66.1 (7.9)

The fact that the standard deviations for the conditions with the highest mean correct classification scores are in the same range as the standard deviations for the conditions with lower means suggests that there is homogeneity of variance. In addition, very few individual scores were above 95%. These means and standard deviations are displayed visually in figure 3.1.

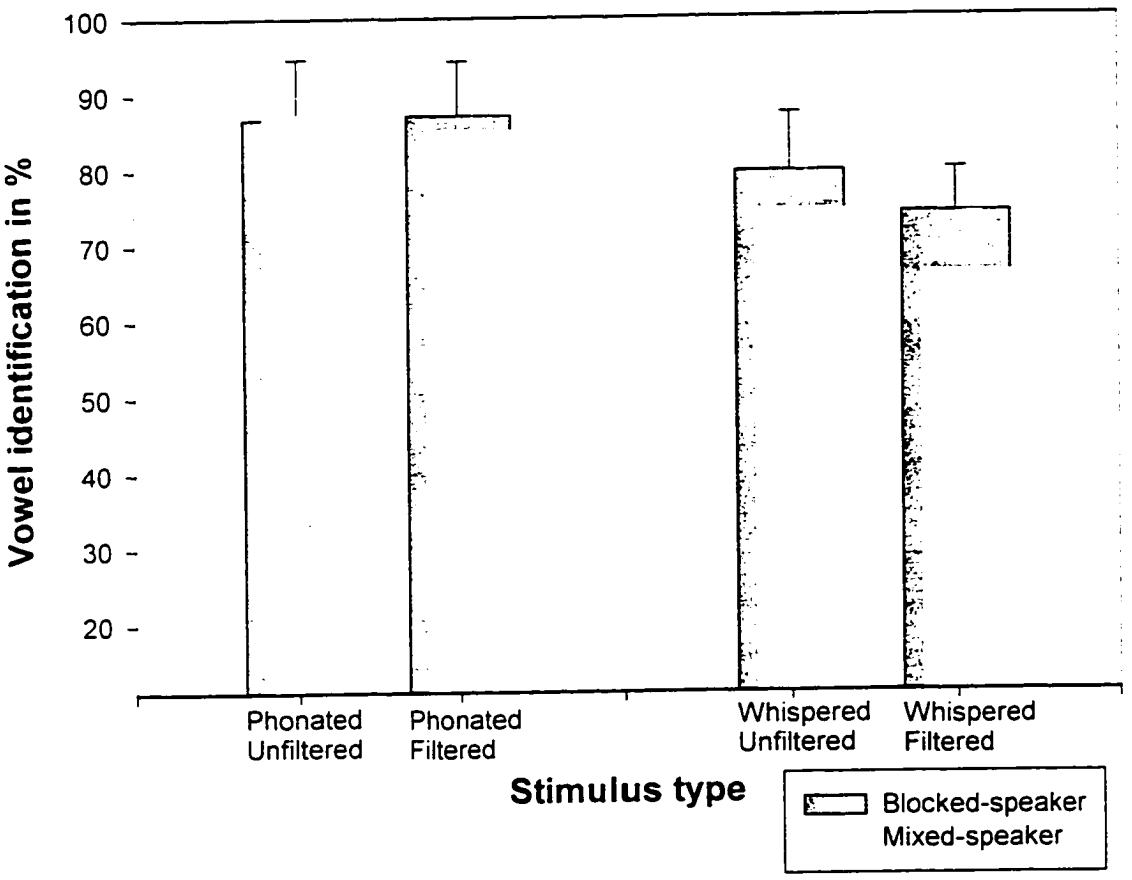


Figure 3.1. Percent correct identification of phonated and whispered vowels, with and without filtering to remove F3 and upper formants, under two experimental conditions. The data displayed are group means and standard deviations for eight listeners.

ANALYSIS OF VARIANCE

A three-way ANOVA was carried out on the mean identification data. The factors in the design were speaker variability (blocked or mixed speakers), phonatory type [fundamental frequency availability] (phonated or whispered) and upper formant availability (unfiltered or filtered). The error terms are provided by interlistener differences. The results of the ANOVA are reported in Table III.ii.

Table III.ii. Three-way ANOVA for presentation condition x phonatory type x availability of upper formants. All effects.

Condition = Blocked-Speaker/Mixed-speaker
Phonation = Phonated/Whispered
Filtering = Unfiltered/Filtered

Source of Variance	Estimated Mean Square	Degrees of Freedom	Error Term	Estimated Mean Square	Degrees of Freedom	F Ratio	p level	
Condition	214.51	1	CxSubject	20.83	7	10.30	0.015	*
Phonation	2725.49	1	PxS	16.60	7	164.16	0.000004	*
Filtering	245.12	1	FxS	20.68	7	11.85	0.011	*
CxP	139.45	1	CxPxS	29.82	7	4.68	0.067	
CxF	28.88	1	CxFxS	13.67	7	2.11	0.189	
PxF	145.41	1	PxFxS	4.93	7	29.48	0.001	*
CxPxF	0.07	1	CxPxFxS	17.20	7	0.00	0.952	

MAIN EFFECTS

All three main effects were significant. These will be reported sequentially.

1. First, percent correct identification was higher in the blocked than in the mixed conditions. On average 82.0% of the vowels were correctly identified in the blocked condition, compared with 78.3% in the mixed condition. The difference of approximately 3.6 percentage points was significant ($p = 0.015$). Thus, stimuli presented in the mixed-speaker condition were more difficult to identify as the intended vowel than stimuli presented in the blocked-speaker condition.

2. Second, an extremely robust effect was found for phonatory type. Percent correct identification was higher for the phonated than for whispered stimulus. On average 86.7% of the phonated vowels were correctly identified, compared with 73.6% of the whispered vowels. The difference of approximately 9 percentage points was significant ($p < 0.001$). Thus, the whispered vowels were more difficult to identify than the phonated vowels.

3. Third, percent correct identification was higher for the unfiltered than for the filtered condition. On average 82.1% of the unfiltered vowels were correctly identified, compared with 78.2% of the filtered vowels. The difference of nearly 4 percentage points was significant ($p = 0.011$). Thus, the vowels with F3 and upper formants filtered out were more difficult to identify than the unfiltered vowels.

INTERACTIONS

The comparisons most important for the crucial questions posed by this dissertation relate to interactions between speaker variability and phonatory type (i.e. fundamental frequency availability), and between speaker variability and upper formant availability. The existence of such interactions would suggest the perceptual importance of FO and of upper formants in vowel normalization.

Figure 3.2 is a graph of the differences in percent correct identification between the blocked-speaker and mixed-speaker conditions for the phonated and whispered vowels. For phonated vowels, percent correct scores are less than one percentage point lower in the mixed-speaker condition (86.3%) than in the blocked-speaker condition (87.0%). For whispered vowels, percent correct scores were approximately 6.5 percentage points lower in the mixed-speaker condition (70.3%) than in the blocked-speaker condition (76.9%). This effect approached the 0.05% level of significance.

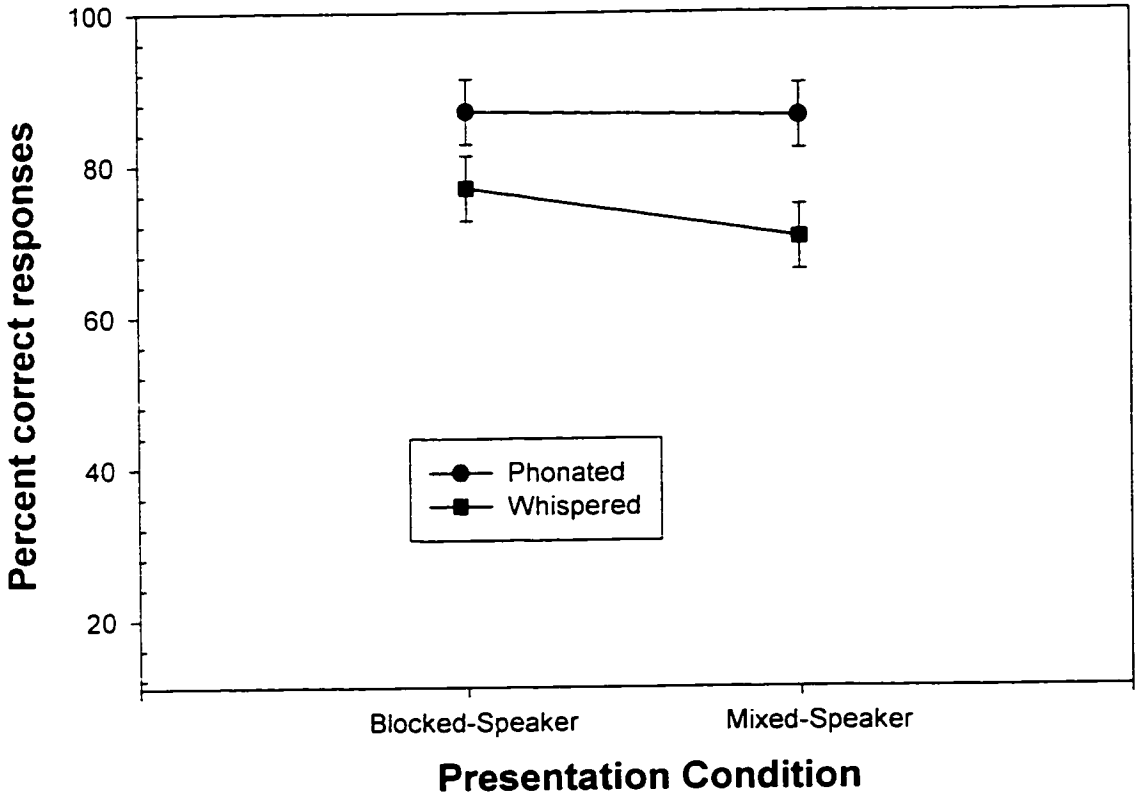


Figure 3.2
Interaction of presentation condition and phonatory type
Error terms are from the pooled group data

The interaction between speaker variability and upper formant availability is of critical importance for the perceptual reality of vowel normalization schemes that use F3 as an additional cue for vowel perception. Figure 3.3 shows a graph of the differences in percent correct identification between the blocked-speaker and the mixed-speaker conditions for the unfiltered and filtered vowels. For unfiltered vowels, percent correct scores were approximately 2.3 percentage points lower in the mixed-speaker condition (80.9%) than in the blocked-speaker condition (83.3%). For filtered vowels, percent correct scores were approximately five percentage points lower in the mixed-speaker condition (75.7%) than in the blocked-speaker condition (80.6%). The magnitude of this interaction is small (a difference between loss in percentage correct scores when switching from blocked-speaker to mixed-speaker conditions for unfiltered and filtered vowels of approximately 2.6%). In addition, this effect failed to reach significance (p. = 0.189).

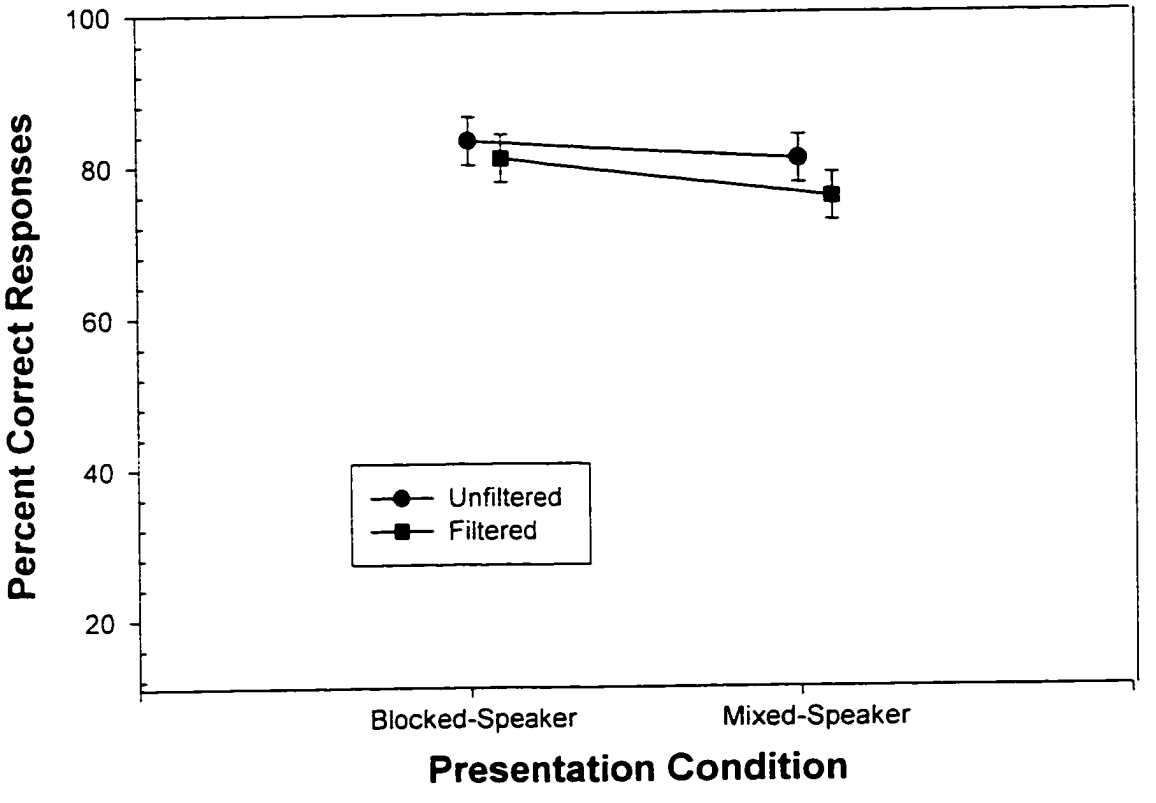


Figure 3.3
Interaction of presentation and upper formant availability
Error terms are from the pooled group data

Possibly the most important finding of this study was one that was not expected: The interaction between phonatory type and availability of upper formants was extremely robust. This significant interaction suggests that the third formant is more important for whispered vowel perception than for phonated vowel perception.

Figure 3.4 is a graph of the differences in percent correct identification associated with unfiltered versus filtered conditions for the voiced and whispered vowels. For phonated vowels, percent correct scores were less than one percentage point lower in the filtered (86.2%) condition than in the unfiltered condition (87.1%). For whispered vowels, percent correct scores were approximately seven percentage points lower for the filtered stimuli (70.1%) than for the unfiltered stimuli (77.1%). This effect was significant ($p = 0.001$).

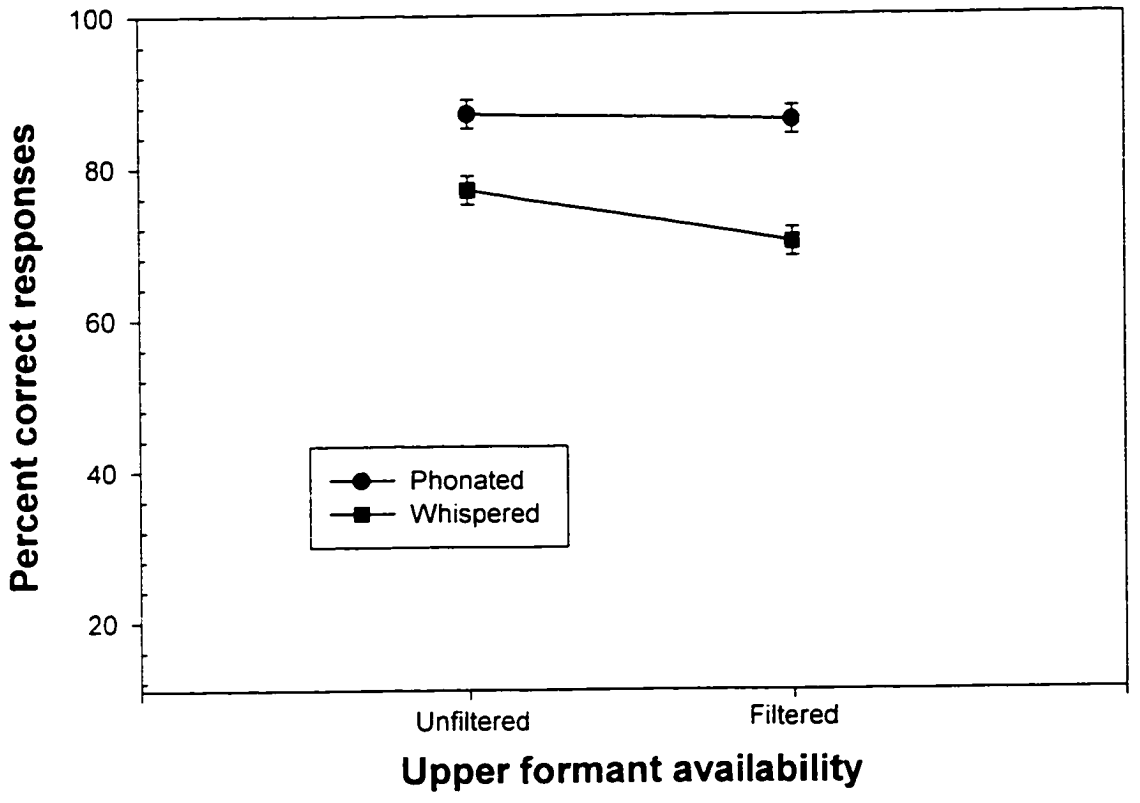


Figure 3.4
Interaction of phonatory type and upper formant availability
Error terms are from the pooled group data

SPEAKER TYPE EFFECTS

It was hypothesized that the effects of the various conditions on classification rates and their interactions might be different for the male, female and child speakers. A four-way ANOVA was carried out on the mean classification data to investigate the effects of speaker type.

A main effect was found for speaker type. Vowels produced by female and child speakers were classified with approximately 82.6% and 81.5% accuracy respectively. Vowels produced by male speakers were classified with approximately 76.1% accuracy. This effect was significant at the 0.001% level. No interactions between speaker type and any other condition were found to be significant.

CONFUSION MATRICES

Confusion matrices were prepared for each speaker for each of the eight conditions. The eight grand confusion matrices for all of the speakers as a group for each of the eight conditions are presented in table III.iii.

Table III.iii. Confusion matrices for the eight conditions (presentation condition x phonatory type x upper formant availability). All speakers and listeners collapsed.

Confusion matrix for blocked-speaker, phonated, unfiltered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	117	0	3	0	0	0	0	0	0	120
/ɪ/	0	120	0	0	0	0	0	0	0	120
/ɛ/	0	10	102	7	1	0	0	0	0	120
/æ/	0	0	28	80	4	0	0	0	0	112
/ɑ/	0	0	0	1	95	14	10	0	0	120
/ɔ/	0	0	0	0	0	111	0	1	0	112
/ʌ/	0	0	0	0	20	1	82	9	0	112
/ʊ/	0	0	0	0	3	1	18	98	0	120
/u/	0	0	0	0	0	1	0	7	112	120
Total Response	117	130	133	88	123	128	110	115	112	1056

Confusion matrix for mixed-speaker, phonated, unfiltered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	120	0	0	0	0	0	0	0	0	120
/ɪ/	0	118	1	0	0	0	0	1	0	120
/ɛ/	0	12	99	7	1	0	1	0	0	120
/æ/	0	0	26	83	2	0	0	1	0	112
/ɑ/	0	0	0	4	88	9	18	1	0	120
/ɔ/	0	0	0	0	0	112	0	0	0	112
/ʌ/	0	0	0	0	13	5	82	12	0	112
/ʊ/	0	0	1	1	2	1	14	101	0	120
/u/	0	0	0	0	0	0	0	0	120	120
Total Response	120	130	127	95	106	127	115	116	120	1056

Table III.iii. (continued). Confusion matrices for the eight conditions (presentation condition x phonatory type x upper formant availability). All speakers and listeners collapsed.

Confusion matrix for blocked-speaker, phonated, filtered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	120	0	0	0	0	0	0	0	0	120
/ɪ/	0	120	0	0	0	0	0	0	0	120
/ɛ/	0	10	100	9	0	0	0	1	0	120
/æ/	0	0	21	89	2	0	0	0	0	112
/ɑ/	0	0	0	0	95	11	14	0	0	120
/ɔ/	0	0	0	0	1	111	0	0	0	112
/ʌ/	0	0	0	0	14	4	82	12	0	112
/ʊ/	0	1	0	0	0	1	18	100	0	120
/u/	0	0	0	0	0	0	0	16	104	120
Total Response	120	131	121	98	112	127	114	129	104	1056

Confusion matrix for mixed-speaker, phonated, filtered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	120	0	0	0	0	0	0	0	0	120
/ɪ/	0	120	0	0	0	0	0	0	0	120
/ɛ/	0	12	94	13	0	0	0	1	0	120
/æ/	0	0	24	88	0	0	0	0	0	112
/ɑ/	0	0	0	7	83	12	15	3	0	120
/ɔ/	0	0	0	0	2	108	0	2	0	112
/ʌ/	0	0	0	2	17	8	73	12	0	112
/ʊ/	0	1	0	0	2	1	18	98	0	120
/u/	0	0	0	0	0	0	0	4	116	120
Total Response	120	133	118	110	104	129	106	120	116	1056

Table III.iii. (continued). Confusion matrices for the eight conditions (presentation condition x phonatory type x upper formant availability). All speakers and listeners collapsed.

Confusion matrix for blocked-speaker, whispered, unfiltered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	112	7	0	0	1	0	0	0	0	120
/ɪ/	0	93	11	0	0	0	0	0	0	104
/ɛ/	0	15	89	8	0	0	0	0	0	112
/æ/	0	3	28	72	0	0	0	1	0	104
/ɑ/	0	0	1	4	62	11	14	4	0	96
/ɔ/	0	0	0	0	5	107	4	4	0	120
/ʌ/	0	0	0	0	15	4	59	18	0	96
/ʊ/	0	0	0	0	1	3	25	56	3	88
/u/	0	0	0	0	0	0	0	5	115	120
Total Response	112	118	129	84	84	125	102	88	118	960

Confusion matrix for mixed-speaker, whispered, unfiltered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	113	6	0	0	0	1	0	0	0	120
/ɪ/	1	84	18	0	0	0	1	0	0	104
/ɛ/	0	22	84	6	0	0	0	0	0	112
/æ/	0	1	38	62	1	1	0	1	0	104
/ɑ/	0	1	0	0	54	13	23	5	0	96
/ɔ/	0	0	0	0	16	95	5	4	0	120
/ʌ/	0	0	1	0	24	5	53	13	0	96
/ʊ/	0	0	0	0	0	1	26	57	4	88
/u/	0	0	0	0	0	1	0	6	113	120
Total Response	114	114	141	68	95	117	108	86	117	960

Table III.iii. (continued). Confusion matrices for the eight conditions (presentation condition x phonatory type x upper formant availability). All speakers and listeners collapsed.

Confusion matrix for blocked-speaker, whispered, filtered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	114	5	1	0	0	0	0	0	0	120
/ɪ/	2	94	8	0	0	0	0	0	0	104
/ɛ/	0	16	78	18	0	0	0	0	0	112
/æ/	1	0	42	61	0	0	0	0	0	104
/ɑ/	0	0	0	3	61	18	10	4	0	96
/ɔ/	0	0	0	1	12	87	9	9	2	120
/ʌ/	0	0	0	0	25	6	45	20	0	96
/ʊ/	0	0	0	0	4	3	23	55	3	88
/u/	0	0	0	0	0	0	0	4	116	120
Total Response:	117	115	129	83	102	114	87	92	121	960

Confusion matrix for mixed-speaker, whispered, filtered vowels.

	Response									
Vowel	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/	Total Stimuli
/i/	105	15	0	0	0	0	0	0	0	120
/ɪ/	8	68	26	2	0	0	0	0	0	104
/ɛ/	0	21	82	8	0	0	0	0	1	112
/æ/	0	2	45	57	0	0	0	0	0	104
/ɑ/	0	0	0	3	55	21	14	3	0	96
/ɔ/	0	0	0	3	14	79	11	11	2	120
/ʌ/	0	0	0	2	36	6	40	12	0	96
/ʊ/	0	0	0	1	4	8	27	46	2	88
/u/	0	0	0	0	0	0	0	17	103	120
Total Response:	113	106	153	76	109	114	92	89	108	960

As can be seen, in general, the patterns of confusions were similar for the various conditions, although there were more errors for some conditions than others. One noteworthy difference in patterns of confusion is that found for /i/, /I/ and /ɔ/. In all conditions containing phonated vowels, /i/, /I/ and /ɔ/ were almost perfectly classified. This was not the case for the whispered vowels, in which /i/ was classified as /I/ approximately 5% of the time. The difference between phonated and whispered conditions for /I/ classification was even more striking, with approximately 10% to 35% errors for the whispered /I/ vowels (primarily /ɛ/ confusions) depending on the condition. Whispered /ɔ/ vowels were also classified with approximately 10% to 35% errors (confused primarily with other back vowels) depending on the condition.

One additional difference in the patterns of errors for the different conditions is that there tended to be more front-back confusions for mixed-speaker conditions than in the corresponding blocked speaker conditions.

Although it was hypothesized that differences for patterns of errors between the phonated and whispered vowels, other than those mentioned above, would be found, little

evidence of such differences can be seen in the confusion matrices.

ACOUSTIC DATA

The focus of this dissertation is on perception, not production or acoustics. However, a sample of the values obtained for the first three formants for each of the phonated and whispered vowels is presented in appendix D. These reported data include an adult male, an adult female and child speaker. In addition, the relevant data from the Peterson and Barney (1952) study are included for comparison. The formant values found in this study are reasonably similar to the mean data reported by Peterson and Barney. The lack of perfect correspondence may stem from a variety of sources. These include: dialectical variation, individual differences, differences in the phonological context and differences in measurement techniques.

CHAPTER IV

DISCUSSION

INTRODUCTION

The critical issues for this dissertation relate to findings regarding the interactions between phonatory type and presentation condition and between filtering and presentation condition. For organizational reasons, however, the main effects will be discussed first, followed by a discussion relating to the interactions, followed by a discussion of general issues relating to this study and to vowel perception research in general.

MAIN EFFECTS

Presentation condition

Stimuli presented in the mixed-speaker condition were more difficult to identify as the intended vowel than stimuli presented in the blocked-speaker condition. Two points, however, must be noted. The first point is that although the interaction between phonatory type and presentation condition failed to reach significance [$F(1,7)=4.68$, with an MS error of 29.82 $p. = 0.068$], the main effect found for presentation condition was largely a result of the whispered vowels. In fact, phonated unfiltered vowels were identified slightly

better in the mixed-speaker condition (87.4%) than phonated unfiltered stimuli presented in the blocked-speaker condition (86.8%). This finding is not consistent with the reports of Strange, Verbrugge, Shankweiler and Edman (1976). Although Assman, Nearey and Hogan (1982) reported significantly worse identification of isolated vowels presented in a mixed-speaker than in a blocked-speaker condition, the actual percent errors for the two conditions in their study were 4.09% and 5.43% respectively. The importance of this 1.34% difference is unclear. (The low error rates for Assman et al. (1982) may be attributable to the use of phonetically trained listeners.) The second point that must be made is that even for the whispered vowels in this study, the magnitude of this difference was less than seven percentage points.

Phonatory type

The second significant main effect that was found was for phonatory type. The whispered vowels were significantly more difficult to identify than the voiced vowels. This finding is consistent with those reported by Kallail and Emanuel (1984), Tartter (1990) and Eklund and Traunmüller (1997).

As few studies have been performed comparing perception of voiced and whispered vowels, a comparison of some of the results for this part of the study with earlier studies is presented. Table IV.i shows the percent correct scores for the mixed unfiltered conditions for both voiced and whispered vowels in this study and for the comparable conditions in Kallail and Emanuel (1984), Tartter (1991) and Eklund and Traunmüller (1997). These studies used different methodologies and vowel sets from the present study; Tartter's (1991) study is more similar to the present study than are Kallail and Emanuel's (1984) and Eklund and Traunmüller's (1997) studies. The results are shown for Tartter's second experiment, in which she used subjects who received training, as did the subjects in the present study. Eklund and Traunmuller (1997) reported male and female listeners' correct classification rates for male and female speakers separately; these have been collapsed for purposes of comparison.

Table IV.i. Mean percent correct classification for the vowels in four studies.

STUDY	Phonated	Whispered	Difference
Halberstam (1998)	87.4%	74.5%	12.9%
Kallail & Emanuel (1984)	80.0%	65.0%	15.0%
Tartter (1991)	92.0%	81.6%	10.4%
Eklund & Traunmüller (1997)	95.3%	88.0%	7.3%

We see that both the mean correct classification scores and the magnitude of the difference between correct classification scores for whispered vs. voiced vowels for the present study fell between the values reported in the Kallail and Emanuel (1984) and the Tartter (1991) studies. Unlike the present study, Tartter's (1991) vowel set included /ə /, which was very well recognized. Other than this vowel, the vowel sets used by Tartter (1991) and the present study were identical, although Tartter's vowels were produced and presented in a [hVd] context. Table IV.ii presents the percent correct scores for the mixed-speaker unfiltered conditions for both phonated and whispered vowels in the present study and for the comparable conditions in Tartter (1991), with the percent correct scores for /ɚ / removed. For the same vowel set, the present study's results are very similar to Tartter's (1991) results.

Table IV.ii. Percent correct classification in this study and in a revised data set adapted from Tartter (1991).

STUDY	Phonated	Whispered	Difference
Halberstam (1998)	87.40%	74.50%	12.90%
Tartter (1991) (/ɚ/ removed)	91.30%	79.60%	11.70%

Eklund and Traunmuller's (1997) findings are problematical, particularly because the formant frequencies of their stimuli are not steady-state. Although neither Tartter (1991), Kallail and Emanuel (1984) nor the present study provide specific evaluations of the degree to which their stimuli are steady-state, all of these studies attempted to elicit quasi-steady-state vowels. This starkly contrasts with the Eklund and Traunmüller (1997) study. Although Eklund and Traunmüller describe their stimuli as "the names of the nine letters which are used in Swedish orthography to represent the vowels," claiming that "with the exception of i, the names of these vowels are pronounced as the long vowels usually represented by the letter in question," they provide an acoustic analysis that makes it clear that most of their stimuli are characterized by a good deal of diphthongization. Their claim that their study is comparable with Kallail and Emanuel (1984) and Tartter (1991) is thus open to question.

Assuming that Eklund and Traunmüller (1997) are basically familiar with the English language, they compound this problem when, in attributing their high recognition rates to the use of "genuine words," they "predict that

similar high recognition rates can be obtained in an experiment using the English names of the letters a, e, i, o, r and u." It is likely that they are correct in their prediction, but for a different reason. The letter names that they mention are highly identifiable sounds and/or diphthongs ([eI], [Ii], [aI], [oU], [aə], and [ju]).

In addition, Eklund and Traunmüller (1997) do not consider a large body of research (see Nearey, 1989 for a review) when they suggest that "the presence of consonants in these stimuli [Tartter (1991)] did not improve, but rather impaired the recognition of the vowels." Some studies have shown advantages for consonantal context (Strange Verbrugge, Shankweiler, & Edman, 1976; Kahn, 1978), and some studies (Assman, Nearey, & Hogan, 1982; Macchi, 1980), have shown no advantage for consonantal context, but no study has suggested that consonantal context places listeners at a disadvantage for vowel perception.

The rationale for Tartter's (1991) study included some reasonable criticisms of Kallail and Emanuel's (1984) design, and the argument that "the results reported by Kallail and Emanuel were a poor estimate of whispered vowel identification" (Tartter, 1991). Of the three studies cited, Tartter's seems to be designed best. Therefore, the

similarity between the results for the whispered vowels in the present study with Tartter's results is viewed as strengthening the validity of both studies.

Returning to the present study, there are at least two possible interpretations of the finding that whispered vowels are not perceived as well as phonated vowels. One possibility is that fundamental frequency information is important in vowel perception. Another possibility is that the difficulties in perceiving whispered vowels relate to the shifted formant values for whispered vowels when compared to the formant values for phonated vowels. Although no statistical analyses were performed on the formant frequencies measured in this study, a clear pattern of higher formant frequencies for the first three formants was noted (see appendix D for a sample of the formant data from some of the speakers from this dissertation demonstrating this pattern). This tendency has also been reported by several researchers (Peterson, 1961; Kallail and Emanuel, 1984; Eklund and Traunmüller, 1997).

There is some evidence that the difficulties in perceiving whispered vowels relate to the shifted formant values for whispered vowels when compared to the formant values for phonated vowels, rather than to a lack of F0

information. Nusbaum and Morin (1992) performed a study similar to this study, but they used synthetic vowels. While Nusbaum and Morin found significant differences in the classification rates for phonated and whispered vowels in a mixed-speaker condition, they found no significant differences in the classification rates for phonated and whispered vowels in a blocked-speaker condition. A possible interpretation of the different findings of the Nusbaum and Morin (1992) study and the present study relates to the formant frequencies of the stimuli. The present study used phonated and whispered vowels with naturally produced formants which, as discussed earlier, tend to be higher for the whispered vowels. Nusbaum and Morin's (1992) study used phonated and whispered synthetic stimuli with identical formant frequencies. It may be that the difference between the findings indicates that the poorer classification for whispered vowels found in this study is caused by the differences between the formant frequencies of the whispered and phonated vowels.

The two possibilities mentioned above to explain the better classification rates for the phonated versus whispered vowels can be more thoroughly explored, at least initially, through the use, within the same study, of synthetic

"whispered" vowels using formant frequency values derived from both phonated and whispered production values. Improved perception for whispered vowels based on phonated formant frequency values would be a strong indication that it is the altered formant frequencies that cause whispered vowel perception to be poorer than phonated vowel perception. (A negative finding would be less easily interpreted.)

It is interesting to note that the general patterns of errors in whispered and phonated vowel classification in the present study were similar (see the confusion matrices in the results section). This finding may suggest that the cause of poorer classification for the whispered vowels may be related to a lack of F0 information, rather than altered formant values. If the altered formant values were involved, the pattern of errors would be expected to be different for the two phonatory types. This is because the alterations in formant frequencies found in the whispered vowels would move whispered vowels into the range of other phonated vowels within the F1 x F2 space.

Upper formant availability

The third significant main effect was for upper formant availability. The filtered vowels were significantly more

difficult to identify than the unfiltered vowels. It is important to note, however, that this effect was largely a result of the whispered vowels. As previously reported, a significant interaction was found for phonatory type and availability of upper formants. In fact, percent correct identification collapsed across blocked-speaker and mixed-speaker conditions for phonated unfiltered vowels (87.1%) and phonated filtered vowels (86.2%) differed by less than one percentage point. This issue will be addressed in more detail in the discussion about the interaction between phonatory type and upper formant availability.

It was hypothesized that the effects of filtering might be different for front and back vowels. Back vowels typically have F3 of lower amplitudes than front vowels. In addition, F3 is typically more distant from F2 than 3.5 Bark. A three to 3.5 Bark range has been shown to be the limit for perceptual integration of two formants (Chistovich and Lubinskaja, 1979; Chistovich, 1985). As a result of the above, it was logical to expect that back vowels would be less affected by filtering. Table IV.iii shows the effect of filtering on percent correct classification rates for each vowel for the phonated and whispered vowels separately and as a group.

Table IV.iii. Difference in percent correct classification in percentage points by vowel as a result of filtering. Data are from group means.

PHONATORY TYPE	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/
Phonated	1.3	0.8	-2.9	6.3	-2.1	-1.8	-4.0	-0.4	-5.0
Whispered	-2.5	-7.2	-5.8	-7.7	0.0	-15.0	-14.1	-6.8	-3.8
Phonated and Whispered	-0.6	-2.9	-4.3	-0.5	-1.2	-8.6	-8.7	-3.1	-4.4

As can be seen from table IV.iii, there was some tendency in the opposite direction from the predicted tendency. Back vowels generally had a greater loss in correct classification rates as a result of filtering than did front vowels. This was a general trend, with some exceptions. No obvious explanation for this trend was recognized.

INTERACTIONS

As mentioned earlier, the most critical issues for this dissertation relate to findings regarding the interactions between phonatory type and presentation condition, and between availability of upper formants and presentation condition. This is because both F0 information (available in the phonated and not in the whispered vowels) and F3 frequency information (available in the unfiltered and not in the filtered vowels) have been used as parameters in vowel normalization schemes. If these schemes reflect actual perceptual processes, then presenting vowels in a mixed-speaker condition rather than a blocked speaker condition would be expected to result in different losses in classification rates for different stimuli types. Specifically, correct classification rates would show greater differences between blocked-speaker and mixed-speaker conditions for stimuli lacking F0 or F3 frequency information than for stimuli containing such information.

Presentation condition x phonatory type

There was a greater difference in mean correct classification scores between blocked-speaker and mixed-speaker condition for the whispered vowels than for the

phonated vowels. For phonated vowels, percent correct scores were less than one percentage point lower in the mixed-speaker condition (86.3%) than in the blocked-speaker condition (87.0%). For whispered vowels, percent correct scores were more than six and a half percentage points lower in the mixed-speaker condition (70.3%) than in the blocked-speaker condition (76.9%).

As previously reported, the interaction between phonatory type and presentation condition narrowly missed significance at the 5% level (p. = 0.068). Note, however, that there was a large error term for this effect (MS error of 29.82), indicating considerable inter-subject variability for this interaction. In other words, some of the subjects may have been more susceptible to the combined effects of phonatory type and presentation condition than others. Even so, the chance that there is no interaction in the means of the populations represented by this sample is less than seven in 100.

To have achieved a significant effect at the 0.05 significance level for presentation condition, a difference of 4.3 percentage points between blocked-speaker and mixed-speaker condition would have to have been obtained. The differences mentioned just above show that for phonated

vowels the difference in percentage points between blocked-speaker and mixed-speaker condition was well below 4.3, while for whispered vowels, the difference was well above 4.3.

Both post hoc LSD and Newman Keuls revealed that mixed-speaker condition vowels were perceived significantly less well for the whispered vowels at the 0.01 significance level, but not for the phonated vowels.

The results suggest that fundamental frequency may be used as a third cue in (addition to F1 and F2) to enhance vowel identification in a perceptually difficult context, such as the mixed-speaker condition of this study. One of the central claims of several vowel normalization schemes (Miller, 1989; Syrdal 1985), that F0 contributes to accurate vowel classification, is supported by the perceptual findings of this study.

Presentation condition and upper formant availability

The interaction between presentation condition and upper formant availability is of critical importance for the perceptual reality of vowel normalization schemes that use F3 as an additional cue for vowel perception. As previously reported, the interaction between phonatory type and

presentation condition was not found to be significant in this study at the 5% level (p. = 0.189).

For unfiltered vowels, percent correct scores were approximately 2.5 percentage points lower in the mixed-speaker condition (80.9%) than in the blocked-speaker condition (83.3%). For filtered vowels, percent correct scores were approximately five percentage points lower in the mixed-speaker condition (75.7%) than in the blocked-speaker condition (80.6%). It is seen that the magnitude of this interaction is small (a difference between loss in percentage correct scores when switching from blocked-speaker to mixed-speaker conditions for unfiltered and filtered vowels of approximately 2.6 percentage points).

Thus, the perceptual results do not provide strong support for the use of upper formants, including F3, as elements in a vowel normalization scheme. Upper formants are particularly useful as a third cue (in addition to F1 and F2) for vowel perception in an perceptually challenging context, such as the mixed-speaker condition of this study. Even if the argument is put forth that the lack of significance found in this study related to insufficient sensitivity in the design, as mentioned, the magnitude of the effect found in this study was quite small.

The finding that upper formants are unlikely to be very important as a normalization parameter must be considered a negative one for the leading intrinsic normalization schemes in which the use of third formant frequency information is viewed as crucial (Miller, 1989; Syrdal and Gopal, 1986). The finding must also be considered a negative one for Wakita's (1977) proposal which uses F4 as a parameter in vowel normalization.

The comments made above should not be construed as suggesting that F3 information has no effect on vowel perception (as opposed to vowel normalization). As we have seen, several studies have found other kinds of effects on vowel perception that relate to F3 frequency (Ainsworth (1975), Fujisaki and Kawashima (1968), Nearey (1989), Slawson (1968) and others. However, none of these effects have been directly related to vowel classification rates.

The fact that F3 information has been shown to affect vowel perception, but has not been shown by this study to have an effect relating to vowel normalization, is not paradoxical. The effects of F3 may actually relate directly to F2 formant perception, rather than influencing vowel perception in the manner suggested by vowel normalization theories. Chistovich has published a number of articles

comparing the perception of two formant and four formant stimuli that imply that a single formant percept is the outcome when more than one acoustic formant falls within a center of gravity of less than 3.5 Bark (Chistovich and Lubinskaja, 1979; Chistovich, 1985). Bladon et al. (1984,) has made more explicit claims suggesting that the acoustic energy contained in formants that are close together on a Bark scale are perceptually incorporated into a unitary peak as a result of the operations of a version of a "model of peripheral auditory representation" based on work by Schroeder. According to this model, F2 and F3 contribute to a sole auditory peak when they are close together on a Bark scale. It is therefore possible that changes in F3 can cause changes in vowel identity as a result of changes in a perceived unitary peak that typically relates to both F2 and F3 information. An effect of this kind would not necessarily have any implications for vowel normalization.

Phonatory type and upper formant availability

As previously stated, what might be considered to be the most interesting finding of this study was one that was not expected: The interaction between phonatory type and availability of upper formants was extremely robust. Upper

formant information was more important for whispered vowel perception than for phonated vowel perception. For phonated vowels, percent correct scores were less than one percentage point lower in the filtered (86.2%) condition than in the unfiltered condition (87.1%). For whispered vowels, percent correct scores were nearly seven percentage points lower in the filtered stimuli (70.1%) than in the unfiltered stimuli (77.1%). The implication of this finding is that a loss of third formant information is relatively inconsequential for the perception of phonated vowels, but of considerable importance for the perception of whispered vowels. This finding is partially consistent with Nusbaum and Morin (1992) who reported that "filtering F3 affected recognition accuracy only when F0 was missing." However, Nusbaum and Morin (1992) reported that filtering influenced classification rates only in a mixed-speaker condition; synthetic whispered vowels presented in a blocked-speaker condition in their study were perceived with no significant effect of filtering. In the present study, filtering had a similar effect on whispered vowel classification in both blocked-speaker and mixed-speaker conditions.

This finding may contain clues about vowel perceptual processes that have not been understood in the past. Namely,

it would appear that the perception of phonated vowels relies on F1, F2 and fundamental frequency, while F3 does not seem to be important in the perception of phonated vowels. Perception of whispered vowels, however, seems to rely much more heavily on F3 and/or higher formants.

An attempt to further interpret this finding is warranted. One way to explain the reliance on F3 in whispered and not in phonated vowels is that there are two perceptual process for vowels, one for and one for whispered vowels, which rely on different parameters. This is unparsimonious.

A second, more plausible way to interpret the perceptual reliance on F3 in whispered and not in voiced vowels is that perception of isolated steady-state vowels reaches its maximum effectiveness when listeners have access to F1, F2 and fundamental frequency. For this reason, access to F3 frequency information is of little or no additional importance in perception of voiced vowels.

Whispered vowels, however, can be considered to be acoustically deficient for purposes of perception. This deficiency may be the result of two factors. The first factor is that, as previously mentioned, the formant structure of whispered vowels differs considerably from that

of phonated vowels (Peterson, 1961; Kallail and Emanuel, 1984; Eklund and Traunmüller, 1997). The second factor that may contribute to the perceptual deficiency of whispered vowels is that the absence of phonation leaves F1 and F2 frequency information without the perceptual anchor provided by fundamental frequency information. This assumes that F0 information is used as a parameter in vowel classification, a hypothesis supported by this study, as well as others (Ainsworth (1975); Fahey, Diehl and Traunmüller (1996); Fujisaki and Kawashima (1968); Hirahara and Kato (1992); Hoemeke and Diehl (1994); Nearey (1989); Slawson (1968), and Traunmüller (1981).

As a result of the deficient nature of whispered vowels, the acoustic information provided by F3 (and possibly higher formants) is effectively used for vowel perception. The result of the ANOVA for the interaction between speaker variability and upper formant availability indicates that the benefit of upper formant information to the perception of whispered vowels may be limited to increasing the availability of useful perceptual cues for individual vowels. The additional information was not shown to be significantly useful for disambiguating vowels in a mixed-speaker condition when compared to a blocked-speaker condition. Thus, although

F3 seems to be used for whispered vowel perception, it does not appear to be useful for perceptual vowel normalization. Only F0 appears to perform that role.

"GOOD" VOWELS AS STIMULI

An additional issue concerning stimuli used in vowel perception research has emerged from this dissertation. This issue will be treated in this section.

Studies examining perception of naturally produced vowels, as well as different conditions within the same studies, have found inconsistent classification rates. It is very common for researchers to explain that these differences in classification rates are caused by differences in elicitation techniques (Tartter, 1991; Macchi, 1980; Kahn, 1978; Strange and Gottfried, 1980). A brief consideration of such concerns brings to light a fundamental problem in using naturally spoken vowels for perceptual research. If one is to label a response other than that of the intended vowel as an "error" and analyze it as such, it is imperative to know that the vowel in question is a good exemplar of such stimuli.

To illustrate: If a speaker is asked to produce an /I/ for the purposes of stimulus creation for a perceptual study and the experimenter perceives the produced vowel to be an /u/, it scarcely needs to be said that the experimenter would consider the produced vowel to be unsuitable for an /I/ stimulus. It would be unreasonable to consider an /u/

response on the part of a subject presented with this stimulus to be an error. Wakita (1977) makes a similar point when he suggests that errors in classification by his formula should logically be considered to be errors only when the vowel stimulus is not "an apparent mispronunciation" [referred to by Wakita (1977) as a Type I error].

Similarly, it is common for published studies to comment on the attempt to match dialect of speakers and listeners (Macchi, 1980; Mullennix, Pisoni and Martin, 1988). The implication is that the authors recognize that labeling a response as an error when speaker and listener have been poorly matched for dialect is inappropriate. A more reasonable interpretation is that the "incorrect" response is a result of the fact that the intended vowel is not a good exemplar of the intended vowel in the listener's vowel system. Again, Wakita (1977) makes a similar point when he suggests that errors in classification by his formula might arguably be considered to be errors only when the vowel stimulus is not "an allophonic variation which deviated considerably from the defined target vowel" [referred to by Wakita (1977) as a Type II error].

Carrying this reasoning an additional step, it becomes clear that the ideal stimuli for research investigating the

cues adequate for vowel normalization and/or perception are "perfectly" produced vowels located at the "center-of-gravity" for each vowel in a particular dialect. It is obvious that naturally produced conversational speech typically will not provide listeners with idealized vowels of this nature. However, perceptual compensation for coarticulation, and the intrinsic redundancy of speech including the contributions of dynamic cues (see page 15 for a review) might explain the fact that listeners have high levels of correct vowel perception despite the fact that naturally produced vowels may never reach these idealized targets. In addition, vowel perception can benefit from the possibility of calibration available in typical single speaker situations, as theorized by extrinsic vowel normalization models. Finally, vowels in conversational speech, that might otherwise be perceived as other than the vowel intended by the speaker, may benefit from a top-down process of enhancement based on linguistic knowledge.

One example of the importance of the use of "good" productions of vowels comes from this dissertation. A reexamination of the stimuli that were perceived by a majority of the subjects as vowels other than the intended vowel revealed that some were poor exemplars of the intended

vowel. In fact, there were three (of 135) voiced and eleven (of 135) whispered vowels that were identified by a trained listener as vowels other than the intended vowel in a forced choice multiple presentation task. It is most logical to assume that vowels such as these cannot be considered to be appropriate stimuli for the intended vowel; they are more logically considered to be misarticulated. The most extreme case known to the author of this dissertation of data of this nature being removed from consideration in vowel perception experimentation is in a recent article by Eklund and Traunmüller (1997). They write that "in some cases, the vowels were not identified correctly by a majority of listeners. However, in such cases the listeners typically agreed in their identifications. All these tokens should therefore be considered as wrongly produced rather than wrongly perceived, and so they have not been included into the acoustic analysis."

It should be understood that although the present author believes that Eklund and Traunmüller's (1997) statement represents a step in the right direction, their criterion of majority "errors" for considering a vowel to be incorrectly produced is unnecessarily arbitrary. Consider that the purpose of their listening experiment was to learn

something about the difficulties of perceiving voiced and whispered vowels in a mixed-speaker condition. Majority "errors" may merely indicate the difficulty of perceiving the stimulus within the experimental conditions. The use of a trained listener, familiar with the nature of the stimuli and given the opportunity for multiple presentations may well be a more rigorous procedure.

A reasonable argument can be made that the removal of vowels considered to be mispronounced by a trained listener under the conditions mentioned above is inappropriate. Arguably, such removal may be viewed as removing correctly articulated vowels that are difficult to perceive (even with an "easy" experimental condition) because of the inherent difficulties in perceiving vowels, specifically steady-state vowels. However, because other tokens of these same vowels, that can be correctly identified, can be elicited from the same speaker, this author believes that the procedure used here to eliminate "incorrectly produced" vowels is justified, and that similar procedures should be adopted in future vowel perceptual research.

Ultimately, the issue may depend to some degree on a priori assumptions. If it is believed that "steady-state" vowels contain cues sufficient for good identification, then

it is logical to insure that stimuli used in "steady-state" vowel experimentation are good exemplars. Researchers with the perspective of Strange, Jenkins and others present arguments such as the following: "Since we are interested in vowel perception under these 'normal' conditions of minor dialectic and idiolectic variation, we have chosen to incorporate such variation into the stimulus materials used in our studies" (Strange and Gottfried, 1980).

The present author accepts the above as a basis for vowel research that investigates vowel perception in coarticulated or conversational speech. But, having "incorporated such variation," it is not surprising that Strange and colleagues have in the past come to the conclusion "consonantal environment specifies vowel identity" (Strange, Jenkins and Edman, 1978). Recognition of the experimental findings of others (Kahn, 1977; Macchi, 1980) who carefully avoided such variation has also led Strange to somewhat different conclusions: "Studies...have shown that under certain conditions of...selection of productions, vowels produced as isolated, sustained tokens (or in neutral /h/-vowel syllables) can be identified with high accuracy" (Strange and Gottfried, 1980).

We may then argue that studies investigating the perceptual mechanisms underlying the highly accurate perception of "steady-state" vowels must attempt to control goodness of production of naturally-produced stimuli. Some of the types of vowel research that logically require this monitoring of goodness of vowel productions are: formant frequency production studies, data analytic vowel normalization studies and steady-state vowel perception studies.

CONCLUSIONS

Perceptual verification of intrinsic normalization

Of the numerous intrinsic vowel normalization schemes that have been proposed, none has been critically examined from the perspective of perception. Considering the purported perceptual processes that these schemes are designed to model, this fact is quite surprising.

Typically, as mentioned earlier, the methods of experimentally testing the validity of normalization techniques has been limited to an analysis of how well published vowel fundamental and formant frequency data can be classified using a proposed normalization schemes (Disner, 1980; Gerstman, 1968; Hillenbrand and Gayvert, 1993; Lobanov, 1971; Miller, 1989; Syrdal and Gopal, 1984 and others). In this sense, these analyses are data-analytic rather than perceptual in nature. The drawback of this approach is that acoustic information that may provide an algorithm with values that enhance vowel classification may have little to do with perception.

Syrdal (1985) and Miller (1989) have proposed elaborate perceptual models using vowel normalization techniques, utilizing F0 and F3 as parameters in addition to F1 and F2. Syrdal (1985) analyzes the validity of her vowel

normalization using a linear discriminant classifier for classification of the Peterson and Barney (1952) data. Miller analyzes the validity of his vowel normalization by the number of vowels from a corpus composed of various vowel data bases that fall within his hand-drawn "vowel slabs" derived from the differences between F3, F2, F1 and the "sensory reference" (based on F0) for each vowel. However, Neither Syrdal nor Miller have published research showing that the variables (F0, F1, F2 and F3) used as input variables in their vowel normalization schemes are actually important for the perceptual processes that these schemes model. Indeed, as mentioned earlier, another data-analytic evaluation of vowel classification based on various parameter sets found that the addition of F0 values to F1 and F2 provides nearly identical benefit for vowel categorization as the addition of both F0 and F3 (Hillenbrand and Gayvert, 1993). This highlights another troubling characteristic of data-analytic evaluations of vowel normalization schemes: The chosen method of analyzing the data can result in different correct classification rates from one study to another.

The fact that F0 and/or F3 can have an effect on vowel and formant perception has been demonstrated (Ainsworth, 1975; Fahey, Diehl and Traunmüller, 1996; Fujisaki and

Kawashima, 1968; Hirahara and Kato, 1992; Hoemeke and Diehl, 1994; Kallail and Emanuel, 1984; Nearey, 1989; Slawson, 1968; Traunmüller, 1981). This research, however, has not been directly relevant to vowel normalization. Particularly in light of the findings of the present study, it is important that future research focus directly on this issue.

Implications for vowel perception

The most profound implications of this research are precisely in the area of the assumptions that have been made in theories of vowel perception. Primary among these is the growing assumption that F3 may be important in the characterization and identification of vowels. This assumption is one that seems to have grown in response to the recognition of the variability of the two lowest formant values for the same vowels for different speakers.

Previous to 1970, vowels were usually assumed to be perceptually characterized by the lowest two formants. For instance, Fujisaki and Kawashima (1968), whose research focused on the perceptual effects of F0 and formants higher than F2 on vowel perception, nonetheless begin their article as follows: "Among various acoustic parameters that constitute the frequency spectrum of the vowel, frequencies

of the two lowest formants are known to be most important in the determination of their phonemic quality."

As a result of the research by Fujisaki and Kawashima (1968) and others, F3 was recognized as a potential cause of perceptual effects on (synthetic) vowels. This has led some researchers to emphasize the importance of F3 in the perception of vowels, particularly those who deal directly with problems related to the aforementioned variability of formant values. Minifie (1973) writes "Even though many formants may be produced during vowel production, only the first three formants appear to be used by listeners in differentiating vowel sounds."

Miller (1989) makes what is possibly the strongest statement supporting the importance of F3 in vowel perception: "Presently, it is well established that the locations of the three prominences of the short-term spectrum of the vowel waveform are highly correlated with the perceived color of the vowel." No reference is given for this "well established" inclusion of F3, and the present author does not know of any research directly supporting the importance of F3 in vowel perception (with the exception of the effects found in this study, primarily relating to whispered vowels). As the data reported on in this study

suggest, F3 may actually have little impact on the perceptual classification of phonated vowels.

It is clear why Miller would like F3's importance to be well established; his (1989) article incorporates F3 into a perceptual model. The present study casts the perceptual validity of this model, as well as that of Syrdal (1985) into serious doubt.

On the other hand, in the case of whispered vowels, this study provides support for a strong effect of F3 on perceptual classification of naturally produced "steady state" vowels.

The findings of this study indicate that F3 provides information about vowel identity, but that this information is redundant in the case of voiced vowels, which are identified equally well whether or not they contain a third formant.

In the case of whispered vowels, it is hypothesized that the loss of FO information and/or the raised formant frequency values when compared to the same vowel produced with phonation by the same speaker, results in a loss of information important to vowel identification. In this case, the information contained in F3 relating to vowel identity is

no longer redundant, and can significantly improve correct identification rates.

The information contained in F3, however, improves vowel classification generally, rather than specifically in a mixed-speaker condition. It appears, therefore, that F3 may not be useful in resolving the acoustic ambiguities encountered when listening to vowels spoken by many speakers. That is to say, F3 may not be useful for perceptual vowel normalization.

F0, on the other hand, appears to be important both for vowel identification across the board, as well as for vowel normalization. Access to F0 information allows listeners to compensate for any ambiguities that are involved in perceiving vowels produced by multiple speakers. This view of a vowel perceptual process that incorporates F0 as a normalizing parameter may be gaining currency, as it is one that exploits a natural and automatic phenomenon: speakers tend to have fundamental frequencies that correlate with their formant frequencies (Honda, 1997). This study represents one step in strengthening the view that F0 has an important role in vowel perception, particularly with regard to vowel normalization.

APPENDIX A

HISTORICAL REVIEW

INTRODUCTION

This dissertation investigates the psychoacoustic reality of some of the assumptions made by vowel normalization schemes. Since vowel normalization schemes are an attempt to compensate for the underspecification of vowels based solely on formant frequencies, an historical review of the development of the source-filter model of vowel production follows. The source-filter model provides the scientific explanation of formants. This review also traces the parallel development of the traditional belief that formants are the primary cue for perception of vowels.

Although the early scientific research into the acoustic basis of vowels was often cited in the past, credit for each of the key aspects of the theoretical underpinnings of the source-filter model has never been properly assigned. This lack of proper assignation is probably due to several factors, including an imperfect understanding of the early state of the science by later scientists, failure to understand the explanations of earlier researchers, and a reliance on secondary sources about the original theories proposed by the early experimenters.

SCIENTIFIC ORIGINS OF SOURCE-FILTER THEORY

The first experiments designed to provide information about vowel acoustics were performed by Willis (1830). He used reeds that vibrated at different frequencies, connected to tubes of different lengths, and listened to the resultant sounds. These sounds were reportedly perceived as vowel-like sounds. Willis demonstrated that irrespective of the reed's frequency, a tone was produced that related to the length of the resonating tube, and that the perceived vowel was related to this secondary tone, or cavity tone.

In 1837, Wheatstone published an article about the history and development of synthetic speech that included a review of Willis' work, as well as an explanation of the way in which the characteristic tones are produced in a tube of a given length. Wheatstone clearly recognized that the frequency of each tone heard in an apparatus such as that used by Willis must be a "simple multiple of that of the original sounding body."

Although these two events have historically been recognized as important in the development of the science of vowel acoustics, neither has been written about, in any detail, in almost 50 years. In addition there was a good

deal of confusion among reviewers of the research at the turn of the century and in the early part of this century about whether Willis' or Wheatstone's view was correct, or whether both were correct.

The idea that Willis was the first to scientifically study vowel acoustics has been widely accepted. Thus Paget (1930) writes "The first systematic investigation of the nature of vowel sounds---verified by their synthetic production by models---was published in 1829 by Robert Willis." With the possible exception of Kratzenstein's work (Kratzenstein, 1780), which shall be mentioned later, this statement is true.

Similarly, Russell (1928) writes that the theory developed by Willis "was the earliest of our really modern scientific theories evolved from and resting on substantial experimental evidence." Others who consider Willis to have begun the scientific study of vowel acoustics include Miller (1937) and Chiba and Kajiyama (1941).

HARMONIC AND INHARMONIC THEORIES

Willis' theory has been referred to as the inharmonic theory, while Wheatestone's theory has been referred to as the harmonic theory (Russell, 1928; Fletcher, 1929; Chiba and

Kajiyama, 1941). The different views that the two terms represented were clear to most early and some later writers who correctly understood these two theories. In fact, "inharmonic" referred to the fact that "the glottal tone might be inharmonic to the cavity tone" (Russell, 1928; Chiba & Kajiyama, 1941). Some of these writers believed that Willis was partially or fully correct. Many later writers assumed meanings for these terms that did not always reflect the original theories, and often made erroneous conclusions about the relative merit of the two views.

Thus, following a lengthy analysis of the views of Willis and Wheatstone, Russell (1928) writes that "we have talked of the two as being in conflict with each other. And...it would appear that they are."

Rayleigh (1896) takes the opposing position. He suggests that "both ways of regarding the subject are legitimate, and not inconsistent with one another." He believed that the harmonic theory was appropriate for vowels with low formants, and the inharmonic theory was appropriate for vowels with high formants. Fletcher (1929), who, unlike Rayleigh, had a view of the acoustics of vowels consistent with vowel acoustics as understood today (see p.10), also saw no conflict between the two theories. He writes that "the

difference in the two theories is not, as some suppose, a difference in the conception of what is going on while the vowel sounds are being produced, but in the method of reporting or describing the motions in definite physical terms" (p.49). He attributes the term "inharmonic" to the fact that "according to this theory, the puffs do not necessarily follow each other."

Both Rayleigh's and Fletcher's statements seem to be unsupported by the original sources, and may partially stem from a reliance on secondary sources. In particular, Fletcher cites Helmholtz as having shown that "these two theories were different only in the point of view and the method of representing the same mechanism of vowel production..." (Fletcher, 46). A similar comment was made by Lindsay (1966): "von Helmholtz later pointed out that both ideas have elements of correctness and modern research has confirmed this view."

Helmholtz actually did not actually provide any experimental or theoretical support for Willis' position (Helmholtz, 1877/1954), and modern research has shown that certain aspects of Willis' theory are incorrect. Helmholtz did state that "Willis's description of the motion of the sound for vowels is certainly not a great way from the truth;

but it only assigns the mode in which the motion of the air ensues, and not the corresponding reactions which this produces in the air." It is difficult to say to what extent Helmholtz believed that Willis' theory was incompatible with his own; in any event, it shall soon be seen that it was.

Since it is difficult for speech scientists today to imagine how an "inharmonic" theory of vowel production might work, and since the theory as originally formulated has rarely been fully described, I will review Willis' theory. As this theory has been shown to be not fully consistent with the contemporary view of the acoustics of vowels, it is necessary to point out that in one key respect, Willis was correct: Willis was the first to argue that the tones associated with vowels were produced as a result of the length of the cavity. Although Helmholtz (1877/1954) and McKendrick (1898) suggest that Donders was the first to state that the vocal tract was "tuned" to different pitches for different vowels, this is not entirely accurate. Both Willis' (1830) and Wheatstone's (1837) articles clearly intend the tubes which they discuss to be analogous to the vocal tract.

At the same time, Willis was not the first to suggest that the oral cavity was "tuned" to different pitches for

different vowels. Although Helmholtz (1877/1954) incorrectly credits Donders (whose work was later than Willis') with this discovery, he cites several earlier researchers with "incomplete observations" of the same nature. These researchers attempted to determine which musical notes were associated with various vowels. Willis' actual contribution turns out to be his recognition that the "cavity tone" relates to the length of the tube and its resultant characteristic resonance. It is only in this sense that Russell (1928) can be considered to be correct when he states that "the cavity tone theory should be credited to Willis."

EULER AND KRATZENSTEIN

As an introduction to Willis' theory, it is important to point out that key aspects of speech acoustics were completely unknown to Willis and his contemporaries. Euler, who is cited by Willis as inspiring his theory (Willis, 1830), writes (Euler, 17--?/1802) "in explaining the theory of sounds, I considered only two respects in which sounds could differ: the one regarded the force of sound...The other difference of sounds is totally independent of this and refers to flat and sharp, according to which we say some are low and some are high... There is still another remarkable

difference among the simple sounds, which seems to have escaped the attention of the philosophers. Two sounds may be of equal force, and in accord with the same note of the harpsichord, and yet very different to the ear...but it is impossible to describe wherein this consists."

Euler, it is seen, was aware only of frequency and intensity at the time that he wrote this letter. He did not recognize that notes played on instruments or sung by voices were actually complex sounds. In fact, he writes that "in music, we employ only those sounds which are denominated simple" (Euler, 17--?/1802). Euler recognized the problems inherent in understanding the perceptual differences in vowels when the acoustic theory is limited to frequency and intensity. He writes "When the vowel 'a' is pronounced or sung, the sound is quite different than e,i,o,u, or ai pronounced or sung, though on the same tone...no investigation of philosophers has hitherto unfolded this mystery."

Kratzenstein had recognized as early as 1779 that vowels were associated with different cavity shapes (Kratzenstein, 1780). Indeed, he won the yearly prize of the Royal Academy of St. Petersburg for his successful demonstration of the synthetic production of vowels using

reeds attached to cavities of various shapes. Nevertheless, his explanation of the acoustical nature of vowels is limited to an analogy comparing the reflection of light off mirrors and the reflection of sound off the walls of the cavity. He does not indicate any recognition of the idea that vowels are associated with tones of a particular frequency.

WILLIS' INHARMONIC SCIENCE

Willis' theory (Willis, 1830) is an attempt to "unfold the mystery" of vowel acoustics described by Euler (17--?/ 1802) but it is based on two principles, both of which fail to recognize the existence or importance of the complex nature of the glottal source. The first deals with the way in which "cavity tones" (formants) are produced. Willis did not recognize that formants are composed of harmonics present in the glottal wave-form, and resonated in the vocal tract. Instead, he believed that these "cavity tones" arise as a result of an excitation of the vocal tract itself, which causes the air in the vocal tract to vibrate at its natural frequency.

Willis (1830) states "According to Euler, if a single pulsation be excited at the bottom of a tube closed at one end, it will travel to the mouth of this tube with the

velocity of sound. Here an echo of the pulsation will be formed which will run back again, be reflected from the bottom of the tube, and again present itself at the mouth where a new echo will be produced, and so on in succession...The effect therefore will be the propagation from the mouth of the tube of a succession of equidistant pulsations alternately condensed and rarefied, at intervals corresponding to the time required for the pulse to travel down the tube and back again; that is to say, a short burst of the musical note corresponding to a stopped pipe of the length in question, will be produced."

A careful reading of this and other passages reveals that Willis believed, since the velocity of sound is a constant, the sound wave produced by bursts of energy at any frequency will arrive at the opening of the tube at a moment in time related solely to the length of the tube. It will be reflected back to the closed end of the tube, and be reflected back to the opening, again arriving at a moment in time relating to the velocity of sound and the length of the tube. Since in his view the only necessary variables to determine the frequency of the resultant tone are the velocity of sound, and the length of the tube, his theory successfully explained why reeds of any frequency applied to

tubes of a fixed length were found to produce the same vowel quality.

Obviously, this aspect of his theory is not fully supported by modern science. Its major flaw is the lack of recognition of the importance of phase. That is to say, that it is not sufficient for a sound wave to arrive at the tube or mouth opening to create a resonant sound. Rather, it is the coincidence of an antinode and the tube or mouth opening that is crucial in determining the most resonant frequency. In addition there is no recognition of the existence of any sound other than a simple tone at the most resonant frequency; harmonics are not considered in this explanation.

The second principle upon which Willis based his theory, is his explanation of how the cavity tone can coexist with a musical tone (actually the fundamental). His explanation follows: "if we examine the nature of our series, we shall find it merely to consist of the repetition of one musical note in such rapid succession as to produce another." (It shall be seen that his meaning here is that the cavity tone, as described above, repeats itself at a frequency matching the fundamental.)

Willis continues "it has long been established that any noise whatever, repeated in such rapid succession at

equidistant intervals, as to make its individual pulses insensible, will produce a musical note For instance, let the musical note of the pipe be g'', and that of the reed c', which is 512 beats in a second, then their combined effect is g''...g''...g''...g''... (512 in a second) in such rapid equidistant succession as to produce c', g'' in this case producing the same effect as any other noise, so that we might expect a priori, that one idea suggested by this compound sound would be the musical note c'."

Willis provides the following illustration "...g' is peculiar to the vowel A: when this is repeated 512 times in a second, the pitch of the sound is c', and the vowel is A: if by means of another reed applied to the same pipe it were repeated 340 times in a second, the pitch would be f, but the vowel still A."

Helmholtz (1877/1954) understood this aspect of Willis' theory when he writes "Willis imagines that the pulses of air which produce the vowel qualities are themselves tones which rapidly die away..."

Fletcher (1929), was not completely accurate in the passage cited earlier : "the difference in the two theories is not, as some suppose, a difference in the conception of what is going on while the vowel sounds are being produced,

but in the method of reporting or describing the motions in definite physical terms" (p.49). It does seem to be the case that Willis (1830) is describing the acoustics in the temporal domain, while Helmholtz (1877/1954) is focused on the frequency domain. However, it appears to be clear that Willis' (1830) science is indeed different from Helmholtz's (1877/1954) in terms of "what is going on while the vowel sounds are being produced," and is also incomplete from a modern perspective. Specifically, it is incomplete in its lack of recognition of the complex nature of the glottal waveform, and the relationship of the components of the cavity tone to the original complex glottal waveform.

In any event, when we wish to assign credit for the development of the source-filter/formant-based model of vowel production and perception, Willis can only be given credit for relating the cavity tones associated with specific vowels to cavity length. He and all who followed his view did not have a comprehensive knowledge of the formation and composition of these complex tones. Additionally, as Helmholtz pointed out, several researchers had previously related vowels to cavity tones of specific frequencies.

However, even in this limited aspect of the theory of the acoustics of vowels, Willis was only partially consistent

with the contemporary view of vowel acoustics. He is only partially consistent with the contemporary view because he believed that vowel resonance is solely a product of tube or vocal tract length. His theory has no role for vocal tract shape. In fact, Willis (1830) criticizes the efforts of Kratzenstein (1780) and Kempelen (1791) by stating "Kempelen's mistake, like that of every other writer on this subject, appears to lie in the tacit assumption, that every illustration is to be sought for in the form and action of the organs of speech themselves, which however paradoxical the assertion may appear, can never, I contend, lead to any accurate knowledge of the subject." This view is related to Willis' (1830) statement that "cylinders of the same length give the same vowel, whatever their diameter and figure."

Despite describing vowel production in terms of oral cavity shape (Wheatstone, 1837), Wheatstone seems to have accepted the idea that cavity shape is unrelated to vowel identity. In reference to Willis' experiments, he writes that "From these experiments it is evident that the forms stated by Kratzenstein, as producing the different vowels, are perfectly arbitrary. The entire series of vowels can be produced from tubes of either of his forms by merely changing its dimensions."

Clearly, this dismissal of the importance of the shape of the vocal tract stemmed from the correct recognition of the importance of the "cavity tone" in vowel acoustics coupled with the mistaken belief that there can only be one characteristic "cavity tone" for a given vowel. If only one cavity tone exists, and the cavity tone rises and falls with decreases and increases in the size of the cavity, there is no place for tract shape in a theory of vowel acoustics. Only the recognition of the possibility of more than one formant could restore the importance of the shape of the vocal tract in a scientific theory of the acoustics of vowels.

THE IMPORTANCE OF AUDIBILITY OF HARMONICS

The development of the "harmonic resonance" theory of vowel acoustics required, at a minimum, the recognition of the fact that tone-producing vibration can result in a complex sound, consisting of a fundamental, and a series of audible harmonics.

Lindsay (1966) reports that both Bernoulli (1755) and Euler (1753) did understand that it was "possible for a string to vibrate in such a way that a multitude of simple-harmonic oscillations are present at the same time and that

each contributes independently to the resultant vibration." This was the "principle of the coexistence of small oscillations, also referred to as the principle of superposition" (Lindsay, 1966).

However, this knowledge did not result in a recognition of audibility of harmonics in a complex tone. In a letter to Lagrange, Euler (1759/1973) writes that "I am completely of your opinion that the harmonic sounds which M. Rameau believes he hears from the same string actually came from other vibrating bodies." It would seem that many researchers at the time believed that only the fundamental frequency could be perceived even when produced by a complex periodic vibration. Clearly, Euler did not have knowledge of auditory perception of complex periodic sound.

For this reason Euler's (1727/1973) explanation of vowel acoustics should not be misinterpreted as an attempt at a description of formants or harmonics. He writes that "the vibratory motion of the air is increased at the head of the windpipe and then is modified in various ways in the oral cavity, by which the low and high pitch voice is modulated and various vowels are formed." From his remarks in his letter to Lagrange (Euler, 1759/1973), he clearly could not

have been describing resonance as it is understood today, since he did not recognize the presence of harmonics at all.

Chladni's (1787/1973) work makes it clear that many scientists in the end of the 1700s were aware that vibrating bodies could vibrate in several modes, producing fundamental and harmonic tones respectively. But it is less clear if many recognized that these tones were commonly produced simultaneously.

Plomp (1964) cites the French scientist Mersenne as reporting in 1636 that freely vibrating strings produced several tones at once, and that such sounds "follow the ratio of 1, 2, 3, 4, 5." Sauveur is cited by both Plomp (1964) and Lindsay (1966) as also recognizing at the turn of the eighteenth century that vibrating strings could produce sounds of several of its harmonics simultaneously. Plomp (1964) writes that Sauveur's research was "used by Rameau about 20 years later as a physical basis for his theory on musical harmony."

However, Plomp (1964) appears to be mistaken when he suggests that following Sauveur's work "the existence of perceptible overtones was generally accepted." From the passage quoted earlier from Euler (1759/1973) stating that "I am completely of your opinion that the harmonic sounds which

M. Rameau believes he hears from the same string actually came from other vibrating bodies," it is clear that while Rameau may have understood this aspect of sound vibration, Euler and Lagrange, two of the greatest scientists who studied acoustics in the eighteenth century (as well as the early nineteenth century in the case of Lagrange), very specifically did not accept Sauveur and Rameau's view.

Ohm's (1843/1973) article was the first to argue that Fourier's theorem could be used to demonstrate that any complex periodic sound could be described as the sum of simple periodic sounds, and that the "particular quality or timbre of actual musical sounds is due to combinations of simple tones of commensurable frequencies" (Lindsay, 1966). He obviously recognized the existence of complex periodic sound, and understood its composition. Ohm's (1843/1973) article was the basis of a revolution in the way in which sound was thought about and studied. It would also allow a more sophisticated view of voiced vowels to develop.

WHEATSTONE AND THE HARMONIC THEORY

Having established the degree of Willis' contribution to vowel acoustics and perception, it is important to determine what Wheatstone added to this endeavor. Although

Wheatstone has been credited with originating the theory that harmonics existing in the source (vocal folds) are reinforced by the filter (vocal tract), it will be shown that his advance was actually incremental in nature; Wheatstone did not formulate his theory precisely in the manner just described.

Wheatstone has traditionally been given credit for the "harmonic theory." Following a statement by Helmholtz (1877/1954) that "the vowels of speech are in reality tones produced by membranous tongues (the vocal chords), with a resonance chamber (the mouth) capable of...reinforcing at different times different partials of the compound tone to which it is applied," Ellis, in a footnote to his translation of Helmholtz, writes that "the theory of vowel tones was first enumerated by Wheatstone in a criticism, unfortunately little known, on Willis's experiments."

Russell (1928) very painstakingly (and accurately) points out the puzzling fact that Wheatstone seemed to see "no conflict between Willis's theory" and his own. He concludes, however, by stating that "Wheatstone clearly postulated the 'overtone' or 'harmonic cavity tone' theory, and is entitled to that credit which is so often given to Helmholtz."

Similarly, although Scripture (1904) believed that Wheatstone was wrong, he nonetheless writes that Wheatstone "supposed that the vowels arose from the vibrations of the vocal cords through the strengthening of certain overtones by the resonance of the mouth."

DID WHEATSTONE ORIGINATE THE SOURCE-FILTER MODEL?

A careful reading of Wheatstone reveals a slightly different picture. Key to a proper understanding of Wheatstone is his term "multiple resonance." Those who have discussed this term have typically misunderstood it.

Paget (1930) writes "Wheatstone made an important contribution to acoustic theory by his discovery of what he called Multiple Resonance i.e. the possibility of obtaining two or more resonant notes simultaneously from the same resonator."

Russell (1928) was confused by the term and struggled to understand what Wheatstone (1837) meant. He writes that it is clear that when Wheatstone used the term "multiple resonances...he is talking of the 'overtones'- - - the term multiple is synonymous with what we might designate as 'exact multiples of the fundamental.'" Later he points out seeming contradictions in Wheatstone's article. He writes of

Wheatstone that "it is evident that he thought of the mouth cavity functioning as one simple resonator. Yet he spoke constantly of 'multiple resonances' and used the plural. So there must have been some idea in his mind implying an involvement of more than one partial."

Paget and Russell probably failed to recognize Wheatstone's actual meaning because they assumed that Wheatstone understood that the voice source was composed of a complex vibration. As we have shown, however, this information was not widely recognized before Wheatstone's time (Ohm's crucial 1843 article appeared six years after Wheatstone's 1837 article), and Wheatstone (1837) himself does not appear to have recognized it either.

A further indication that Wheatstone did not know about the nature of complex vibrations is to be found in an article written in 1823. Wheatstone (1823) believed that various acoustic properties of sounds beyond frequency and amplitude were caused by the number and amplitude of individual "molecular vibrations" that "pervade the entire substance of a phonic." (Phonic was Wheatstone's term for bodies "which being properly excited, make those sensible oscillations, which have been thought to be the proximate cause of all the phenomena of sound.")

In commenting on these important "minute motions," Wheatstone (1823/1973) makes it clear that he does not mean simultaneous harmonic vibration. He criticizes Perrault for mistaking "for these vibrations the oscillations of the subdivisions of the long string that he employed." It would not be until Ohm's article in 1843 that these harmonic "oscillations" would be widely recognized as key to understanding the nature of complex sounds, and their perception.

THE MEANING OF "MULTIPLE RESONANCE"

A close examination of Wheatstone's (1837) article shows what he intended by "multiple resonances" and reveals that he did not originate the "harmonic" theory as it is understood today. It should be recognized from the above that the requisite scientific knowledge simply was not yet in existence, or at least was not widely known.

Wheatstone uses the term "multiple resonance" contrastively with the term "simple or unisonant resonance." He explains that simple resonance refers to the fact that when a vibrating body is brought near a column of air whose natural frequency coincides with that of the vibrating body,

the column of air begins to sympathetically vibrate at the same frequency.

The following is Wheatstone"s description of the "new facts of resonance" which amounts to his definition of "multiple resonance": "A column of air will not only enter into vibration, when it is capable of producing the same sound as the vibrating body which causes the resonance, but also, when the number of vibrations which it is capable of making is any simple multiple of that of the original body, or in other words, if the sound to which the tube is fitted is any harmonic of the original sound."

The key point to notice in this passage is that Wheatstone believes that the source is composed of a single simple tone, and it is the column of air which vibrates at a multiple of the source frequency. Thus, although Wheatstone was the first to recognize that the tones important in voiced vowel acoustics must be at frequencies that are multiples of the fundamental frequency, he did not recognize that these tones are actually present in the original voice source.

Recognition of Wheatstone's theoretical perspective provides a possible explanation for the observation made by Russell (1928), alluded to above: "It is particularly interesting to observe that in Wheatstone's mind, he

apparently saw no conflict between Willis's theory and that he proceeds to expound. We might even go farther and say that to all appearances Wheatstone feels he is merely clarifying and elaborating the theory as Willis proposed it. Yet generally speaking, we have talked of the two as being in conflict with each other. And regardless of how it appeared to Wheatstone, it would appear that they are."

Wheatstone's seeming obliviousness to the differences between his own and Willis' theory is more understandable once Wheatstone's theory is properly delineated. Clearly, there are important differences between them, particularly whether the frequency of the "cavity tone" must be a multiple of the fundamental frequency. However, the two theories are similar in one very significant way. Both theories suppose that a source containing a single frequency gives rise to an additional tone as a result of proximity to a column of air of a certain length.

As a consequence of this explanation of Wheatstone's theory, it becomes apparent that Wheatstone cannot be given credit for the source-filter theory of vowel production, because he never conceived of the vocal tract as a filter for a complex source. He was the first to recognize that the "cavity tone" must be a multiple of the fundamental, a

significant advance, but he did not correctly recognize why this is so.

WHO CREATED THE SOURCE-FILTER THEORY?

If Wheatstone is not the originator of the source-filter theory, it must be established who the true originator is. Helmholtz (1877/1954) is an obvious candidate. He writes "Vowels of speech are in reality tones produced by membranous tongues (the vocal chords), with a resonant chamber (the mouth) capable of altering in length, width, and pitch of resonance, and hence capable also of reinforcement at different times different partials of the compound tone to which it is applied." This is an elegant and accurate description of the source-filter theory.

However, historically, Wheatstone has been credited as the first, Grassman as the second, and Helmholtz as the third to describe the source-filter theory of vowel acoustics. Having excluded Wheatsone from consideration, the crucial issue becomes whether or not Grassman's description of the science of vowel acoustics was an advance over Wheatstone's. It becomes necessary to examine Grassman's (1854/1904) original article to determine if he was aware that the

perceived "cavity tone" is composed of harmonics present in the original source.

Because of the wide-spread belief that Wheatstone had previously delineated the source-filter theory at the time of Grassman's (1854/1904) article, Grassman is typically credited historically as having played a minor role in the development of this theory. Miller (1937) writes that Willis' theory "was extended by Wheatstone (1837) and Grassmann (1854)...Helmholtz (1862-1877) expounded the theory, a development of those given before..." Similarly, Scripture (1904) writes "Wheatstone's view was expounded as a general hypothesis by Grassmann and
developed into a theory by Helmholtz."

Chiba and Kajiyama (1941) just list Grassmannn's (1854) article as one that "discussed theories of the vowels." They refer to the harmonic resonance theory as the Helmholtz-Wheatstone theory; apparently Grassmann's contribution is considered to be a minor one. Incidentally, the only articles which provide a reference for Grassman's (1854/1904) article, give the reference as "Programme des Stettiner Gymnasiums, Leitfadens der Akustik." This title is actually the title that Grassman himself provides in a later article (1870/1904). The actual title under which the original

article was published was "Uebersicht der Akustik und der niedern Optik. von Professor Hermann Grassman: Programm des Koniglichen und Stadtgymnasiums zu Stettin, September 1854." The obvious implication is that none of those who assign various degrees of credit to Grassman and Helmholtz actually read Grassman's (1854/1904) original article.

In fact, Grassmann (1854) clearly understood all of the critical elements of the source-filter theory. Unlike Wheatstone (1823), Grassman recognized that a vibrating body can vibrate in a complex fashion, containing components including a fundamental frequency and a series of harmonics; he also recognized that several of these components can be audible. He writes (Grassman 1854/1904, my translation): "Yes, it is possible for a string to move so that it vibrates as a whole, and, nevertheless, simultaneously divide itself into a quantity of equal divisions which also vibrate independently, so that besides the primary tone that the string produces when it vibrates as a whole, additional intrinsic higher tones that are harmonic to the primary tone can be simultaneously sounded."

Although his description of vowel acoustics is not very concise, it is clear that his basic explanation is correct. He writes (Grassman 1854/1904, my translation; slashes [/]

indicating phonemic designations are added, and not found in the original text): "Simultaneously, the vocal cords set the air contained in the oral cavity into vibration, through which various quiet secondary tones are emitted, which follow the position one gives the oral cavity, and in which the series of harmonic tones has the tone of the vocal cords as the basic tone in this way, the vowels are produced. In the transition from /u/ to /ü/ to /i/, an attentive ear easily hears a series of quiet harmonic secondary tones that go from the two stroked c to the five stroked c and higher; one can produce these through the same oral position. For the vowel /a/ an entire series of harmonic secondary tones are sounded of which the ear can usually perceive until the fourth octave from the fundamental tone, so that for an /a/ a full chord of secondary tones is sounded together. In this manner, the transition from /a/ through /o/ to /u/, as well as that from /a/ through /e/ to /i/ or through /ö/ to /ü/, are simultaneously explained."

It is clear from this passage, that Grassman was cognizant of the fact that vowels are characterized by harmonics made audible by a particular vocal tract configuration. Indeed his comments regarding the vowel /a/ make it clear that he was aware that the bandwidth of the

most prominent resonance in the vocal tract could allow it to cause several harmonics to resonate within a particular frequency band.

It does not appear that Grassmann was even aware of Wheatstone's work in this area. In any event, his treatment goes beyond Wheatstone's in its recognition that the source contains numerous harmonics and that the vocal tract resonates the harmonics associated with particular vowels.

Although some current scholarship downplays the view of Grassman as a frustrated schoolteacher, who never received recognition for his mathematical and scientific accomplishments (Schubring, 1996; Rowe, 1996), the history of the development of the source-filter theory of vowel acoustics seems to buttress this view. Grassman (1870/1904) himself seems clearly angered by Helmholtz receiving credit for the theory which he had originated. After quoting his earlier article and crediting it with providing the first correct description of vowel acoustics, he writes (1870/1904, my translation): "This passage in my program...although it is a complete theory of vowel tones...remained completely disregarded...Five years later Helmholtz drafted..." He proceeds to argue that Helmholtz's theory is "incomplete" and "defective."

In any event, whether the cause was Helholtz's prominence, his "customary thoroughness" (Jenkins, 1987) or luck, the source-filter theory of vowel acoustics rose to prominence following his writings. Unlike Grassman, Helmholtz was a scientist who was rarely ignored.

RECEPTION OF THE SOURCE FILTER THEORY

It should not be thought that as soon as the views of Wheatstone, Grassman and Helmholtz had been publicized that their approach was accepted, and Willis' inharmonic theory rejected. For approximately one hundred years following their work there were still important scientists who considered Willis' position to have merit.

Thus, as previously cited, Rayleigh (1896) writes that "both ways of regarding the subject are legitimate, and not inconsistent with one another." He suggests that the harmonic theory is more appropriate for vowels, with low frequency peaks of resonance, while the inharmonic theory is more appropriate for vowels with high frequency peaks of resonance. Lloyd (1898) gives a complex explanation defending the inharmonic theory, despite the phonographic evidence that he had, which showed no evidence of the existence of any inharmonic energy. He does not take a conclusive position, preferring a "middle position."

Scripture (1904) credits Hermann with supporting "Willis in asserting that the cavity tone is completely independent of the cord tone." He also writes that "Hermann has objected to the overtone theory of the cavity tone that in many voices it is so high above the cord tone that it

cannot be supposed that an overtone of that pitch can possibly be present."

Scripture was so confident that the view of Willis and Hermann was correct, that he writes "in the face of such conclusive evidence it is hard to see any point in which a decision in favor of the theory proposed by Willis and developed by Hermann can possibly be attacked" (Scripture, 1904). Later he writes "it has been shown that the mouth tone is inharmonic to the cord tone and that it is a free vibration. It follows that the cord vibrations are not of the nature of the sum of a series of vibrations." He is willing to concede only that "some cases...in my study...may be reconciled with the Helmholtz view."

The view held by Hermann and Scripture (1904) should not be understood as stemming from a lack of knowledge of the possibility of complex periodic vibration. It stemmed, rather, from the belief that this type of vibration is not characteristic of vocal fold movement. Scripture (1904) writes "According to one theory the note produced directly by the vibrations of bands consists of a series of partial tones...The second theory would suppose the bands not to vibrate but to open momentarily and then close again in a series of movements, whereby the resultant air movement is

not a vibratory one of a sinusoidal or a harmonic nature but is a series of brief puffs...That for the chest register this theory is certainly the correct one has been shown..."

Chiba and Kajiyama (1941), although completely cognizant of the fact that voiced vowels are composed of a fundamental and its harmonic frequencies, with areas of resonance causing individual harmonics to be reinforced, and possessing a fair understanding of the essential nature of the two theories, do not unequivocally state that the harmonic theory is correct and the inharmonic theory is incorrect. It seems that they object to the harmonic theory on the grounds that vowels can be produced by whisper or by synthetic inharmonic tones (p.175). This objection is trivial, in that the theory merely describes the acoustic nature of naturally produced voiced vowels. Nevertheless, Chiba and Kajiyama do recognize that "it is scarcely necessary to say that the 'steady-state' theory is, if anything to be preferred to the transient theory." ("steady-state" and "transient" theories are considered by Chiba and Kajiyama to be other terms for "harmonic" and "inharmonic" theories respectively.)

Following the work of Chiba and Kajiyama (1941), the source-filter conception of vowel acoustics has never been

questioned. It has become the standard description of vowel acoustics.

ONE FORMANT OR TWO (OR THREE)?

The issue of how many cavity tones, or formants are present in vowels, was one that was studied and debated almost from the very beginning of the field of vowel acoustics. As this issue was crucial in the development of an understanding of vowel perception, particularly as it relates to vowel normalization, some of the basic history will be reviewed.

It should be obvious, based on the acoustic theories of Willis (1830) and Wheatsone (1837) reviewed above, that the earliest scientific hypotheses for the basis of cavity tones or formants had no room for the existence of more than one formant. In both articles, the cavity tone was seen as arising from the resonating cavity, based on its length. Since there was no recognition of the existence of numerous harmonics in the source, it was unlikely that anyone would hypothesize the existence of multiple formants when approaching the issue from a theory-driven perspective.

Many investigators, however, studied the issue empirically, an approach which made the discovery of

additional peaks of resonance possible. As mentioned above Helmholtz (1877/1954) cites a number of researchers who attempted to determine the frequency of the "cavity tones" of vowels. In fact, judging by the table of values provided by Ellis in a footnote to Helmholtz (1877/1954), it would appear that Donders may have detected the presence of a second formant for one vowel. Helmholtz himself divides the vowels into those in which he could detect two resonant peaks, and those in which he could not. These are front and back vowels respectively.

The first researcher to argue that all vowels were characterized by two peaks of resonance was A. G. Bell (1879). Lloyd came to a similar conclusion shortly thereafter (1890, 1898), and it is possible that he came to this conclusion independent of Bell's work. Lloyd also describes additional resonances and attempts to explain in which part of the vocal tract they arise.

Miller (1937) used an instrument called the phonodeik to analyze vowels and concludes that Helmholtz was correct in asserting that some vowels are characterized by one peak of resonance, and others by two.

Paget (1930) gives an extended description of his own discovery of the existence of two peaks of resonance. He

also gives a detailed description of the production of synthetic vowels using two resonators.

It seems likely that part of the reason that Bell (1879) and Paget (1930) recognized this fact while Helmholtz did not, relates to their differing research techniques. Helmholtz found his resonances by holding his oral cavity in a position appropriate for a given vowel, and holding tuning forks of different pitches next to his open mouth (1877/1854). Helmholtz's method made it likely for a back resonance to be missed when it was close in frequency to that of the front resonance.

Bell's method was to tap a finger placed against the neck above the larynx to detect the lower frequency resonance and to tap a finger placed just in front of the upper teeth to detect the higher frequency resonance mouth (1911, quoted by Jenkins, 1987; Ellis in a footnote to Helmholtz, 1877/1954). The a priori assumption is clearly that the back cavity is primarily related to a lower resonance and the oral cavity to an upper resonance. Paget (1930) used a similar percussion method. Not surprisingly, Bell (1911, quoted by Jenkins, 1987; Ellis in a footnote to Helmholtz, 1877/1954), and Paget, (1930) believed that they could detect two resonances for each vowel.

Interestingly, Fletcher (1929), writing around the same time as Paget (1930), states as a simple fact the same concept that Paget (1930) agonizes over for several pages. Fletcher (1929) writes that vowels "pass through two variable resonating cavities, namely the mouth and the throat. For this reason, all voiced sounds are characterized by having component frequencies magnified in two particular regions."

Lloyd (1898) provides an excellent explanation for why phonographic evidence did not show the presence of two peaks of resonance despite the theoretical model that requires their presence. He writes "in a vowel having two resonances differing only about 300 v.d. it is useless to look for any sign of doubleness in the reinforcements evidenced by the Fourierian analysis, unless the vowel is sung below 150 v.d." ["v.d." is apparently equivalent to Hertz.] He claims that whenever vowels are recorded at frequencies that are sufficiently low, one should find "that palpable cleavage in the reinforcements which is the constant sign of the presence of two separate resonances."

Lloyd's (1898) position is quite modern, as Chiba and Kajiyama (1941) take an almost identical position. They make reference to "some vowels - one of whose two formants does not often manifest itself or whose two formants are near each

other (i.e. vowels which were formerly called single formant vowels)..."

Although the focus of much of the early research reviewed here was vowel production, the tacit assumption throughout this period of time was that the cavity tones were the primary cues of vowel identity. Thus, in a sense, this historical overview brings the field of vowel perception to the point of the "simple target model" (Strange, 1989). In this model "vowel targets are represented as points in a multidimensional acoustic space whose coordinates are the first two or three oral formants."

APPENDIX B

REMOVAL OF "MISARTICULATED VOWELS FROM ANALYSIS

As mentioned on page 45, after the data were collected, it became clear that, even under the blocked, unfiltered condition, there were certain vowels that were consistently misidentified. The following protocol was used to decide which of these vowels would be removed from the reporting and analysis of the data.

All stimuli that were perceived in the blocked unfiltered conditions by a majority of the eight subjects as a vowel other than the intended vowel, were reexamined by the experimenter. As many of these were perceived as imperfect exemplars of the intended vowels, they were presented to a trained listener in a forced-choice listening task. The choices were the intended vowel and the majority "incorrect" response of the eight subjects. The listener was told the speaker type (adult male, adult female, or child), and was allowed to listen to each vowel stimulus as many times as he wished. All vowels that the trained listener perceived as other than the intended vowel were assumed to have been mispronounced by the speaker, and unsuited for inclusion in an analysis of perceptual errors. As a result of this procedure, three phonated and 15 whispered vowels (from a

total of 135 phonated and 135 whispered vowels) were removed from the analysis.

Table B.a shows the means of correct identification scores for all eight conditions for the original set of all vowels and the means for the set with the vowels judged to be "mispronounced" removed. All vowels (other than /I/ and /u/) and all speaker types had at least one stimulus removed as a result of this procedure. The effect of removal of these vowels is to improve both the phonated and whispered vowel scores; the whispered vowels showed a greater increase. This is to be expected, as more "incorrectly perceived" vowels were removed from the whispered set than from the phonated set. However, no main effect or interaction effect was changed from significant to insignificant or from insignificant to significant as a result of the removal. Table B.b shows the ANOVA for the original data set.

Table B.a. Mean percent correct identification for all conditions with and without "misarticulated" vowels removed

	Blocked-speaker				Mixed-speaker			
	Phonated		Whispered		Phonated		Whispered	
VOWEL SET	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
All data	86.6	86.5	74.9	70.2	86.6	85.1	70.4	62.6
Misarticulated removed	86.8	87.2	79.7	74.2	87.4	85.2	74.5	66.1

Table B.b. Three way ANOVA for presentation condition x phonatory type x availability of upper formants for original data. All effects.

Condition = Blocked-Speaker/Mixed-Speaker
Phonation = Phonated/Whispered
Filtering = Unfiltered/Filtered

Source of Variance	Estimated Mean Square	Degrees of Freedom	Error Term	Estimated Mean Square	Degrees of Freedom	F Ratio	p level	
Condition	178.55	1	CxSubject	17.65	7	10.12	0.015	*
Phonation	4440.62	1	PxS	18.87	7	235.31	0.0000012	*
Filtering	202.33	1	FxS	17.07	7	11.85	0.011	*
CxP	116.12	1	CxPxS	23.54	7	4.93	0.062	
CxF	20.88	1	CxFxS	13.33	7	1.57	0.251	
PxF	119.86	1	PxFxS	4.63	7	25.90	0.001	*
CxPxF	2.68	1	CxPxFxS	11.01	7	0.24	0.637	

As previously indicated, the data analysis was not significantly affected by the removal of the misarticulated stimuli from the analysis. Eliminating these data points allowed the analysis to focus on the "correctly" pronounced vowels. Because elimination of these vowels provides the most meaningful analysis, all statistics and figures were based on this data set. (See the Discussion section for other comments on this topic.)

APPENDIX C

Table C.a.Percent correct identification for all subjects and conditions

	BLOCKED-SPEAKER				MIXED-SPEAKER			
	PHONATED		WHISPERED		PHONATED		WHISPERED	
Subject	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
1	96.21	95.45	93.33	79.17	90.15	90.15	81.67	76.67
2	91.67	90.15	81.67	74.17	90.91	89.39	81.67	65.83
3	93.94	94.70	84.17	78.33	92.42	90.91	74.17	65.00
4	89.39	91.67	84.17	82.50	87.12	84.09	74.17	74.17
5	81.82	74.24	75.00	74.17	90.91	85.61	70.83	52.50
6	74.24	82.58	70.83	65.83	79.55	81.82	66.67	68.33
7	89.39	85.61	77.50	73.33	88.64	91.67	76.67	68.33
8	78.03	83.33	70.83	65.83	79.55	68.18	70.00	58.33
AVG	86.84	87.22	79.69	74.17	87.41	85.23	74.48	66.15

APPENDIX D

Although the focus of this dissertation is on perception, a sample of acoustic data is reported for comparison purposes. The measurements for all formant frequencies reported in this appendix were made using the SPECTO program, written by Mark Weiss.

Averaged amplitude spectra through 8000 Hz were made for each of the vowels and the center frequencies containing peak amplitudes were estimated by eye to provide an estimate of formant frequencies. The typically high amplitudes of the lowest harmonics, particularly for vowels with low first formants sometimes made the judgment of the center frequency of the first formant more difficult than for the second and third formants. Fundamental frequency was estimated for the phonated vowels by estimating the frequency of the twentieth harmonic from the averaged spectra and dividing by 20.

The fundamental frequency and formant frequencies for all of the phonated vowels of one adult male, one adult female and one child speaker are presented in table D.a. Comparison data, adapted from Peterson and Barney (1952) is presented in table D.b.

Table D.a. F1, F2 and F3 for one adult male, one adult female and one child speaker for the phonated vowels (All values in Hz).

		/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/
Formant frequencies (cps)										
F1										
Phonated	Male	290	440	590	620	710	500	590	440	330
	Female	390	440	740	980	1000	590	830	470	380
	Child	360	680	910	1010	860	650	920	620	530
F2										
Phonated	Male	2220	1740	1690	1630	1200	1070	1160	1100	890
	Female	2810	2000	1940	1780	1440	1010	1330	1370	1030
	Child	3260	2400	2300	2400	1480	1100	1480	1540	1700
F3										
Phonated	Male	2680	2540	2590	2430	2300	2540	2250	2490	2400
	Female	3350	2930	2970	2760	2650	2790	2340	2880	2990
	Child	3850	3500	3540	3700	3110	3200	3200	3120	3000

Table D.b. F1, F2 and F3 for the adult male, adult female and child speakers for the phonated vowels, reported by Peterson and Barney (1950) (All values in Hz).

		/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/
Formant frequencies (cps)										
F1										
Phonated	Males	270	390	530	660	730	570	640	440	300
	Females	310	430	610	860	850	590	760	470	370
	Children	370	530	690	1010	1030	680	850	560	430
F2										
Phonated	Males	2290	1990	1840	1720	1090	840	1190	1020	870
	Females	2790	2480	2330	2050	1220	920	1400	1160	950
	Children	3200	2730	2610	2320	1370	1060	1590	1410	1170
F3										
Phonated	Males	3010	2550	2480	2410	2440	2410	2390	2240	2240
	Females	3310	3070	2990	2850	2810	2710	2780	2680	2670
	Children	3730	3600	3570	3320	3170	3180	3360	3310	3260

Although no statistical analysis was performed on the acoustic data, the formant frequencies for the whispered vowels appeared to have a tendency to be higher than the formant frequencies for the phonated vowels. Table D.c. illustrates this tendency for the same three speakers whose phonated formant frequencies were provided in table D.a.

Table D.c. F1, F2 and F3 for one adult male, one adult female and one child speaker from this study for the phonated and whispered vowels (All values in Hz).

		/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʌ/	/ʊ/	/u/
Formant frequencies (cps)										
F1										
Phonated	Male	290	440	590	620	710	500	590	440	330
	Female	390	440	740	980	1000	590	830	470	380
	Child	360	680	910	1010	860	650	920	620	530
Whispered	Male	350	740	770	710	800	620	840	580	470
	Female	390	780	890	1140	1020	820	920	880	350
	Child	470	740	890	970	1000	740	1000	920	470
F2										
Phonated	Male	2220	1740	1690	1630	1200	1070	1160	1100	890
	Female	2810	2000	1940	1780	1440	1010	1330	1370	1030
	Child	3260	2400	2300	2400	1480	1100	1480	1540	1300
Whispered	Male	2260	1870	1810	1780	1360	1130	1420	1210	1000
	Female	2930	2190	2200	2010	1420	1210	1450	1360	1050
	Child	3290	2580	2520	2280	1510	1170	1630	1480	940
	Child									
F3										
Phonated	Male	2680	2540	2590	2430	2300	2540	2250	2490	2400
	Female	3350	2930	2970	2760	2650	2790	2340	2880	2990
	Child	3850	3500	3540	3700	3110	3200	3200	3120	2800
Whispered	Male	2750	2720	2680	2610	2540	2610	2400	2580	2520
	Female	3530	3110	3170	3080	2950	2810	2730	2760	2840
	Child	3850	3440	3350	3400	2960	3170	3170	3100	2800

This finding of a tendency toward higher formant frequency values for whispered vowels compared to phonated vowels is also consistent with the results reported by Kallail and Emanuel (1994) and Eklund and Traunmüller (1997).

REFERENCES

Ainsworth, W.A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant and M. Tatham (eds.) *Auditory Analysis and Perception of Speech*. London, Academic Press. pp. 103-113.

Assman, P.F. and Nearey, T.M. (1987). Perception of front vowels; The role of harmonics in the first formant region. *Journal of the Acoustical Society of America 81(2)*, 520-534.

Assman, P.F., Nearey, TM. & Hogan, J.T. (1982). Vowel Identification: Orthographic, perceptual, and acoustic aspects. *Journal Acoustical Soc. America 71(4)*, 975-989.

Bell, A.G. (1879). Vowel Theories. *American Journal of Otology 1*, 163-180.

Bell, A.G. (1911). *The Mechanism of Speech (5th edition)*. New York, Funk and Wagnalls.

Bernoulli, D. (1755). *Reflexions et eclaircissements sur les nounelles vibrations des cordes exposees dansles memoires de l'Academie, de 1747 et 1748*. Berlin, Royal Academy.

Bladon, A., Hendon, C. and Pickering J.B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication 4 (1)*, 459-69.

Carlson, R., Fant, G. and Grandstrom, B. (1975). Two-formant models, pitch and vowel perception. In *Auditory Analysis and Perception of Speech*, edited by G. Fant and M.A.A. Tatham (London, Academic Press) pp. 55-82.

Chiba,T. and Kajiyama,M. (1941). *The Vowel: Its Nature and Structure* (Tokyo, Tokyo-Kaiseikan).

Chistovich, L.A. (1985). Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America 77(3)*, 789-805.

Chistovich, L.A. and Lubinskaja, V.V. (1979). The center of gravity effect in vowel spectra and critical distance between the formants. *Hearing Research 1*, 185-195.

Chladni, E.F.F. (1787/1973). Discoveries in the Theory of Sound (Selections, translated by R.B Lindsay). In *Acoustics: Historical and Philosophical Development*, edited by R.B. Lindsay (Stroudsburg, PA, Dowden, Hutchinson and Ross), pp. 156-165.

Delattre, P.C., Liberman, A.M. and Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America 27(4)*, 769-773.

Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America 67(1)*, 253-261.

Eklund, I. and Traunmuller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, 54 (1), 1-21.

Euler, L. (1727/1973). Dissertation on sound (Translated by R.B. Lindsay). In *Acoustics: Historical and Philosophical Development*, edited by R.B. Lindsay (Stroudsburg, PA, Dowden, Hutchinson and Ross), pp. 104-117.

Euler, L. (17--/1802). *Letters on Different Subjects in Physics and Philosophy Addressed to a German Princess (vol. 2, 2nd edition.* Translated by H. Hunter). (London, Murray and Highly).

Euler,L. (1753). *Remarques sur les memoires precedents de M. Bernoulli*. (Berlin, Royal Academy), pp. 1-196.

Euler,L. (1759/1973). Letter to Joseph Louis Lagrange, from Berlin (Translated by R.B. Lindsay). In *Acoustics: Historical and Philosophical Development*, edited by R.B. Lindsay (Stroudsburg, PA, Dowden, Hutchinson and Ross), pp. 131-135.

Fahey, R.P., Diehl, R.L. and Transmüller, H. (1996). Perception of back vowels: Effects of varying F1-F0 Bark distance. *Journal of the Acoustical Society of America 99(4)*, 2350-2357.

Fant, G. (1975). Nonuniform vowel normalization. *Speech Transmission Lab., Royal Institute of Technology Quarterly Progress Status Report 2/3*, 1-19.

Fletcher,H. (1929). *Speech and Hearing* (Van Nostrand, New York).

Fujisaki, H & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Audio Electroacoustics. AU-16 1,* 73-77.

Gerstman, L.J. (1968). Classification of self-normalized vowels. *IEEE Transactions Audio & Electroacoustics, vol. AU-16, no. 1,* 78-80.

Grassman, H. (1854/1904). Uebersicht der akustik und der niedern optik. In *Hermann Grassmans Gesammelte Mathematische und Physikalische Werke,* edited by E. Study, G. Scheffers and F. Engel (Leipzig, B.G. Teubner).

Grassman, H. (1870/1904). Ueber die Physikalische Natur der Sprachlaute. In *Hermann Grassmans Gesammelte Mathematische und Physikalische Werke,* edited by E. Study, G. Scheffers and F. Engel (Leipzig, B.G. Teubner).

Helmholtz, H.L.F. (1877/1954). *On the Sensations of Tone (Translated and with an additional appendix by A.J. Ellis)* (New York, Dover).

Hillenbrand, J. & Gayvert, R.T. (1993). Vowel classification based on fundamental frequency and formant frequencies. *Journal of Speech and Hearing Research 36,* 694-700.

Hirahara, T. & Kato, H. (1992). The effect of FO on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (eds.) *Speech Perception, Production and Linguistic Structure* (Burke, VA, IOS Press), pp. 88-111.

Hoemeke, K.A. and Diehl, R.L. (1994). Perception of vowel height: The role of Fl-FO distance. *Journal of the Acoustical Society of America 96,* 661-674.

Honda, Kiyoshi (1997). Form and function: A view of speech production. Paper presented at the Conference for Professor Katherine S. Harris, CUNY Graduate School and University Center.

Javkin, H.R., Hermansky, H. and Wakita, H. (1987). Interaction between formant and harmonic peaks in vowel perception. *Proceedings of the Eleventh International Congress of Phonetic Sciences vol. 5,* 186-189.

Jenkins, J.J. (1987). A selective history of issues in vowel perception. *Journal of Memory and Language 26*, 542-549.

Jenkins, J.J. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America 85(5)*, 2135-2154.

Jenkins, J.J., Strange, W. and Edman, T.R. (1983). Identification of vowels in 'vowelless' syllables. *Perception and Psychophysics 34*, 441-450.

Joos, M.A. (1948). Acoustic phonetics. *Language 24 (Suppl.)*, 1-136.

Kahn, D. (1978). On the identifiability of isolated vowels. *UCLA Working Papers in Phonetics, 41*, 26-31.

Kallail, K.J. and Emanuel, F.W. (1984). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *Journal of Phonetics 12*, 175-186.

Kempelen, (1791) *Le Mecanisme se la Parole suivi de la Description d'une Machine Parlante*. Par M. de Kempelen, Conseiller Aulique Actuel de sa Majeste l'Empereur Roi. (Vienna).

Koenig, W. (1949). A new frequency scale for acoustic measurements. *Bell Labs Records 27*, 299-301.

Kratzenstein, C.G. (1780). Experiment: Solution proposed at a public meeting for the year 1870 at the St. Petersburg Imperial Academy of Sciences concerning the following assignments. *Akademiya Nauk, Acta* (St. Petersburg) 188-252.

Ladefoged, P. & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America 29(1)*, 98-104.

Lehiste, I. and Peterson, G.E. (1959). The identification of filtered vowels. *Phonetica 4*, 161-177.

Lindsay, R.B. (1966). The story of acoustics. *Journal of the Acoustical Society of America, 39*, 629-644.

Lloyd, R.J. (1890). *Some Researches into the Nature of Vowel-Sound.* (Liverpool, Turner and Dunnet).

Lloyd, R.J (1896). The genesis of vowels. *Transactions of the Royal Society of Edinburgh, 1896,* 972-973.

Lloyd, R.J. (1898). On the Fourierian analysis of phonographic tracings of vowels. *Proceedings of the Royal Society of Edinburgh, 1898,* 97-117.

Lobanov, B.M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America 49,* 606 -608.

Macchi, M.J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America 68(6),* 1636-1642.

McKendrick, J.G. (1898). Observations on the theories of vowel sounds. *Proceedings of the Royal Society of Edinburgh, 1898,* 71-96.

Miller, D.C. (1937). *The Science of Musical Sounds* (MacMillan, New York).

Miller, J.D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America 85(5),* 2114-2133.

Minifie, F.D. (1973). Speech acoustics. In *Normal Aspects of Speech, Hearing, and Language,* edited by F.D. Minifie, T.J. Hixon, and F. Williams (Prentice-Hall, Englewood Cliffs, N.J.), pp.235-284.

Mullennix, J.W., Pisoni, D.B. and Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America 85,* 365-378.

Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America 85(5),* 2088-2113.

Nusbaum, H.C. & Morin, T.M. (1992). Paying attention to differences among talkers. In *Speech Perception, Production and Linguistic Structure,* edited by Y. Tohkura, E.

Vatikiotis-Bateson & Y. Sagisaka (Burke. VA, IOS Press), 113-123.

Ohm, G.S. (1843/1973). On the definition of a tone with the associated theory of the siren and similar sound producing devices (Translated by R.B. Lindsay). In *Acoustics: Historical and Philosophical Development*, edited by R.B. Lindsay (Dowden, Hutchinson and Ross, Stroudsburg, PA) 243-247.

Paget, R. (1930). *Human Speech* (New York, Harcourt Brace).

Papçun, G. (1980). How do different speakers say the same vowels?: Discriminant analysis of four imitation dialects. *UCLA Working Papers in Phonetics 48*, 8-13.

Peterson, G.E. (1954). Systematic research in experimental phonetics: 4, The evaluation of speech signals. *Journal of Speech and Hearing Disorders, 19*, 158-168.

Peterson, G.E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research 4 (1)*, 10-29.

Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America 24*, 175-184.

Plomp, R. (1964) The ear as a frequency analyzer. *Journal of the Acoustical Society of America, 36*, 629-644.

Potter, R., Kopp, G. and Green, H. (1947). *Visible speech* (New York, Van Nostrand-Reinhold).

Potter and Steinberg (1950). Toward the specification of speech. *Journal of the Acoustical Society of America 22*, 807-820.

Rayleigh, J.W.S (1896). *The Theory of Sound vol. 2* (MacMillan, New York).

Rowe, D. (1996). The reception of Grassman's work in Germany during the 1870s. In *Hermann Gunther Grassmann (1809-1877): Visionary, Mathemetician, Scientist and Neohumanist*, edited by G. Schubring (Kluwer Academic, Boston).

Russell,G.O. (1928). *The Vowel: Its Physiological Mechanism as Shown by X-ray* (McGrath, College Park, MD).

Schubring, G. (1996). Introduction: Reflections on the complex history of Grassman's reception. In *Hermann Gunther Grassmann (1809-1877): Visionary, Mathemetician, Scientist and Neohumanist*, edited by G. Schubring (Kluwer Academic, Boston).

Scripture,E.W. (1904). *Elements of Experimental Phonetics* (Scribners, New York).

Shepard, R.N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review* 89 (4), 305-333.

Slawson, A.W. (1968). Vowel quality and musical timbre as functions of spectrum envelopes and fundamental frequency. *Journal of the Acoustical Society of America 43*, 87-101.

Stevens, 5.5., Volkman, J. (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology 53*, 329-353.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America 85 (5)*, 2081-2087.

Strange, W. and Gottfried, T.J. (1980). Task variables in the study of vowel perception. *Journal of the Acoustical Society of America 68*, 1622-1625.

Strange, W. Jenkins, J.J. and Johnson, T.L. (1983) Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America 74*, 695-705.

Strange, W., Verbrugge, R.R., Shankweiler, D.P. & Edman, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America 60(1)*, 213-224.

Suomi,K. (1984). On talker and phoneme information conveyed by vowels: a whole spectrum approach to the normalization problem. *Speech Communication 3*, 199-209.

Syrdal, A.K. (1985) Aspects of a model of the auditory representation of America English vowels. *Speech Communication 4*, 121-135.

Syrdal, A.K. and Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of America English vowels. *Journal of the Acoustical Society of America 79*, 1086-1100.

Tiffany, W.R. (1952). Vowel recognition as a function of duration, frequency modulation and phonetic context. *Journal of Speech and Hearing Disorders 17*, 289-301.

Traunmüller, H. (1990). A note on hidden factors in vowel perception experiments. *Journal of the Acoustical Society of America 88(4)*, 2015-2018.

Traunmüller,H. (1981) Perceptual dimension of openness in *vowels. Journal of the Acoustical Society of America 69(5)*, 1465-1475.

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-25, 2*, 183-192.

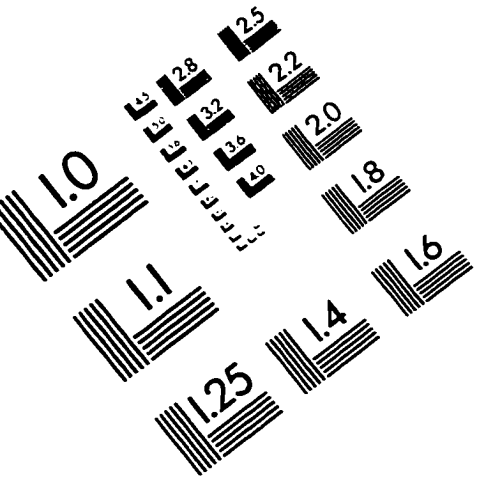
Wheatstone, C. (1823/1973). New experiments on sound, in *Acoustics: Historical and Philosophical Development*, edited by R.B. Lindsay (Dowden, Hutchinson and Ross, Stroudsburg, PA), 184-193.

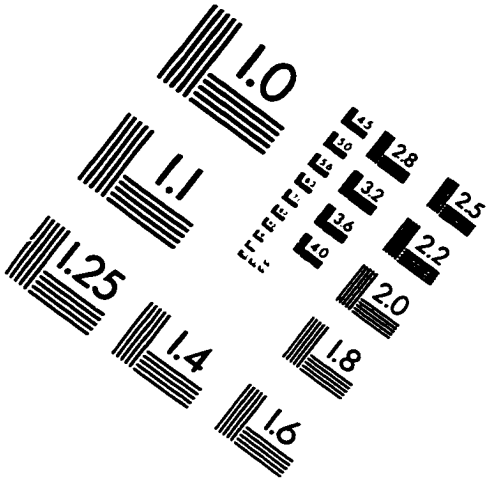
Wheatstone, C. (1837). Willis on reed organ pipes, speaking machines, etc. *London and Westminster Review, Oct. 1837*, 27-41.

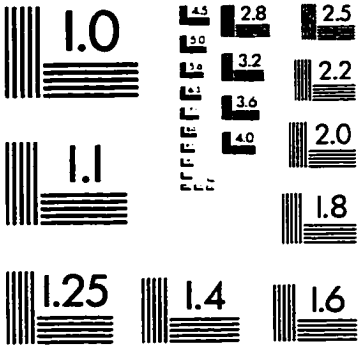
Willis, R. (1830). On the vowel sounds and on reed organ pipes. *Transactions of the Cambridge Philosophical Society iii*, 231-268.

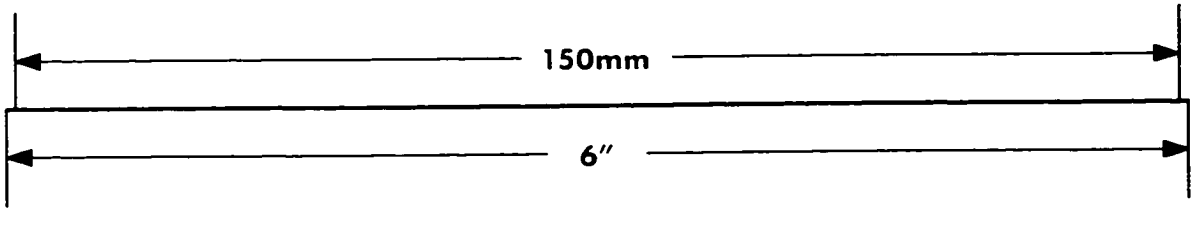
Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America 33(2)*, 248.

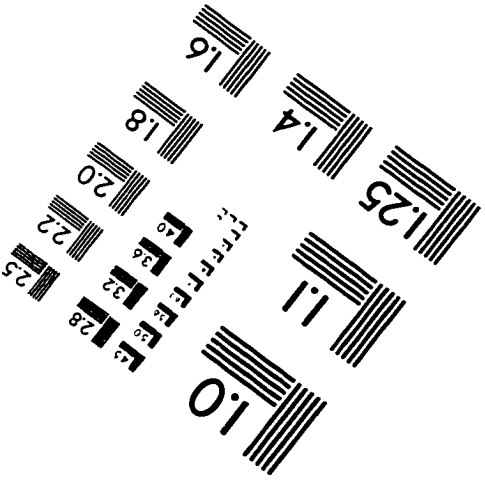
IMAGE EVALUATION
TEST TARGET (QA-3)











APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989