Tactual Display of Consonant Voicing to Supplement Lipreading

by

Hanfeng Yuan

S.B., Shanghai Jiao Tong University (1995)
S.M. Massachusetts Institute of Technology (1999)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2003

Signature of Author_____
Department of Electrical Engineering and Computer Science
August 28, 2003

Certified by_____
Nathaniel I. Durlach
Senior Research Scientist
Thesis Supervisor

Certified by____
Charlotte M. Reed
Senior Research Scientist
Thesis Supervisor

Accepted by_
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Tactual Display of Consonant Voicing to Supplement Lipreading

By

Hanfeng Yuan

Submitted to the Department of Electrical Engineering and Computer Science
on August 28, 2003, in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

ABSTRACT

This research is concerned with the development of tactual displays to supplement the information available through lipreading. Because voicing carries a high informational load in speech and is not well transmitted through lipreading, the efforts are focused on providing tactual displays of voicing to supplement the information available on the lips of the talker. This research includes exploration of 1) signal-processing schemes to extract information about voicing from the acoustic speech signal, 2) methods of displaying this information through a multi-finger tactual display, and 3) perceptual evaluations of voicing reception through the tactual display alone (T), lipreading alone (L), and the combined condition (L+T).

Signal processing for the extraction of voicing information used amplitude-envelope signals derived from filtered bands of speech (i.e., envelopes derived from a lowpass-filtered band at 350 Hz and from a highpass-filtered band at 3000 Hz). Acoustic measurements made on the envelope signals of a set of 16 initial consonants represented through multiple tokens of C_1VC_2 syllables indicate that the onset-timing difference between the low- and high-frequency envelopes (EOA: envelope-onset asynchrony) provides a reliable and robust cue for distinguishing voiced from voiceless consonants.

This acoustic cue was presented through a two-finger tactual display such that the envelope of the high-frequency band was used to modulate a 250-Hz carrier signal delivered to the index finger (250-I) and the envelope of the low-frequency band was used to modulate a 50-Hz carrier delivered to the thumb (50T). The temporal-onset order threshold for these two signals, measured with roving signal amplitude and duration, averaged 34 msec, sufficiently small for use of the EOA cue. Perceptual evaluations of the tactual display of EOA with speech signal indicated: 1) that the cue was highly effective for discrimination of pairs of voicing contrasts; 2) that the identification of 16 consonants was improved by roughly 15 percentage points with the addition of the tactual cue over L alone; and 3) that no improvements in L+T over L were observed for reception of words in sentences, indicating the need for further training on this task.

Thesis Supervisor: Nathaniel I. Durlach, Senior Research Scientist.
Thesis Supervisor: Charlotte M. Reed, Senior Research Scientist.

Acknowledgments

It is a great pleasure to acknowledge those who have helped me in the past years at MIT.

I would first like to express my sincere thanks to my advisors Charlotte and Nat. They have not only taught me how to conduct scientific research, but also taught me how to be a nice person. The research work would have been impossible without their advice, insights, encouragement and patience. Working with them has been enjoyable and rewarding. I am very glad to have the opportunity to continue to work with them in the future.

I would like to thank my thesis committee, Lou Braida and Ken Stevens, for their careful reading of the thesis, and many creative and valuable comments and suggestions. I am also grateful to their great introductions to two amazing areas: speech communication introduced by Ken, and auditory perception introduced by Lou.

A special thank to Andy Brughera for the funny times working together, and to Lorraine Delhorne for her help in speech segmentation.

I thank Thomas Weiss for providing me the opportunity of a teaching assistant in 6.003. I enjoyed working as the teaching assistant with Alan Oppenheim for both his great lectures and magic.

I thank every member in Sensory Communication Group and on the 7^{th} floor in Building 36 for providing such a friendly environment to live and study in.

Finally, I would like to thank my parents, husband, and other family members for their love and support. This thesis is dedicated to them.

This work was supported by a grant from the National Institutes of Health (DC00126).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Hearing Impairment

Hearing impairment refers to a decreased sensitivity to auditory stimulation. A classification of the degree of hearing impairment, which can range from mild to profound, is shown in Table 1-1. Hearing loss can be caused by damage at various points in the auditory pathway, leading to four primary types of hearing impairment: 1) Conductive hearing loss due to a problem in the outer or middle ear, 2) Sensory hearing loss due to the damage or destruction of the hair cells in the cochlea, 3) Mixed hearing loss arising from a combination of conductive and sensory hearing loss, and 4) Neural hearing loss due to damage in the neural pathway from the cochlea to the brain.

Table 1-1. Primary classification of hearing impairment.

	Very mild	Mild	Moderate	Moderately severe	Severe	Profound
Hearing loss (dB)	15~25	26~40	41~55	56~70	71~90	>91

More than 20 million Americans or 8.6 percent of total U.S. population are reported to have hearing problems (on the basis of data collected in the 1990-1991 census, Ries, 1994). By the mid-21st century, a projected 36 million people will be affected by hearing impairment due both to the increase in the overall population as well

as to the aging of this population. Over half of the hearing-impaired population currently consists of working-age adults, roughly one-third of whom are over the age of 65. The major consequence of hearing impairment lies in its interference with the ability to communicate through speech, leading to a series of negative effects on the lives of hearing-impaired people in many aspects such as income, activity and education.

The current research is concerned with the development of communication aids for persons with profound hearing impairment.

1.2 Communication Methods for Persons with Profound Hearing Impairment

Methods of communication used by hearing-impaired and deaf individuals can be grouped into two major categories: manual methods and oral methods.

Manual methods of communication for the deaf make use of manually produced gestures that are typically received through vision. Three main types of manual methods include sign systems that parallel the English language (as in Signed English), sign systems that are based on languages that are distinct from English (as in American Sign Language), and manual supplements to English (as in Cued Speech). While each of these methods has been proven to be an effective means of communication for deaf people, they require special training in the reception and production of the manual signals.

The goal of oral methods is to allow deaf people to communicate through the use of orally produced speech. Many of these methods are based on providing supplemental information to the cues available through lipreading. Aids to oral communication include conventional electroacoustic hearing aids, cochlear implants, auditory brainstem

implants, and tactual aids. Conventional electroacoustic hearing aids are typically used by persons with mild to severe hearing loss (25-90 dB HL). These devices receive sounds through small microphones, amplify the sounds overall or in selective frequency regions (according to the impaired auditory system) by an amplifier, and deliver the amplified acoustic signal to the external ear canal. People with severe hearing loss (above 90 dB HL) receive limited benefit from acoustic amplification. Hearing aids currently are being designed to provide information about specific parameters of speech as aids to lipreading for listeners with severe loss (e.g., Faulkner et al., 1992).

Other devices available for people with profound hearing impairment include cochlear implants, auditory brainstem implants (Rauschecker & Shannon, 2002), and tactual aids. Cochlear implants are designed to replace the function of the cochlea through direct stimulation of the auditory nerve. In cochlear implants, electrodes are surgically inserted into the tubelike ducts of the snail-shaped cochlea. The electrodes are actuated by a signal processor designed to translate acoustic energy into electrical waveforms that are delivered to the auditory nerve. However, cochlear implantation is not available to some profoundly deaf individuals due to absence or destruction of the auditory nerve. In such cases, auditory brainstem implants might be helpful by bypassing the cochlea and auditory nerve and stimulating the auditory processing centers of the brainstem.

In tactual aids, the acoustic signal is transformed into patterns of vibration presented to the skin. In general, the data indicate that the performance achieved through cochlear implants is superior to that through tactual aids. However, there are still many reasons to continue research on tactual aids. The first is that it is still difficult to compare

the relative effectiveness of cochlear implants and sensory substitution due in part to the limited data available for comparison. The second is that tactual aids are reversible. This is important especially for children because of the possible auditory changes associated with their growth, as well as the aid technology. The third is that there are individuals who are not well suited to cochlear implantation. And, finally, the cost of tactual aids is substantially less than that required for surgical implantation by roughly a factor of 10.

The research proposed here is directed towards the development of improved tactual displays of speech focused on the display of consonantal voicing as a supplement to lipreading. This research includes work on signal-processing schemes to extract information about voicing from the acoustic speech signal, methods of displaying this information through a tactual display, and perceptual evaluations of the effectiveness of the tactual cue for the reception of consonant voicing.

1.3 Outline of the Thesis

Chapter 2 provides a brief description of research on tactual perception and the history of tactual communication of speech, including the natural method of Tadoma and artificial tactile displays.

Chapter 3 describes the method of lipreading, and motivation for pursuing study of the feature of voicing as a supplement to lipreading. This chapter also includes a description of the articulatory properties of voicing and the acoustic cues associated with this feature.

Chapter 4 provides a description of three major approaches that have been applied to obtain information about the voicing property of speech: amplitude envelope

extraction, fundamental frequency extraction, and feature-based automatic speech recognition (ASR).

Chapter 5 describes a set of acoustical measurements of speech syllables undertaken to identify an acoustic cue that is robust for voicing detection, can obtained in real time (such that it can be synchronized with lip movements), and can be coded for delivery through the tactual sense.

Chapter 6 provides a description of the methods used in each of five perceptual experiments: 1) Absolute threshold measurements, 2) Temporal-onset order discrimination, 3) Pair-wise discrimination of voicing, 4) 16-consonant identification, and 5) Sentence recognition.

Chapter 7 provides a description of the results of each of the five experiments, and Chapter 8 presents a discussion of these results. Finally, Chapter 9 summarizes the major findings and Chapter 10 proposes some directions for future research.

Chapter 2

Background

Interest in the use of the tactual sense as a means of communication has existed for many years (e.g., see review by Reed, Durlach and Delhorne, 1992b). One key issue concerning tactual communication is whether the tactual sense is adequate for the task of transmitting and processing complex information such as speech. Assuming that the tactual sense does have this information-bearing capacity, it becomes important to develop methods for encoding the acoustic signals (typically used by the auditory system) into signals that are well suited for the tactual channel. An understanding of the basic properties of the tactual sense is essential for understanding both the transmission capacity and the limitations of the tactual channel and for background in the development of effective tactual displays.

In Section 2.1, a summary of studies of the basic characteristics of the tactual sense is provided in order to shed some light on the information transmission properties of the tactual channel. In Section 2.2, a natural method of tactual communication used by deaf-blind individuals (the Tadoma method) is briefly described. To some extent, this method provides evidence that the tactual sense alone is capable of receiving continuous discourse at near-normal communication rates. In Section 2.3, a brief description of research on the development of artificial tactual displays is provided along with a summary of performance through such displays.

2.1 Psychophysical Characteristics of the Tactual Sense

· Detection Thresholds

The response of the tactual system to different frequencies of sinusoidal stimulation is one of the most fundamental characteristics of a sensory system and is important in designing a tactual communication channel as a substitute for hearing. The study of this property has a long history (see Knudsen, 1928; Verrillo, 1963; 1966a; Bolanowski, 1988; Gescheider, Bolanowski, Pope and Verrillo, 2002).

The general findings are listed here:

1) The skin is differentially sensitive to frequency for large contactor areas (> 0.08 cm^2) with the most sensitive frequency in the vicinity of 250 Hz.

2) The tactual sense has a narrower operating frequency range (from roughly 0 to 1000 Hz) than hearing (from 20 to above 10000 Hz, although the spectrum below 8000 Hz is sufficient for speech communication).

3) Thresholds depend on temperature. Over a certain range of temperatures (15°C to 40°C), the higher is the temperature, the higher the sensitivity is (Bolanowski and Verrillo, 1982; Verrillo and Bolanowski, 1986).

4) Thresholds vary as a function of body site. Verrillo (1963, 1966b, and 1971) investigated the sensitivity of the skin at three body sites (thenar eminence, volar forearm, and middle fingertip) under identical experimental conditions. The results indicate that the fingertip is the most sensitive of the three sites, and the forearm is the least.

5) Thresholds depend on the stimulus duration. For shorter stimulus durations (< 200 msec) and large contactor area (2.9 cm^2), a 3-dB decline in thresholds occurs for each doubling of the stimulus duration (Verrillo, 1965).

6) Thresholds also depend on the area of contactor stimulation. Verrillo (1963) found that thresholds decrease at the rate of approximately 3 dB per doubling of contactor area at higher frequencies (80 ~ 320 Hz); however, the size of contactor has no effect on thresholds at lower frequencies (< 40 Hz).

· Tactile Sense vs. Kinesthetic Sense

The term "tactile" or "cutaneous" is used to refer to low-intensity, high frequency vibrotactile stimulation of the skin surface, whereas the term "kinesthetic" refers to the awareness of body postures, movements, and muscle tensions. The term "tactual" encompasses both components of the sensory system. Sensations are quite different at lower frequencies (below 100 Hz) than at higher frequencies. Subjects report a sensation of periodicity or "buzzing" at low frequencies and a more diffuse, "smooth" sensation at higher frequencies. Particularly, Tan (1996) noted that subjects could naturally categorize motions over a frequency range of near DC to 300 Hz into three perceptually distinctive categories: slow motion (up to about 6 Hz), fluttering motion (about 10 Hz to 70 Hz), and smooth vibration (above about 150 Hz).

· Frequency Discrimination

The skin is quite limited in its frequency discrimination ability compared with the auditory system (a difference limen of 30% for tactual versus 0.3% for auditory, Goff,

1967). Tan (1996) reported that subjects could reliably identify only two frequencies within each of the three distinct frequency regions defined above.

· Intensity Discrimination

The dynamic range of the tactual system is limited (roughly 55 dB dynamic range for touch versus 130 dB for the auditory system), and is also inferior in intensity discrimination. Different studies have reported different values of the detectable intensity difference for tactual stimulation, ranging from 0.4 dB (Knudson, 1928) to 2.3 dB (Verrillo and Gescheider, 1992).

· Amplitude Modulation Sensitivity

Amplitude modulation sensitivity refers to the ability to detect amplitude modulation, which is another measure of temporal sensitivity. Maximal sensitivity occurs at modulation frequencies in the region of 20-40 Hz and sensitivity is greater for sinusoidal carriers than for narrow or wide-band noise carriers (Weisenberger, 1986).

· Gap Detection

Gap detection is defined as the ability to detect the silent temporal interval (gap) between two tactual stimuli (markers). Tactile gap detection thresholds are in general substantially higher than auditory thresholds. Gap-detection thresholds vary from as low as 5 msec for highly damped mechanical pulses (Sherrick, 1982) to several tens of msec using sinusoidal markers (Formby et al., 1991), and depend on both physical attributes of the maskers and subjective factors such as age (Van Doren et al., 1990).

Temporal Order

Temporal order refers to the ability to determine the order of two different signals. Various experiments concerning temporal-order discrimination ability have been conducted through different modalities (including hearing, vision, and touch), and with different stimulus parameters. Among them, Hirsh and Sherrick (1961) found the threshold for temporal-order judgments between the onsets of two brief stimuli to be roughly 20 msec, independent of modality. A more complete summary of previous studies is provided in section 8.2.1.

2.2 Tadoma

Tadoma is one of the natural methods of tactual communication for the deaf-blind relying on the tactual sense alone. It was once employed as an educational method for deaf-blind children (Hansen, 1930). There are now only about 20 deaf-blind adults in the USA who are competent in the use of Tadoma. However, this method has theoretical importance regarding the capability of the tactual sense as a channel for speech communication. In the Tadoma method, the receiver places a hand on the face and neck of the talker with the thumb resting lightly on the lips and the fingers fanning out over the face and neck. The receiver comprehends speech by monitoring various actions associated with speech production (e.g., lip and jaw movements, airflow at the mouth, vibrations on the neck).

Previous research (Reed et al., 1985) has documented the performance of experienced Tadoma users. Average Tadoma scores from a group of 9 experienced Tadoma users indicate the following: roughly 60% correct for 24-consonant identification

in C-/a/ syllables, 45% for 15-vowel identification in /h/-/V/-/d/ syllables, 40% for identification of open-set monosyllabic words, and 65% for key words in CID sentences. The structure of confusion matrices derived from segmental identification tests (Reed et al., 1982b) indicates that the features of voicing, place of articulation, frication, and lip rounding are well perceived for consonants, as are the features of tenseness, vertical lip separation and lip rounding for vowels. In order to gain further insight into relationship between the articulatory cues and the specific features perceived through Tadoma, Reed et al. (1989b) conducted an analytic study employing systematic variations in the accessibility of four major articulatory components of the Tadoma display. The percent-correct scores increased uniformly with the number of cues available to the subjects. Results also indicate that the reception of specific features is related to particular articulatory cues (e.g., voicing perception is highly dependent on access to laryngeal vibration on the neck). The observation that removal of one cue from the normal Tadoma display did not have a large differential effect on performance across the various consonant or vowel features indicates that redundant information is provided by the cues in normal Tadoma.

Although Tadoma is no longer widely used, it has a very important place in the history of tactual speech communication by providing evidence that the tactual sense can serve as an effective channel for speech. Tadoma also provides important background information for the development of artificial tactile devices.

2.3 Artificial Tactile Displays

The development of artificial tactile devices to present acoustic information to persons with profound hearing impairment has a long history. One of the first devices employed the diaphragm of a special telephone receiver (Gault, 1924, 1926) to deliver unprocessed sound vibrations to the fingers or hands. Since then, various artificial tactile displays have been developed as aids for speech communication (see reviews by Kirman, 1973; Reed et al., 1982a; Reed et al., 1989a; Bernstein, 1992). These devices convey aspects of the speech signal by means of tactile patterns presented to arrays of stimulators applied to the skin.

These devices have included different types of stimulation (either mechanical or electrocutaneous), have been applied to a variety of body sites (e.g., finger, hand, forearm, abdomen, thigh), and have employed various numbers and configurations of stimulators (e.g., single-channel or multi-channel stimulation, linear or two-dimensional arrays). There are several characteristics that are common to most of these devices, however. First, they rely on place of stimulation to transmit information. Second, they stimulate the cutaneous component of the tactual sense, with little involvement of the kinesthetic component. Third, they tend to employ homogeneous stimulation of all the vibrators in the stimulatory array.

Much of the evaluation of the performance of tactile speech communication aids has been concerned with the benefits they provide to lipreading. The size of the benefits provided by tactile supplements to lipreading is fairly similar across a variety of types of tactile devices. For example, the reception of words in sentences is improved by roughly 5 to 15 percentage points over lipreading alone through the addition of tactile information

in a variety of different forms. Such tactile devices include single-channel displays of envelope information (Besing et al., 1995), multi-channel displays of the energy in filtered bands of speech (Reed et al., 1992a), as well as single or multi-channel displays of fundamental frequency (Hanin et al., 1988).

Most of the tactile displays obtained significant improvement as a supplement to lipreading; however, the performance obtained with artificial displays is generally inferior to that obtained through Tadoma. The difference may be attributed to the following factors: 1) In Tadoma, the hand is used for sensory input, while less sensitive body sites such as abdomen and forearm are often used in artificial displays. 2) The Tadoma signal is multi-dimensional, engaging kinesthetic as well as cutaneous components of the tactual system, while most artificial displays stimulate only the cutaneous system in a homogeneous manner. 3) Tadoma is tied to the articulatory process, while most tactual displays are spectral-based. 4) Tadoma users have received more training time than users of most artificial tactual displays.

Chapter 3

Supplementing Lipreading

Most tactual displays have been used to supplement lipreading. One reasonable strategy is to determine the information that is difficult to obtain through lipreading and to present this information to the skin through a tactual display. Speech features that are not readily available through lipreading include both suprasegmental and segmental properties. At the suprasegmental level, these features include properties such as stress, intonation and inflection. At the segmental level, features can be identified which are necessary to differentiate between visually similar phonemes (i.e., visemes). Section 3.1 provides a brief summary of speech reception through lipreading alone. In Section 3.2, the articulatory process related to voicing and the acoustic cues of voicing are discussed, serving as the basis of the design of a tactual voicing display. Although, the focus of the current research is on tactual communication of the feature voicing, similar methods can be applied to other features, such as nasality and manner, which are also not well conveyed through lipreading alone.

3.1 Lipreading

Lipreading is an important means of obtaining information about speech for hearing-impaired persons. The process of lipreading itself is robust in many aspects. Lipreading, or the ability to obtain speech information from the face, is not significantly

compromised by a number of variables. Humans are fairly good at lipreading from various angles and can perform well even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred, when the image is rotated toward a profile view, when it is viewed from above or below, or when there is a large distance between the talker and the viewer (Massaro, 1998, p. 429). Furthermore, people can naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about 200 msec.

Lipreading is largely based on the visibility of various lip features that arise during articulation of speech. A list of lip features is provided in Table 3-1 for both vowels and consonants (Jeffers and Barley, 1971). This information divides the phonemes into several groups with identical or highly similar visible articulation movements, but is insufficient for absolute speech perception.

Table 3-1. Lip features.

Vowel	**Consonant**
Lip shape (puckered, back, rounded, or relaxed)	Lip articulation
Lip opening (narrow or moderate)	Degree of oral cavity opening
Movement (present in diphthongs or absent).	Place of articulation

Various phonemes in the same group are visually noncontrastive or indiscriminable and are referred to as visemes (Fisher, 1968; Walden, et al., 1981; Owens and Blazek, 1985). Although the exact composition of viseme classifications can vary across studies (depending on factors such as speaker, stimuli, subjects, and statistical criteria), there is some consistency across studies. The consonant viseme groups found in the majority of studies are /p,b,m/, /f,v/, /w,r/, /th,tx/, /ch,j,sh,zh/, /k,g/ and /t,d,s,z/.

Reed et al. (1978) summarized the performance on identification of consonants by lipreading from three studies (Heider and Heider, 1940; Erber, 1974; Walden, et al., 1977), listed in Table 3-2. Also included in Table 3-2 are more recent data from Bratakos et al. (2000). Heider and Heider (1940) studied the identification of 20 consonants in CV syllables with V=/oy/ or /I/. A teacher presented the stimuli to 39 deaf students, who each contributed one response to each consonant in each context. In Erber's study (1974), 20 consonants in VCV syllables with V=/i/, /a/, or /u/ were presented to 6 deaf students. Each student contributed 10 responses to each consonant in each context. In the study of Walden et al. (1977), 20 consonants were presented in C/a/ syllables, and each of 31 subjects contributed 20 responses for each consonant. Bratakos et al. (2000) employed 24 initial consonants in C-/a/-C syllables, and each of 3 subjects contributed 16 responses per consonant. By examining the confusion matrices obtained through these studies in terms of information transfer on various features, it is found that the features of voicing and nasality are poorly perceived, whereas the features place, frication, and duration are better perceived by lipreading. By providing supplemental information for voicing and nasality with lipreading, the viseme groups can be reduced to smaller groups, thereby leading to improved lipreading ability.

Table 3-2. Proportion of unconditional information transfer on various features through lipreading.

Feature	Heider & Heider	Erber	Walden et al.	Bratakos et al.
Voicing	0.16	0.06	0.02	0.12
Nasality	0.16	0.12	0.02	0.21
Frication	0.78	0.57	0.80	0.63
Duration	0.28	0.63	0.75	0.29
Place	0.81	0.80	0.68	0.78

In this research, the focus is on tactual transmission of the feature voicing. The next section summarizes previous research on production and acoustic properties related to consonant voicing.

3.2 Consonant Voicing

3.2.1 Conditions for Voicing Production

Typically, voicing refers to the distinction made between two classes of segments, related to the presence (voiced) or absence (voiceless) of vocal fold vibration. Voicing is a distinctive feature in the case of obstruent consonants, i.e., stops, fricatives and affricates. That is, two obstruent consonants may be the same in all features except voicing. There are three conditions for vocal-fold vibration (Stevens, 1998): the first is sufficient transglottal pressure, the second is that the vocal folds are placed together, and the third is that the vocal folds are slack. In the production of voiced consonants, to maintain vocal-fold vibration, the volume of the vocal tract must expand so that the transglottal pressure is sufficient to maintain the vibration. Part of the expansion is obtained by lowering the larynx. This is generated by contraction of the strap muscles that extend from the hyoid bone and the thyroid cartilage to the sternum. The lowering of the larynx can also cause a downward tilt of the cricoid cartilage. The tilting of the cricoid cartilage causes a shortening and hence a slackening and thickening of the vocal folds. On the other hand, in the production of the voiceless consonants, the volume increase is inhibited by stiffening the vocal tract wall and the vocal folds, and the glottis is spread apart. Therefore, the air pressure above the glottis is allowed to build up, thus

causing cessation in the airflow through the glottis, and cessation of vocal fold vibration. And it can be accompanied by aspiration noise if there is a spread glottis after the release.

3.2.2 Acoustic Cues of Voicing

The acoustic cues for voicing distinction in English obstruent consonants have been examined extensively both through studies in which measurements are made from naturally spoken utterances and through studies utilizing edited natural speech or synthetic stimuli in which different acoustic parameters are manipulated. A variety of acoustic cues have been found that provide information for the distinction of voicing between pairs of obstruent consonants produced at the same place in the vocal tract in various contexts. This section is organized according to individual voicing cues: voice onset time, frication duration, formant transitions, F0 perturbations, vowel duration, amplitude of the first harmonic and harmonic structure. With each voicing cue, a description of the acoustic properties of these cues is provided, along with an explanation of their origin in production and their role in perception.

3.2.2.1 VOT

Voice onset time (VOT) is a cue extensively investigated both by measurement and synthetic experiments for pre-stressed English stops. It refers to the time from the release of the closure made by an articulator at a particular point along the vocal tract to the onset of vocal-fold vibration (see Fig. 3-1, taken from Zue, 1976). The measurements from natural utterances show that the VOTs for the pre-stressed stops are clustered around 0-20 msec for the voiced stops and around 50 msec for the voiceless stops. The

results also show that VOT varies as a function of the place of the articulation of the stop, smaller for labials and larger for velars (Lisker and Abramson, 1964; Zue, 1976). The data from synthetic experiments also show that VOT provides a direct cue for voicing perception. If this time is greater than about 25-40 msec (depending on consonantal place of articulation), the consonant is identified as voiceless; otherwise it is identified as voiced. These data also show that there is a sharp perceptual boundary between voiced and voiceless responses. The discrimination within a given class is poor, while discrimination near the phoneme boundary is good.

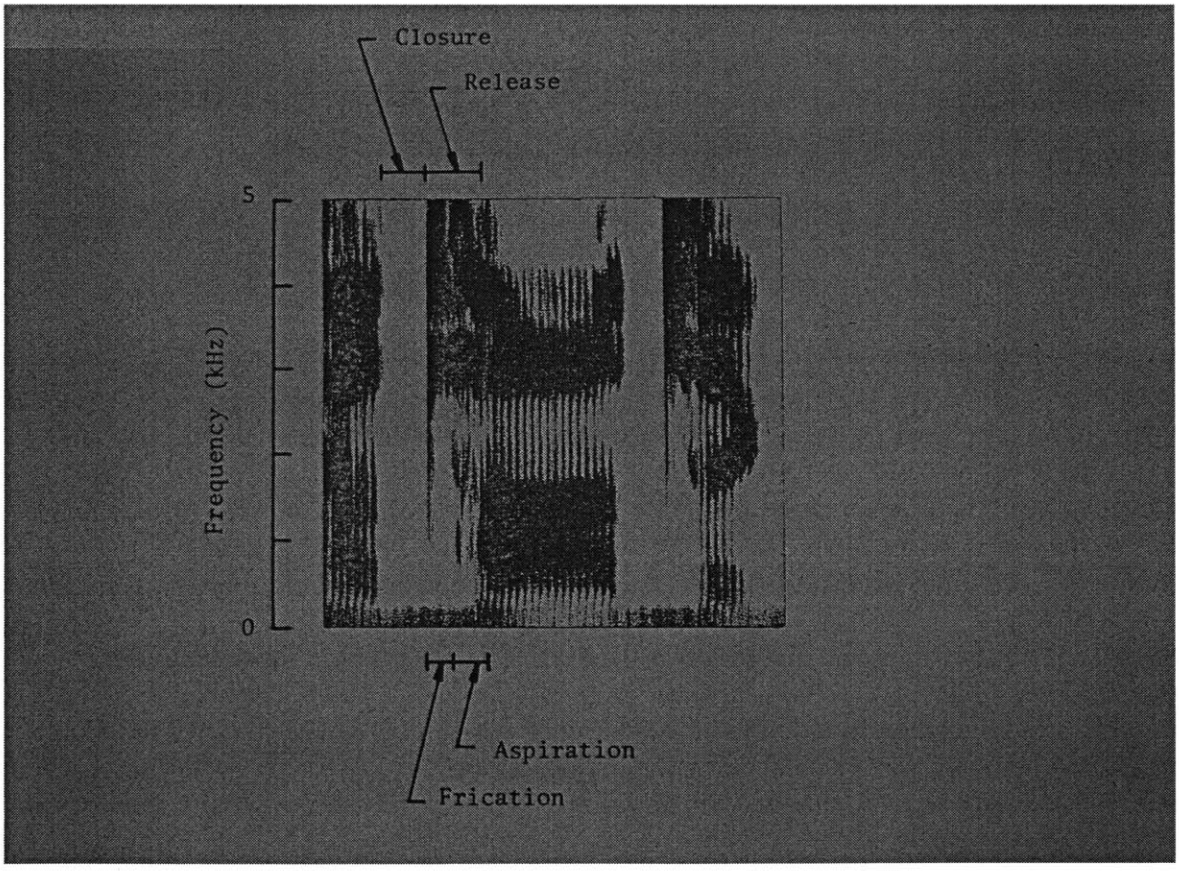

Fig. 3-1. Illustration of a series of events during the production of a voiceless stop. (Taken from Zue, 1976, p. 76).

The following VOT data are selected from measurements of natural speech stimuli taken by Zue (1976):

Table 3-3. VOT of the voiced stops.

	/b/	/d/	/g/	Avg
VOT (msec)	13	19	30	20.6

Table 3-4. VOT of the voiceless stops.

	/p/	/t/	/k/	Avg
VOT (msec)	58.5	69.5	74.5	67.5
Frication (msec)	27	35	43	35
Aspiration (msec)	31.5	34.5	31.5	32.5

From Tables 3-3 and 3-4, it can be seen that the VOT varies as a function of the place of the articulation. It is shortest for the labial and longest for the velar both for voiced and voiceless stops. The aspiration duration for the voiceless stops is independent of the place of the articulation.

The variation of VOT with the place of the articulation can be explained by the different articulators and the length of the region of contact of the oral constriction. In general, the movement of the tongue body is slower than that of the tongue tip and lips for the alveolars and labials; moreover, the length of the velar constriction is longer than the lengths of the labial and alveolar constrictions. The constriction for the labials is formed at the lips, therefore no front resonator is formed. This can explain why the labial release is usually weak in intensity. In addition, the lips can move away quite rapidly following the release, which also contributes to the appearance of a short burst. The constriction for velars is formed by the tongue body, which is rather massive and cannot move away from the palate too rapidly following the release. A tapered, narrow opening

is maintained for a longer period of time; therefore the constriction for /g/ opens slowly, allowing turbulence noise to be generated for a longer period of time (Houde, 1967). Moreover, the intraoral pressure builds up more rapidly for a velar closure due to the smaller compliance of the vocal tract walls. This intraoral pressure will exert force on the surface of tongue body, leading to an earlier release than could be achieved without the intraoral pressure. This effect of intraoral pressure on the time of release is greatest for the velar stops.

A possible explanation for the aspiration durations (the time from completion of the initial burst of frication noise to voicing onset) being approximately equal for different places of constriction is given by Klatt (1975). During the production of the voiceless stops, the glottis is spread apart at the closure. After the release, the drop of the intraoral pressure initiates the closure of the glottis. The time required for the glottis to change from a spread position to a position appropriate for vocal-fold vibration is around 50 msec and is independent of the supraglottal movement.

3.2.2.2 Frication Duration

A narrow constriction in the vocal tract is formed in the production of both fricatives and affricates. A rapid flow of air through the constriction creates turbulence in the flow that acts as a source of sound (i.e., the source is in the vicinity of the constriction instead of at the glottis). The frication is then generated by the turbulence filtered by the acoustic cavity in front of the constriction.

Acoustic measurements in previous studies indicate that frication duration provides a robust cue to the voicing distinction in syllable-initial position, with voiceless

fricatives having longer frication durations than voiced fricatives (although with considerable overlapping of the two classes), both for fricatives in isolated syllables and in connected speech (Behrens and Blumstein, 1988; Baum and Blumstein, 1987; Jongman, Wayland, and Wong, 2000; Crystal and House, 1988). Stevens et al. (1992) measured the frication duration of fricatives in intervocalic position, and found that the frication duration of a voiceless fricative was greater than that for a voiced fricative. However, they pointed out that this difference in duration of the cognate fricatives is due to differences in adjustment of glottal configuration rather than to differences in the timing of supraglottal movements. In other words, the difference in duration will disappear when the falling F1 transition in the first vowel and rising F1 transition in the second vowel are included in the measurements.

Cole and Cooper (1975) examined the effect of frication duration on the perception of voicing distinction for the English affricates /ch-j/, and the fricatives /f-v/ and /s-z/. The frication durations of the voiceless consonants /ch/, /f/ and /s/ in C/a/ context were systematically manipulated by removing part of the frication just prior to the vowel. Their results indicated that the shortening of the frication of the voiceless consonants produced a change in the perception from voiceless to voiced. On the other hand, Jongman's (1989) perceptual results indicated that the frication duration had only a minor effect on the subjects' judgments of consonant voicing, thus rejecting the findings of Cole and Cooper (1975). Stevens et al. (1992) also studied the perceptual effect of fricative duration on consonant voicing with both synthetic and edited natural speech in VCV context. The results using synthetic speech indicated that frication duration played a role in the perception of voicing; however, the results of the experiment using edited

natural speech were similar to those obtained by Jongman, i.e., the effect of frication duration on consonant voicing was small.

Thus, it seems that the claim that frication duration provides a primary cue of consonant voicing in English fricatives is still under discussion.

3.2.2.3 Formant Transition

a). Stops in Pre-stressed Position

The presence of formant transitions after voicing onset has been found to be a cue for voicing for stops in pre-stressed position (Liberman, Delattre, and Cooper, 1958; Stevens and Klatt, 1974). Liberman et al. examined the responses of listeners to a series of CV synthetic stimuli that differed only in the amount of F1 transition by cutting back the beginning of the F1 transition. The cutback of the first formant raises its starting frequency and also delays the time at which it begins relative to the other formants. The results show that the stimuli with larger F1 cutback were identified as voiceless. The boundary between voiced and voiceless responses occurred when the cutback was at a point where most of the F1 transition was completed. In Stevens and Klatt's (1974) synthetic experiment, they manipulated the VOT and transition duration independently to compare the role of VOT and formant transition following voicing onset for the voicing distinction. The data indicate that there is a significant trading relationship between these two cues. The presence or absence of a significant spectral change following voice onset produces up to a 15-msec change in the boundary in terms of VOT.

In the production of the voiced stops, there is vocal-fold vibration after the release, and it is accompanied by supraglottal movements. According to the perturbation

theory, the shift in a formant frequency due to a localized perturbation in the cross-sectional area of the vocal tract depends on the distribution of sound pressure and volume velocity amplitude along the tract. At a point where the volume velocity is at a maximum, an increase in cross-sectional area causes an increase in the formant frequency. For the labial stop, the constriction is made at the lips; after the release, all of the formants will increase in frequency as the area of the constriction becomes larger, with the higher-frequency formants increasing somewhat less rapidly than the lower frequencies. Similar formant transitions will occur for both alveolar and velar stops. However, in the production of the voiceless stops, after the release, the glottis is still spread; thus, there is no vocal-fold vibration. The vibration begins only after the glottis is sufficiently adducted. By this time, the rapid movements of the supraglottal articulators are essentially complete. Therefore, no well-defined formant transition will be found in the voiceless stops, although the transition may be reflected in the aspiration.

b). Stops in Final Position (F1 structure)

Hillenbrand et al. (1984) made acoustic analyses of naturally produced speech that suggested that syllable-final consonants identified as voiceless had higher F1 offset frequencies than consonants perceived as voiced. This effect is greatest for open vowels like /e, a/ with a high-F1 steady state and least for more constricted vowels such as /i, u/ with low-F1 steady state values (see Table 3-5). Perceptual research has shown that the first formant transitions (F1FT) influence final-stop voicing decisions. When vowel durations are approximately equal, utterances containing falling F1FTs and low F1 offset frequencies are judged as ending in voiced stops more often than utterances without

F1FTs or with gradual F1FTs, which terminate at higher frequencies. Fischer and Ohde's experiment (1990) shows that both vowel duration and F1 offset influence perception of final stop voicing, with the salience of the F1 offset property higher for vowels with high-F1 steady-state frequencies than for these with low-F1 steady-state frequencies.

Table 3-5. Frequency of F1 at offset for the vowels /i/, /a/, /e/, and /u/ as a function of voicing value of the following consonant (vd= voiced, uv=voiceless). Measurement of F1 offset frequency was based the linear predictive coding analysis. (From Hillenbrand et al., 1984).

Vowel	/e/ vd	/e/ uv	/a/ vd	/a/ uv	/i/ vd	/i/ uv	/u/ vd	/u/ uv
F1 Offset (Hz)	305	486	334	410	232	249	159	182

From Table 3-5, it is seen that F1 offset frequency is higher for the vowels preceding voiceless compared to voiced stops. The differences in F1 offset are larger for vowels with higher-frequency first formants than with lower first formants.

For syllables ending in voiceless stops, a laryngeal gesture typically terminates voicing at about the same time that articulatory closure is achieved. For syllables ending in voiced stops, glottal vibration generally continues into at least some portion of the closure interval. Because of the early termination of glottal vibration in final voiceless stops, these syllables are generally characterized by (1) relatively high first formant terminating frequencies and (2) shorter intensity decay times (Wolf, 1978).

c). Fricatives

Stevens et al. (1992) systematically studied the effects of F1 transitions on voicing distinction for fricatives. Initially, predictions were made for the extent and

duration of F1 transitions for voiceless and voiced fricatives. For the voiced fricatives, glottal vibration continues as the supraglottal constriction is formed; thus, a change in F1 (either an increase in the case of C-V context or a decrease in the case of V-C context) can be observed according to the perturbation theory. For voiceless fricatives, however, glottal vibration ceases during the formation of the supraglottal constriction; thus, only a part of the change in F1 can be observed, resulting in shorter duration and smaller extent of the F1 transition. Acoustical measurements of the extent and duration of the F1 transitions were obtained at the VC and CV boundary for /s/ and /z/ in a VCV environment for three speakers. The results of acoustical measurements were consistent with the theoretical predictions. The change in F1 was substantially smaller in V-/s/ and /s/-V boundaries than that in V-/z/ and /z/-V boundaries. In addition, the duration of F1 transitions was relatively shorter for /s/ than for /z/. Stevens et al. (1992) also conducted perceptual experiments on the effect of the F1 transition on voicing distinction of fricatives in VCV context. The results indicated that the F1 transition did affect the perception of voicing, although this effect was interactive with other factors such as frication duration.

3.2.2.4 F0 Perturbations

F0 perturbation refers to the pattern of change in fundamental frequency (F0) as a function of time in vowels adjacent to obstruent consonants. A falling fundamental frequency usually occurs after a voiceless consonant, while a flat or rising F0 usually accompanies voiced consonants (House and Fairbanks, 1953; Ohde, 1984; Whalen et al., 1993).

The possible origin of the production of F0 perturbation is due to differences in tension in the vocal folds and to positions of the larynx and hyoid bone for voiced and voiceless consonants (Stevens, 1998). In the production of voiced consonants, the larynx is lowered, leading to a downward tilt of the cricoid cartilage. The tilting of the cricoid cartilage causes a shortening and hence a slackening and thickening of the vocal folds, which carries over to the following vowel. The fundamental frequency is lower at the beginning of the vowel and then increases. On the contrary, in the production of voiceless consonants, the vocal folds are stiffened. This stiffening carries over to the following vowel, and therefore the fundamental frequency is higher at the beginning and then falls off.

3.2.2.5 Vowel Duration

The extent to which the preceding vowel duration can act as an acoustic cue for the voicing distinction of consonants in the final position has been investigated in several studies. House (1961) found that vowels preceding voiceless consonants are shorter in duration than those before voiced consonants. House (1961) suggested that this shortening of the vowels before voiceless consonants is probably an articulatory activity arbitrarily imposed by the phonology of English and is learned by persons who speak English.

Perceptual experiments on the effects of vowel duration on the perception of consonant voicing have also been conducted. Denes (1955) investigated the effect of vowel duration on the perception of voicing by physically manipulating the pair of words: /juz/ and /jus/, which were recorded natural utterances. He found that both

frication duration and vowel duration (or the relative duration of vowel and final consonant) can be used as an effective cue for voicing. Raphael (1972) investigated the effect of vowel duration on voicing perception of word-final stops, fricatives, and consonant clusters using synthetic speech. He found that the final consonants or consonant clusters were perceived as voiceless when the preceding vowels had short duration, and as voiced when the preceding vowels had long duration, regardless of other cues used in the synthesis of the final consonant or cluster. Vowel duration has only a small effect on the perception of voicing for syllable-initial affricates and fricatives (Cole and Cooper, 1975).

3.2.2.6 Amplitude of the First Harmonic

Theoretical analysis indicates that the glottal vibration extends over a longer time in the obstruent interval for voiced fricatives than for voiceless fricatives, and the glottal vibration always occurs at the CV or VC boundary of the fricatives (Stevens et al., 1992).

The amplitude of the first harmonic reflects the strength of the glottal vibration. Thus, the amplitude difference between the first harmonic H1 in the frication noise and that of the adjacent vowel can be used as a measure of the absence or presence of glottal vibration (Stevens et al., 1992). Specifically, a decrease by 10 dB of the amplitude of H1 relative to its value during the vowel at the VC boundary or an increase by 10 dB at the CV boundary was used as the threshold for the absence or presence of glottal vibration. Based on this, Stevens et al. (1992) proposed an operational criterion to distinguish the voicing feature of a fricative, i.e., a fricative was classified as voiced if at least 30 msec of glottal vibration was present at either the onset or offset of the frication noise;

otherwise, it was voiceless. Acoustical measurements made on VC(C)V utterances indicated that this criterion classified the majority of the fricatives in terms of voicing.

The stability of this criterion was tested by Pirello et al. (1997) under a variety of other contextual conditions (CV syllables in isolation or in contexts following voiced and voiceless velar stops). Results of the acoustical measurements indicated that this acoustic measure reliably classified the consonants in terms of voicing under different contextual conditions, despite the fact that the carryover influence of the context is short-lived.

3.2.2.7 Harmonic Structure

In the production of the voiceless consonants, the glottis is spread. A spread glottis gives rise to a larger decline in the amplitude of higher harmonics than does an adducted glottis. A larger difference in the amplitude between the first two harmonics is observed for vowels near voiceless compared to voiced consonants (Choi, 1999). This harmonic-amplitude difference serves as yet another acoustic cue for the voicing distinction for initial consonants in the initial position.

A variety of acoustic cues to voicing can be found in the large and still growing literature. Production of the voicing contrast consists of a sequence of articulatory and aerodynamic events. Each of the underlying events has multiple acoustic manifestations. Therefore, it is not surprising to find many acoustic cues that play a role in voicing contrast. The challenge of the current research is to determine an effective method for reliably capturing these cues in the acoustic speech signal and displaying this information to the skin.

Chapter 4

Methods for Obtaining Voicing Information from Speech

Three possible approaches towards obtaining information about voicing from speech are described here. These approaches include two acoustically-based methods of signal processing (to obtain information about the amplitude envelope of filtered bands of speech or to specify fundamental frequency). In addition, we discuss the use of automatic speech recognition as a tool for identifying the voicing property.

4.1 Amplitude Envelope

A brief description of a general signal-processing scheme for extraction of the amplitude-envelope signal is provided in Section 4.1.1, as well as a discussion of the effects of various parameters on the resulting envelope signal. In Section 4.1.2, a review of previous studies of the reception of amplitude-envelope cues derived from speech is presented.

4.1.1 Signal Processing

The amplitude envelope preserves information related to the gross temporal-intensive contour of a signal, but greatly reduces the information concerning rapid fluctuations in the signal. It is typically generated using the method described by Horii,

House, and Hughes (1971) in which the speech signal is pre-filtered and rectified subsequent to being smoothed by a low-pass filter. The envelope signal is then multiplied by a carrier signal that is used to convey fluctuations in envelope amplitude (see Fig. 4-1).

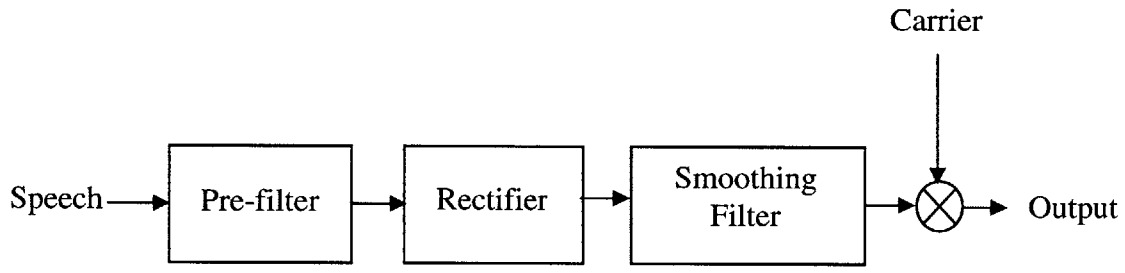

Fig. 4-1. Block diagram for speech processing.

The functions of each block in Fig. 4-1 are summarized by Grant et al. (1991).

The pre-filter determines the spectral region(s) of the speech signal from which the envelope is derived. It can be an all-pass filter such that the entire speech signal is used, or it can be a band-pass filter such that only a narrow spectral region of the original speech signal is used.

The rectifier flips the negative amplitudes of the pre-filtered speech signal into positive amplitudes, thus retaining only information about the absolute value of amplitudes. The rectified signal reflects the energy level of the speech in the spectral region determined by the pre-filter. A side effect of the rectification is that the bandwidth of the rectified signal can be increased substantially.

The smoothing filter determines the envelope bandwidth, which in turn determines the degree to which the envelope can follow rapid fluctuations in speech energy.

The carrier determines the spectral center and timbre of the resulting amplitude-modulated signal and can be used to transpose the envelope information to any frequency region. Pure tones, complex tones, and noise carriers have been used to vary the quality and bandwidth of the resulting amplitude-modulated signal. It is found that both warping and shifting make the speech less intelligible and have a more detrimental effect on consonants than on vowels (Shannon et al., 1998; Grant et al., 1994).

4.1.2 Perceptual Studies of Envelope-Based Speech

a). Auditory Alone

Hill et al. (1968) investigated the auditory recognition of speech as a function of the number of channels of spectral information provided by a vocoder system. In this system, the speech signal was passed through N band-pass filters with the center frequencies arranged logarithmically between 180 and 4200 Hz. The output of each band-pass filter was rectified and low-pass-filtered at 15 Hz. The output of each band was then used to control the output amplitude of an audio oscillator tuned to the center frequency of the band-pass filter. The oscillator outputs were then added, and the resultant signal was used to drive the earphones. Essentially, the speech signal is distorted by lumping the powers originally spread over finite frequency bands into single frequencies. The ability to identify consonant and vowel speech segments increased as the number of filters increased. Six to eight channels of spectral information led to roughly 70% correct

intelligibility. The experience gained in listening to male-voiced phonemes was not easily transferred to the recognition of female-voiced phonemes. Vowels and periodic consonants were almost never confused with aperiodic consonants.

Van Tasell et al. (1987, 1992) studied the recognition of consonants in /aCa/ context using cues available in the time-intensity envelope of speech. Three different envelope signals were studied, derived using three low-pass filters with different cutoff frequency (20, 200, and 2000 Hz) applied to the full-wave rectified speech. The product of the resulting envelope signal and a 3000-Hz low-pass filtered pink noise was then smoothed with a 3000 Hz LPF. Increasing envelope bandwidth from 20 to 200 Hz significantly improved the performance of the normal-hearing listeners, while the further increase from 200 to 2000 Hz did not. This implies that lower-frequency temporal information (F0) can be used to support consonant phonemic identification; higher-frequency temporal information in the amplitude envelope of the speech waveform apparently is not useful for consonant identification. Inter-subject variability was fairly large, indicating that individual subjects may differ in their ability to use envelope cues.

In Turner et al.'s (1995) experiments, speech was filtered into a high-pass band and a low-pass band, divided at 1500 Hz. The envelopes from the high-pass and low-pass band of speech were used to modulate a high-pass band of noise and a low-pass band of noise, respectively. Recognition of vowels and consonants improved dramatically with two modulated noise bands compared to one. This result demonstrated that temporal envelope information, quantified into two spectral bands, was sufficient for overall high recognition performance on consonants and was sufficient for recognition of almost 100% of the voicing and manner information. When the number of the bands was

increased from 2 to 4 (Shannon et al., 1995), the further improvement was primarily due to increased information on place of articulation. The two bands of noise provided a severely degraded spectral representation of vowel and consonant formants: almost no formant structure and formant frequency transitions were presented. Thus, the high recognition of voicing and manner indicates that those cues require only minimal spectral information. However, consonantal place of articulation and vowel recognition depend more heavily on spectral cues than on temporal cues for high levels of recognition.

b). Envelope Cues as an Auditory Supplement to Lipreading

Breeuwer and Plomp (1984) investigated the benefits of auditorily supplementing lipreading with envelope-based signals in normal-hearing listeners. Envelopes were derived from bands of speech with center frequencies of 500, 1600, and 3160 Hz (with 1- or 1/3 –oct bandwidth), and with a 20-Hz smoothing filter. The envelopes then modulated a tone at the center frequency of the pre-filter. The maximal single-band benefit was achieved with the 1-oct band at 500 Hz, where the mean number of correctly perceived syllables in sentences increased from 22.8% for lipreading alone to 65.7% for supplemented lipreading. The maximal benefit for a two-band envelope signal was achieved with two 1-oct bands at 500 Hz and 3160 Hz, where the mean number of correctly perceived syllables was 86.7%.

Grant et al. (1991) investigated the effects of the variation of several parameters (including pre-filter, carrier, and smoothing filter) on the benefits obtained by auditory presentation of single-band envelope signals as a supplement to lipreading in normal-hearing subjects. The signals were generated from different combinations of pre-filters

(including wide-band speech and octave bands of speech centered at 500, 1600 and 3150 Hz), smoothing filters (ranging from 12.5 to 200 Hz), and carrier signals (including wide-band noise and pure-tones ranging from 200-3150 Hz). The signal that produced the largest benefits to lipreading of sentences was derived from the 500-Hz octave-band of speech, which was filtered by a 100-Hz smoothing filter, and finally multiplied by a 200-Hz carrier signal. With this signal, the ability to recognize words in sentences improved by roughly 35-40 percentage points for the aided condition compared to lipreading alone for both high- and low-context sentence materials

Grant et al. (1994) investigated the benefits to lipreading provided by simultaneous auditory presentation of amplitude-envelope cues across two different frequency regions (using octave bands of speech centered around 500 Hz and 3150 Hz, a 50-Hz smoothing filter, and carrier signals at or below the center frequency of the pre-filter). Aided lipreading scores for sentences with two-band envelopes were superior to single-band performance only when the carrier frequencies of the two-band cues were close to the center frequencies from which the envelopes were derived. Downward transposition of the carrier frequencies in two-band envelope signals resulted in poorer performance than non-transposed envelope cues.

c). Envelope Cues as a Tactual Supplement to Lipreading

Besing, Reed, and Durlach (1995) investigated the capability of the tactual sense to convey the information carried by an envelope signal as a supplement to lipreading. The best envelope from the study of Grant et al. (1991) was selected for auditory and tactual presentation of the supplementary cue (500-Hz octave band of speech with a 200-

Hz carrier). The benefit to lipreading of sentences from the tactual supplement (L+T)[1] was roughly 1/3 of that obtained for auditory presentation (L+A) in the same group of normal-hearing listeners. Further experiments examined the reception of several prosodic features. For the Pitch Rise/Fall test [in which the subject's task was to determine whether a given sentence was spoken as a question (rising) or as a statement (falling intonation)], scores across subjects showed no systematic pattern for improvements to lipreading with either of the two supplements (51% for L, 48% for L+T, and 52% for L+A). Performance was somewhat better for determining the location of stress in one of three possible words (58% for L, 63% for L+T, and 60% for L+A; the chance level was 33%), but again showed no significant improvements for aided lipreading over lipreading alone. These results suggest that factors other than suprasegmental cues may contribute towards the difference in performance observed between the auditory and tactual supplements for lipreading of sentences.

To investigate the role of segmental cues in the envelope signals, Bratakos, Reed, Delhorne and Denesvich (2001) studied the effects of the same envelope signal presented as a supplement to lipreading through either the auditory or tactual modality. The reception of segmental and sentence materials was studied in normal-hearing subjects under 5 conditions: lipreading alone (L), auditory supplement alone (A), tactual supplement alone (T), lipreading combined with the auditory supplement (L+A), and lipreading combined with the tactual supplement (L+T). Vowel and consonant identification was compared across the 5 conditions. Generally, performance was ordered as follows: L < L+T < L+A. Performance on various speech features was examined to

[1] L+T: lipreading supplemented by tactual cue; L+A: lipreading supplemented by auditory cue; L: lipreading alone.

determine the amount of information provided by the envelope signal. The largest improvements for the tactual and auditory supplements to lipreading of vowels were for the features tense and low (Table 4-1).

Table 4-1. Vowel feature performance. (From Bratakos et al., 2001).

	L	L+T	L+A
Low	65.2%	75%	80.8%
Tense	53%	68%	75.7

The largest improvements for the tactual and auditory supplements to lipreading of consonants were for the features voicing, nasality, and plosion (Table 4-2).

Table 4-2. Consonant feature performance. (From Bratakos et al., 2001).

	L	L+T	L+A
Voicing	12%	32%	42%
Nasality	21%	39%	78%
Plosion	34%	44%	67%

Apparently, the amplitude envelope does contain segmental information that can augment the information received through lipreading. Specifically, speech recognition performance is improved significantly for the information transmission of the features tense and low in the vowels and the features voicing, nasality and plosion in the consonants. In each of these cases, the benefits provided by the auditory supplement always exceeded those derived from the tactual supplement. Despite the benefits observed, there is still substantial room for improvement in voicing reception under both the auditory and tactual modalities.

4.2 Fundamental Frequency

Fundamental frequency or pitch is an important parameter in speech analysis, synthesis, and encoding. It is used in various applications such as (1) speaker identification and verification, (2) speech analysis and synthesis, (3) linguistic and phonetic knowledge acquisition, (4) voice disease diagnostics, and (5) speech encoding (see Kadambe and Boudreaux-Bartels, 1992).

Fundamental frequency is an acoustic manifestation of vocal-fold activity that provides both segmental and suprasegmental information. On the segmental level, it indicates the presence or absence of vocal-fold phonation, thus contributing to the discrimination of the feature voicing. For fricatives, one difference between voiced and voiceless cognates, such as /v/ and /f/, is that vocal-fold vibration is present throughout at least part of voiced fricatives and is absent in voiceless fricatives; for stops such as /b/ and /p/, the relevant cue is in the timing of the onset of vocal-fold activity relative to the opening of the lips. Therefore, segmental information is carried by the presence or absence of F0. Furthermore, F0 is higher in the following vowel for a given voiceless consonant than for its voiced counterpart.

On the suprasegmental level, the F0 contour that spreads over two or more segments plays an important role in signaling syntactic structure (such as questions versus statements, or location of stressed words in a sentence) and emotional attitude (such as anger or sadness). F0 information has particular significance in lipreading, because it is largely uncorrelated with visible oral correlates of place of articulation, and is hence complementary to the information contained in lipreading.

4.2.1 Extraction of Pitch Contour

Reliable estimation of the pitch contour is very difficult because it is affected by a variety of factors that include the physical characteristics of the vocal folds, the emotional state, and accent (the way the word is pronounced) of the speaker. Although a wide variety of pitch-estimation algorithms have been developed, none is capable of perfect estimation of pitch across a wide range of speech utterances and speakers. The existing algorithms can be broadly classified into two categories: event-detection and nonevent-detection (Kadambe and Boudreaux-Bartels, 1992).

In the event-detection method, pitch is estimated by locating the instant at which the glottis closes (called an event) and then measuring the time interval between two such events. There are several methods of detecting the event including autocovariance (Strube, 1974), epoch extraction (Ananthapadmanabha and Yegnanarayana, 1975, 1979), maximum-likelihood epoch determination (Cheng and O'Shaughnessy, 1989) and wavelet-transform methods (Kadambe and Boudreaux-Bartels, 1992).

Nonevent-detection methods estimate the pitch by estimating the average pitch over a segment of a speech signal with fixed length. For each segment the average pitch is achieved using one of the following methods: (1) compute the autocorrelation of the infinitely and centrally clipped signal (Sondhi, 1968); (2) compute the autocorrelation of an inverse filtered signal (Markel, 1972); (3) compute the cepstrum of a given segment of a signal (Noll, 1967); (4) compute the average magnitude difference function of a given signal (Ross, Shafer, Cohen, Frendberg, and Manley, 1974).

Each detector has its own advantages and disadvantages. In general, event-detection methods are more computationally intensive than nonevent-detection methods.

However, the nonevent algorithms are insensitive to nonstationary variations in the pitch period over the segment length.

4.2.2 F0 as an Auditory Supplement to Lipreading

Research on the improvements to lipreading provided by auditory supplements of F0 contour has examined word recognition in continuous speech of known topic (Grant, Ardell, Kuhl, & Sparks, 1985), recognition of words in sentences of known topic (Boothroyd, Hnath-Chisolm, Hanin, & Kishon-Rabin, 1988), and identification of phonological speech contrasts, both segmental and suprasegmental (Boothroyd, 1988).

Grant et al. (1985), using CDT (connected-discourse tracking), investigated the ability to combine speechreading with prosodic information extracted from the low-frequency regions of speech. Three normally hearing subjects were tested under 5 receptive conditions: speechreading alone (S), speechreading plus amplitude envelope cues (AM), speechreading plus fundamental frequency cues (FM), speechreading plus intensity-modulated fundamental frequency cues (AM+FM), and speechreading plus a voicing duration cue (DUR). Average performance in words/min tracking score were: 41.1 (SA), 73.7 (SA+AM), 73.6 (SA+FM), 83.6 (SA+AM+FM) and 65.4 (SA+DUR). Thus, equivalent benefits were provided by the AM and FM cues and the combination of these two cues led to an additional modest improvement.

Boothroyd et al. (1988) compared the effectiveness of three acoustic supplements to lipreading: 1) the low-pass-filtered output of an electroglottograph, 2) a variable-frequency, constant-amplitude sinusoidal substitute for voice fundamental frequency, and 3) a constant-frequency and constant-amplitude sinusoidal substitute that served as a

representation of voicing. Mean recognition of words in known-topic sentences across 12 normally hearing adults increased by between 30 and 35 percentage points with the electroglottograph signal and the variable-frequency sinusoidal F0 substitute, and the increase was greater for longer sentences. The constant-frequency and constant-amplitude sinusoidal substitute provided only a 13 percentage-point increase accounting for voicing detection alone.

Boothroyd (1988) also investigated the effects of audible F0 as a supplement to lipreading of minimal speech contrasts. For the vowel, consonant place, and final consonant continuance contrasts, performance was dominated by visual information from lipreading alone, and the addition of audible F0 had little effect. In contrast, for the suprasegmentals and final consonant voicing, perception was dominated by audible F0, and the addition of lipreading had little effect. For initial consonant voicing and continuance, however, there was positive interaction between lipreading and hearing. This positive interaction was attributed to the integration of voice onset time (perceived auditorily) with articulatory movements (perceived visually).

4.2.3 F0 as a Tactual Supplement to Lipreading

The investigations of lipreading supplemented by tactile F0 information have used procedures similar to those employed in auditory F0 supplement studies (Hnath-Chisolm and Kishon-Rabin, 1988; Hnath-Chosolm & Medwetsky, 1988; Hanin, Boothroyd & Hnath-Chisolm, 1988; Yeung, Boothroyd & Redmond, 1988; Waldstein & Boothroyd, 1995). Different tactual coding schemes for F0 contour have been used across studies, including: (1) temporal single-channel schemes in which input frequency is represented

as rate of vibration to a single vibrator (Hnath-Chisolm & Kishon-Rabin, 1988); (2) schemes in which F0 is coded as the vertical displacement of a small finger-rest, engaging kinesthetic/proprioceptive pathways (Waldstein & Boothroyd, 1995); and (3) spatial multichannel displays in which F0 is represented as location of vibration in a linear array of transducers (Yeung, Boothroyd & Redmond, 1988; Hnath-Chisolm & Medwetsky, 1988).

The results of these studies indicate that tactually transmitted information about F0 can be used to enhance lipreading performance, at least in tasks requiring the perception of speech pattern contrasts. However, the magnitude of the enhancement provided by a tactual supplement of F0 is substantially less than that achieved by auditory presentation of F0 (Boothroyd, 1988). Only performance on final consonant voicing was comparable under tactile and auditory presentation. Because this contrast is cued by durational differences, it suggests that the tactual and auditory senses are comparable for duration discrimination. For suprasegmental contrasts (i.e., roving stress and rise/fall) and other segmental contrasts (i.e., vowels, consonant place, initial voicing and continuance), performance with the tactual enhancement was significantly higher than for lipreading alone, but less than with the auditory enhancement.

The magnitude difference of the enhancement effect for tactual versus auditory presentation may arise from several reasons: (1) The relatively small amount of training and experience provided through the tactile sense compared with that of the well-established auditory sense. (2) The quantization of F0 contour inevitably introduced by tactual coding, resulting in reduced frequency resolution. In a separate study on the effect of quantization of an auditory F0 supplement to lipreading (Hnath-Chisolm &

Boothroyd, 1988), the results showed that as the number of steps in the quantized F0 contours increased from 1 to 12, the lipreading enhancement effect increased. (3) The inferior ability of touch in frequency discrimination and temporal sensitivity may also contribute to reduced performance.

It is worth noting that for the initial voicing contrast, the results varied across studies (Hnath-Chisolm and Kishon-Rabin, 1988; Kishon-Rabin, et al., 1996). The subjects of Hnath-Chisolm and Kishon-Rabin (1988) were able to successfully integrate visual and tactual information, thus perceiving the initial voicing contrast. In contrast, the subjects of Kishon-Rabin et al. (1996) were not able to integrate tactile and visual information with enough temporal precision to perceive differences of voice-onset-time, thus failing to perform well on the initial consonant voicing contrast. This difference may be due to the different materials and subjects involved in the two studies.

4.3 ASR Based Voicing Detection

Automatic speech recognition (ASR) can be used to aid communication of the deaf in several ways. A simple way is to provide orthographic displays of the recognized phones or words. However, this requires a high level of performance of the recognizer (above 90%), and good reading skill of the users. This level of recognizer performance is well in excess of the capabilities of currently existing systems. Also, young children and some adults may not possess sufficient reading skills to use such displays. Another approach is to present a small set of discrete symbols that are extracted from the acoustic speech signal via ASR as an aid to lipreading. Several visual aids have been studied

(Upton, 1968; Ebrahimi and Kunov, 1991; Duchnowski, Lum, Krause, Sexton, Brataos, and Braida, 2000).

Ebrahimi and Kunov (1991) used a 5×7 LED matrix to present three features including voice pitch, high-frequency energy (above 3000 Hz), and total energy of the speech signal. Eight young adults (3 normal hearing and 5 profoundly deaf) evaluated this device in a 12-consonant (/a/-C-/a/) context) identification task. Scores improved from 41% for lipreading alone to 76% for aided lipreading. In particular, the voicing distinction was improved from 55% with SA to 88% with the aid.

Duchnowski et al. (2000) produced a form of Cued Speech (Cornett, 1967) based on real-time automatic recognition. Cues were derived by a Hidden Markov Model (HMM)-based speaker-dependent phonetic speech recognizer that used context-dependent phone models. These cues were then presented visually by superimposing animated handshapes on the face of the talker. The benefit provided by these cues was found to be strongly dependent on articulation of hand movements and on precise synchronization of the actions of the hands and the face. Experienced cue receivers were able to recognize roughly two-thirds of the key words in cued low-context sentences compared to one-third by lipreading alone.

There has been some research aimed specifically at the use of automatic speech recognition for detection of voicing. Choi (1999) implemented algorithms for automatic detection of voicing and evaluated the performance of this recognition system.

4.3.1 Voicing Detection in Isolated Utterances

The initial stage of Choi's work was based on assessing voicing recognition performance on a set of measurements made manually from a corpus of isolated VCV and CVC utterances. The consonants /C/ were from the set of 16 obstruent consonants (/p, b, t, d, k, g, f, v, th, tx, ch, j, s, z, sh, zh/) and the vowels were either /a/ or /e/. These utterances were spoken once by two speakers, one male (ks) and one female (cb).

The final measurements used in this stage of the research were a set of measurements at points near the closure and release including H1 amplitude at +30 ms after closure and –30 ms before release, F0, H1-H2, H1-A1, H1-A3, and F1 frequency at –10 ms before voice offset (at closure) and at +10 ms after voice onset (at release) (Choi, 1999, p. 56). As noted before, these measurements were made manually. The mean of each measurement was calculated for voiced and unvoiced stops and fricatives (either at closure or release). A voicing decision was made for each measurement according to both manner (fricatives or stops) and landmark type (closure or release) by comparing the measurement with the mean and choosing the closest group. These voicing decisions for individual measurements were then consolidated into a single decision for the landmarks, and further consolidated into overall voicing decision for the segments.

Recognition rates ranged from 66 – 100% accuracy for various sets of training and test data. The percent-correct recognition of voicing of the landmark[2] for CVC are listed in Table 4-3.

[2] Landmarks are points in the speech signal where acoustic cues for these features are most evident. There are three types of landmarks: vowel, glide, and consonant.

Table 4-3. Percent correct recognition of voicing of the landmark for CVC syllables. (From Choi, 1999).

	ks (speaker 1)			cb (speaker 2)		
Training set	CVC/all	CVC/aa	CVC/eh	CVC/all	CVC/aa	CVC/eh
CVC/aa	92%	94%	91%	87.5%	84%	91%
CVC/eh	89%	84%	94%	83%	72%	94%
CVC/all	87.5%	84%	91%	82%	75%	87%

4.3.2 Voicing Detection in Continuous Speech

Detection of voicing in continuous speech was examined using a subset of the LAFF database. The entire database consists of 100 grammatical sentences from four speakers. It consists of approximately 200 one-to-three syllable words with few instances of consonant clusters. The entire database includes 758 obstruent consonants, of which 374 are lexically voiced and 384 are unvoiced. The subset of the database consists of the first ten sentences for speaker ks (male) and ss (female). There are 67 consonants in the subset, of which 25 are voiced and 42 are unvoiced.

The final measurements employed in this stage of the study (and obtained manually) were the same as those in the initial stage. The same classification procedure as in the initial stage was used to make a voicing decision in the continuous speech database.

For continuous speech, the error rates of voicing decision of the landmarks ranged from 25.9% (train: LAFF/1-5, ks; test: LAFF/6-10, ks) to 16.7% (train: LAFF/1-5, ks; test: LAFF/1-5, ks). When measurements from isolated syllables were used in training, the error rates for voicing detection of the landmarks in LAFF sentences for speaker ks

ranged from 24.5% for training with VCV/all utterances to roughly 43% for training with CVC/all utterances.

4.3.3 Automatic Detection of Consonant Voicing

In this stage, the measures were extracted using an automatic algorithm at times in the signal corresponding to consonant landmarks, with the landmarks located manually at times for closures, releases, and voice onset/offsets, as well as for events related to the manner of the underlying consonant. A larger subset of LAFF consisting of the first 30 sentences for speakers ks and ss were used in this stage. This larger set contains 228 consonants, of which 112 are voiced and 116 are unvoiced.

For isolated utterances, the results for training with the hand-made measurements led to a large error rate of roughly 30% for voicing decision of the landmarks. Training with the automatic measurements lowered the error rate to roughly 13%; however, this is larger than the error rates for training and testing on hand-made measurements of isolated utterances (less than 10%). The error rates of testing on the automatic measurements of the first 30 sentences of the LAFF database were roughly 40% for training on the manual measurements of the isolated utterances and 30% for training on the automatic measurements of the isolated utterances. Error rates of the voicing decision of segments for training and testing on the same automatic measurements of the first 30 sentences of the LAFF database were roughly 16% for ks and 14% for cb.

Chapter 5

Acoustic Analysis

The first step of the current study is to identify a reliable acoustic cue for consonant voicing that can be processed in real time, and thus can be synchronized with lipreading. Although a number of acoustic correlates of consonant voicing have been discussed in Chapter 4, none of them is satisfactory for the current purpose. In this chapter, a novel acoustic cue of consonant voicing is proposed based on speech-production theory. Acoustic measurements were obtained on a C_1VC_2 nonsense-syllable database. The predicted performance of an ideal observer using the acoustic measurements indicates it serves as an excellent acoustic cue of voicing for initial segmental consonants. The perceptual effectiveness of this cue is evaluated in a series of experiments described in Chapter 6.

5.1 Stimuli

5.1.1 Nonsense Syllable Database

The nonsense syllables were obtained from a corpus of audiovisual speech stimuli recorded at MIT over four sessions in 1987-1988. The speakers for these recordings were two females, SR and RK, both of whom were teachers of the deaf and were approximately 30 years of age at the time of the recordings. The materials used for the current study were taken from 1-inch U-matic videotapes recorded with a narrow angle

camera lens (the face occupied the top five-sixths of the vertical viewing area and nine-sixteenths of the horizontal viewing area). Two audio channels were recorded onto the tapes: (1) the acoustic speech signal through a Lavalier microphone and (2) a laryngograph signal. Only the acoustic speech signal was used in the current experiments.

The materials used in the current study were taken from two sets of C_1VC_2 syllables that were balanced for C_1. In these syllables, C_1 =/p, b, t, d, k, g, f, v, th[1], tx[2], s, z, sh[3], zh[4], ch[5], j[6], m, n, l, r, w, y, h, hw/, and C_2 is selected randomly from a set of 21 consonants: /p, t, k, b, d, g, f, th, s, sh, v, tx, z, zh, ch, j, m, n, ng, l, r/. One set of eight lists per talker was recorded in 3 different vowel contexts: /i/, /a/, and /u/; another set of two lists per talker was recorded in 16 different vowel contexts. Each list contains one representation of C_1V combination. Thus, each list in the first set contains 72 tokens (24 $C_1 \times 3$ V) and each list in the second set contains 384 tokens (24 $C_1 \times 16$ V).

5.1.2 Procedures for Digitizing Speech Materials

The C_1VC_2 nonsense syllables were digitized for real-time speech signal processing and for rapid-access control in the experiments. The audio-visual stimuli were originally available on videotapes. These analog recordings were digitized using the Pinnacle DV500 Plus system and then stored in individual files on the host computer. A block diagram of the system used in digitizing the signals is provided in Fig. 5-1. The resulting digitized files typically hold roughly 10 minutes of analog recordings. The

[1] th is voiceless "th" as in path or thumb.
[2] tx is voiced "th" as in that or those.
[3] sh as in ship.
[4] zh as in azure.
[5] ch as in church.
[6] j as in jump.

video was digitized in NTSC standard, with frame size 720 × 480, frame rate 29.97 frames per second, and pixel depth of 24 bits. The audio was initially sampled at a rate of 48000 Hz, with16-bit resolution.

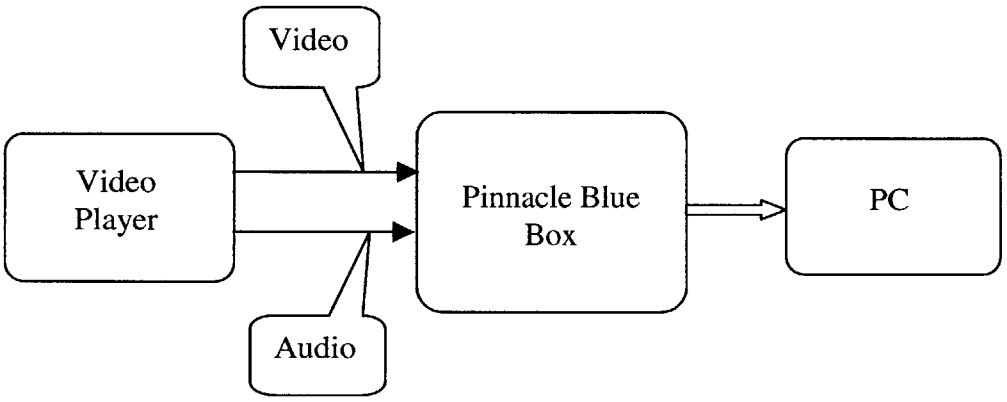

Fig. 5-1. Block diagram of the video digitizing system.

The resulting digitized files were then edited by the Adobe Premiere 6.0, a professional digital video-editing software package. The files were decomposed into a video track and an audio track and the two tracks were saved into separate files in the Timeline window of Adobe Premiere for separate processing. The audio was filtered at 11000 Hz by an anti-aliasing FIR filter to eliminate a 16-kHz noise (presumably an artifact of the recording process), and down-sampled at 22000 Hz. The sound level of each file was normalized to remove overall level differences between the two speakers as well as within the same speaker across different recording sessions. In the normalization, the sound level of each file was adjusted proportionally to maintain the same maximum sound level across files (i.e., the maximum sample value of each file was set to 0.95, and all other sample values were adjusted by the ratio 0.95/maximum sample value of the

given file). The filtering, sampling-rate conversion, and sound-level normalization of the digitized audio signals were carried out using MATLAB software. The video track and the processed audio track were combined into one file using the Timeline window of Adobe Premiere.

These processed files were then segmented into individual files each of which contained one token of the C_1VC_2 syllables. Segmentation was carried out using the video track and was based on lip opening and lip closure: the lips always move from a closed to an open state at the initiation of an utterance and from an open to a closed state at the end of an utterance. The following criteria were used to segment the utterances. The starting point of the file was specified at a location roughly 2 to 4 frames prior to the initiation of lip opening. The end point of the file was specified as roughly 2 to 4 frames following complete lip closure. The number of frames for each segment of the C_1VC_2 syllables averaged approximately 50 (or a duration of roughly 1.67 sec), and the average file size was approximately 6 Mbytes.

To reduce their size, the segmented files were compressed using the Sorenson Video compressor and converted to QuickTime format. The video was cut to 440 × 480 from the original 720 × 480, the audio was set to 22050 samples/sec (16-bit per sample), and the keyframe was set to every 60 frames. Following compression, the size of the C_1VC_2 syllables averaged roughly 0.43 Mbytes. The resulting files could be accessed and processed directly by the computer.

5.2 Acoustic Measurements

5.2.1 Envelope Extraction

Two envelopes (a low-frequency energy envelope and a high-frequency energy envelope) that are believed to carry voicing information were extracted from the acoustic speech signals, as shown in the block diagram of Fig. 5-2.

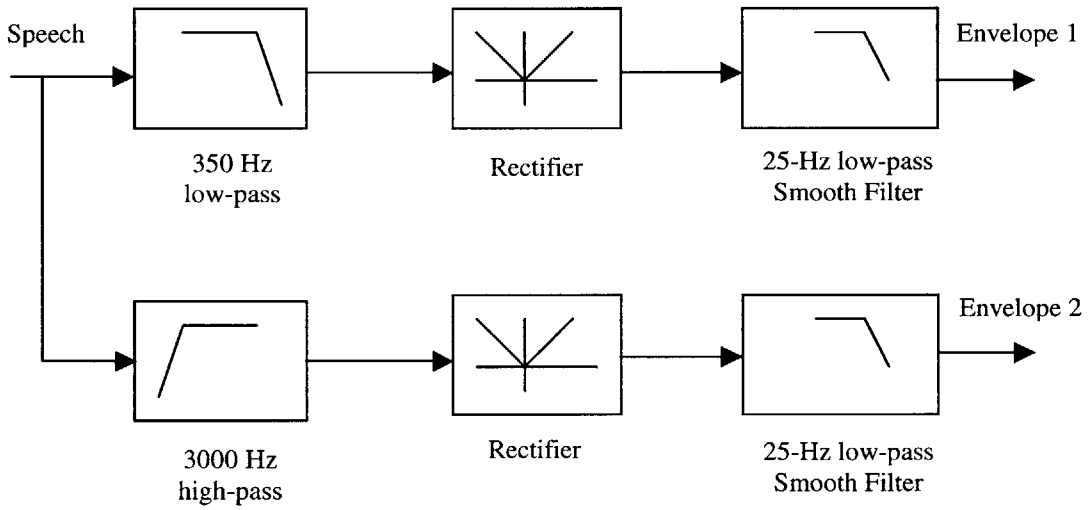

Fig. 5-2. Block diagram of envelope-extraction system.

The function of the 350-Hz low-pass filter and 3-kHz high-pass filter in Fig. 5-2 is to extract the speech signal in the frequency bands of interest. A second-order low-pass Butterworth filter with cutoff frequency 350-Hz is used for the low-frequency energy envelope, and a second-order high-pass Butterworth filter with cutoff frequency 3000 Hz is used for the high-energy envelope. Each filter is followed by rectification to flip the negative input to positive output to keep only the level information of the selected speech

spectrum. The rectified signals are then passed through a smoothing filter to reduce the fluctuations in the envelope. The smoothing filter in each channel is a second-order low-pass Butterworth filter, with cutoff frequency of 25 Hz (see Grant et al., 1991).

5.2.2 Measurement of Envelope-Onset Asynchrony (EOA)

Different patterns of spectral energy are observed for voiced versus voiceless consonants (regardless of their place or manner of production). The voiceless consonants tend to have significant energy above 3000 Hz, and less energy below 350 Hz. The voiced consonants, on the other hand, tend to exhibit significant energy below 350 Hz (at least during part of their production).

The low-frequency energy is associated with the presence of vocal-fold vibration, while the high-frequency energy is associated with aspiration or frication. The timing of the onset of high-frequency energy relative to low frequency energy tends to differ for voiced versus voiceless consonants. Typically, for initial voiceless consonants, the onset of the high-frequency energy occurs before the onset of the low-frequency energy. For initial voiced consonants, on the other hand, the onset of the high-frequency energy either follows or occurs almost simultaneously with the low-frequency energy. Therefore, the onset asynchrony of the two envelopes is expected to be a good indicator of the voicing distinction.

In order to measure the asynchrony of the two envelopes, a threshold must be selected to define the onset of the envelope. In general, there is less energy in the high-frequency band than in the low-frequency band, leading to a corresponding amplitude difference between the two envelopes. A threshold of 0.02 (on a scale from 0 to 1) was

selected for the low-frequency energy envelope, and 0.002 (on a scale from 0 to 1) for the high-frequency energy envelope. The onset time for each envelope is defined as the point in time at which the amplitude of the envelope first exceeds the threshold value. (It should be noted that this threshold value is not related to the tactual perceptual threshold).

Envelope-onset asynchrony (EOA) is defined as the difference in time between the onset of the high-frequency envelope and the onset of the low-frequency envelope: $EOA = OnsetTime_{LP} - OnsetTime_{HP}$. For the voiceless consonants, the onset of the high-frequency energy leads the onset of the low-frequency energy in most cases; therefore, the EOA of voiceless consonants is positive. On the contrary, for the voiced consonants, the onset of the low-frequency energy either leads or is nearly simultaneous with the onset of the high-frequency energy; therefore, the EOA of voiced consonants is negative or around zero.

An illustration of the measurement procedure is shown in Fig. 5-3 for two C_1VC_2 syllables: sheek (voiceless initial consonant /sh/) and zhatx (voiced initial consonant /zh/). For the syllable sheek (left panel of Fig. 5-3), the low-frequency energy envelope exceeds the threshold of 0.02 at roughly 0.325 sec (defining its onset). The high-frequency energy envelope exceeds the threshold of 0.002 at roughly 0.090 sec (defining its onset). Therefore, the EOA is EOA = 0.325-0.090 = 0.235 sec. Similarly, for the syllable zhatx (right panel of Fig. 5-3), the onsets of the low-frequency energy envelope and high-frequency energy envelope are 0.050 sec and 0.100 sec respectively. Thus, the EOA is EOA = 0.050-0.100 = -0.050 sec.

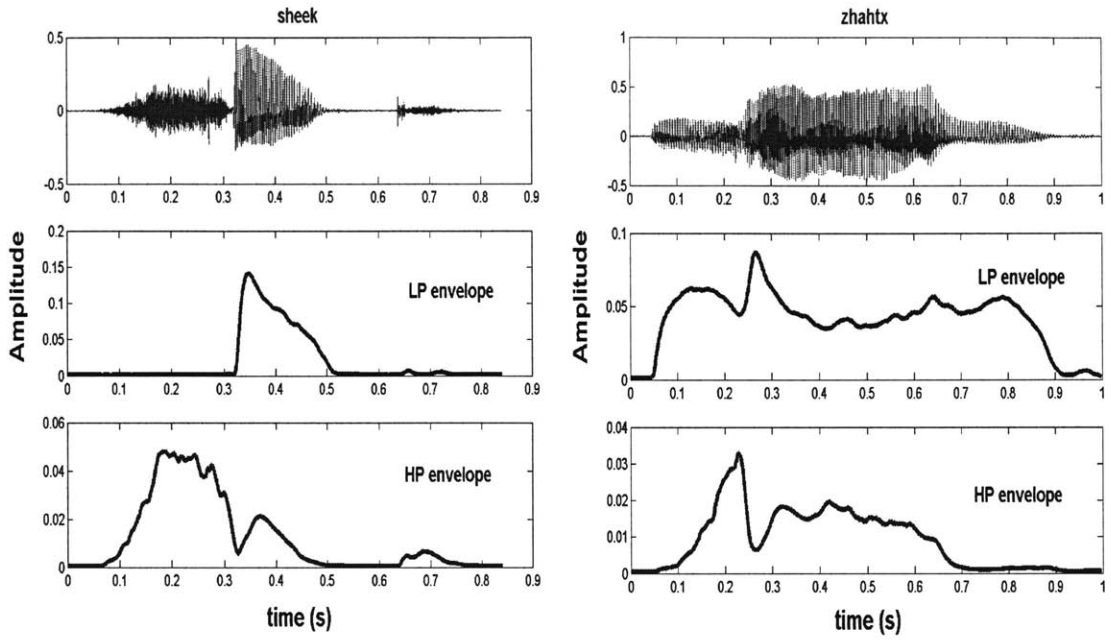

Fig. 5-3. Illustration of EOA measurements for two syllables. Upper trace is the original speech signal, middle trace is the low-pass band envelope, and lower trace is the high-pass band envelope.

5.2.3 Relation between EOA and VOT

Voice onset time (VOT) is the duration of the time between the release of a stop closure and the beginning of vocal-fold vibration. There are typically three types of VOT across languages (see Fig. 5-4):

1. short positive VOT: where the onset of the vocal-fold vibration coincides with the stop release (burst).
2. long positive VOT: where the onset of the vocal-fold vibration occurs after the stop release (burst).
3. negative VOT: where the onset of vocal-fold vibration precedes the stop release or burst, referred to as prevoicing.

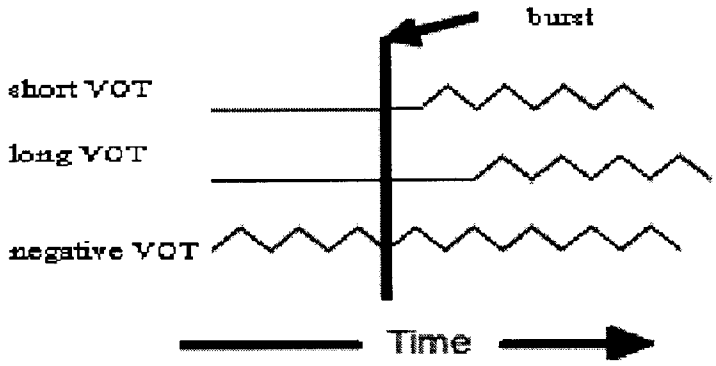

Fig. 5-4. Illustration of three types of VOT.

In English, VOTs for the voiced stops are in general less than 20 ms or even negative, and greater than this value for voiceless stops (Zue, 1976; Kiefte, 2003). It should be noted that this definition is valid only for initial stops. This acoustic property of VOT is regarded as a phonetic anchor associated with the feature voicing. It serves as a measure to categorize voicing in stop consonants: stops with long VOT are classified as voiceless, while those with short or negative VOT as voiced.

EOA and VOT for English stops are closely related to each other in two aspects: 1) The burst and the following frication and aspiration of a voiceless stop are composed of high-frequency energy; and 2) The vocal-fold vibration is usually composed of low-frequency energy. Although the energy onsets of the high- and low- frequency bands selected for envelope processing do not coincide with the timing of the burst and the onset of vocal-fold vibration respectively, they are nonetheless good indicators of the two events in production.

Real-time measurement of VOT is relatively difficult. It requires detecting the burst that corresponds to the release of the stops, as well as the onset of glottal vibration from the speech waveform. Burst detection is difficult, and even in cases where the burst can be reliably detected, its short duration poses an obstacle to perception through touch. On the contrary, the real-time measurement of EOA is quite straightforward (see Fig. 5-2). The envelopes of the two different frequency bands can be derived automatically in real-time, and their onset asynchrony can be used naturally as a cue to voicing.

Furthermore, EOA is applicable to English stop consonants, affricates and fricatives, whereas VOT is a concept applicable only to English stop consonants. The most salient acoustic property associated with voicing in fricatives is the presence of vocal-fold activity during the fricative noise interval. In particular, 20-30 ms of glottal excitation present during the fricative interval at either the boundary of the consonant implosion or at the release in consonant seems to be a critical attribute. This measure accurately classified approximately 83% of voiced and voiceless fricatives in the singleton intervocalic position and in intervocalic clusters matched for voicing, but was less successful in categorizing utterance-final fricatives and fricative clusters with mixed voicing (Stevens et al., 1992).

VOT can be affected by speech rate: the magnitude of the VOT is longest for isolated words, intermediate for slow speaking rate, and shortest for fast speaking rate. A similar effect is expected on the EOA.

5.3 Distribution of the EOA Values

The EOA was measured for each of the 112 tokens (8 lists × 2 speakers × 3 vowels + 2 lists × 2 speakers × 16 vowels) of each initial consonant from the two sets of recordings. The measurement of EOA was made automatically by a computer program.

The distribution of the EOA values for each consonant was derived by dividing the total range of the EOA within a given consonant or across consonants into equal duration intervals. The number of tokens with EOA within each given interval was then tabulated. Finally, the proportion of occurrences within each interval was obtained by dividing the number of counts in each interval bin by the total number of tokens for each consonant. Since the range of EOA of the voiceless stops is larger than that of the voiced stops, a bin size of 10 msec was selected for displaying the EOA of the voiceless consonants, and 5 msec for the voiced consonants (to obtain sufficient sample points).

5.3.1 3-Vowel Stimulus Set

The distributions of EOA for 8 pairs of voiced-voiceless contrasts are shown in Figs. 5-5 and 5-6 for measurements from both speakers with three vowels (/i /, /a /, /u /). Each consonant is represented by 48 tokens. Fig. 5-5 shows EOA distributions for the three stop contrasts (/p-b/, /t-d/, /k-g/) and for the affricate contrast (/ch-j/). Fig. 5-6 shows EOA distributions for the four fricative contrasts (/f-v/, /th-tx/, /s-z/, /sh-zh/). Two observations may be made from these plots. First, there appears to be very little overlap in the EOA values for voiced and voiceless contrasts. Second, there appears to be greater variability associated with the EOA value of voiceless compared to voiced consonants.

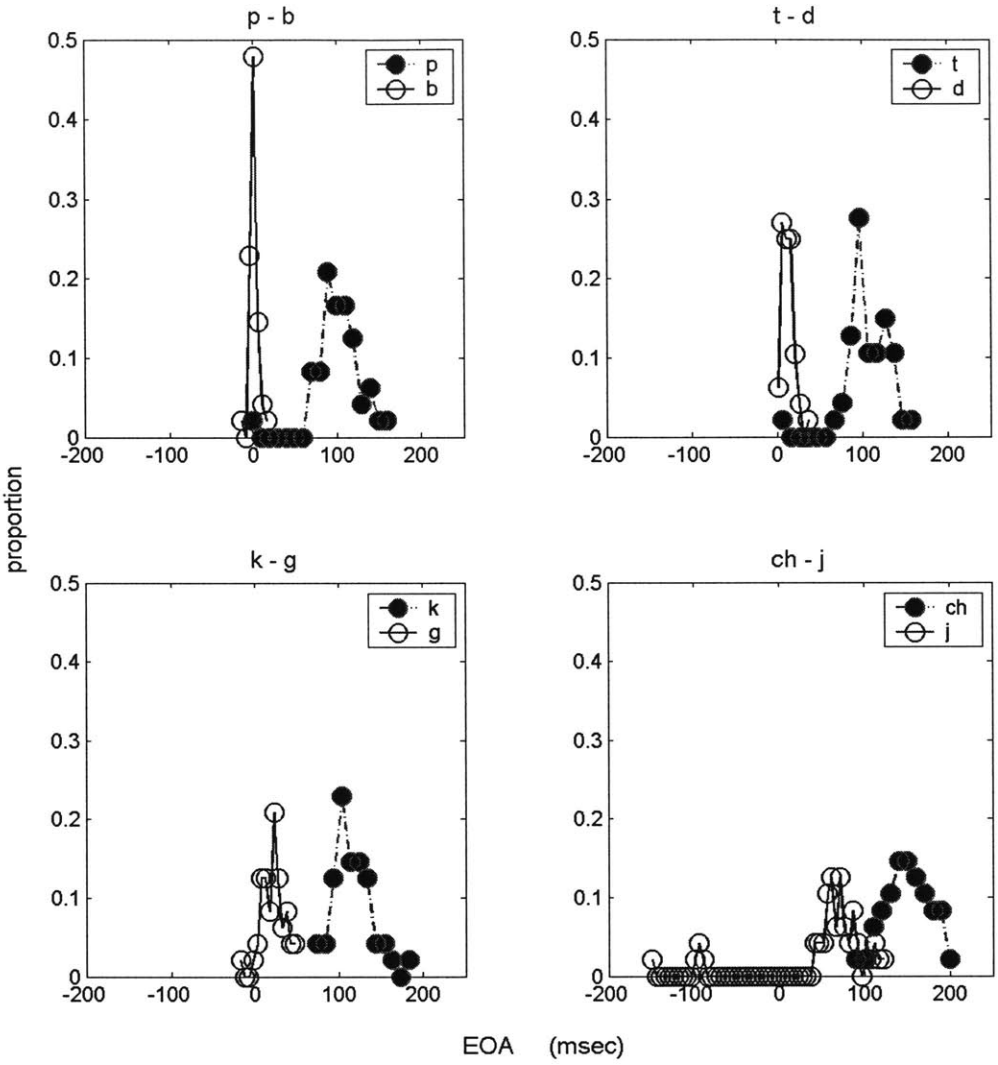

Fig. 5-5. EOA probability distributions of the stops and affricates in the 3-vowel stimulus set.

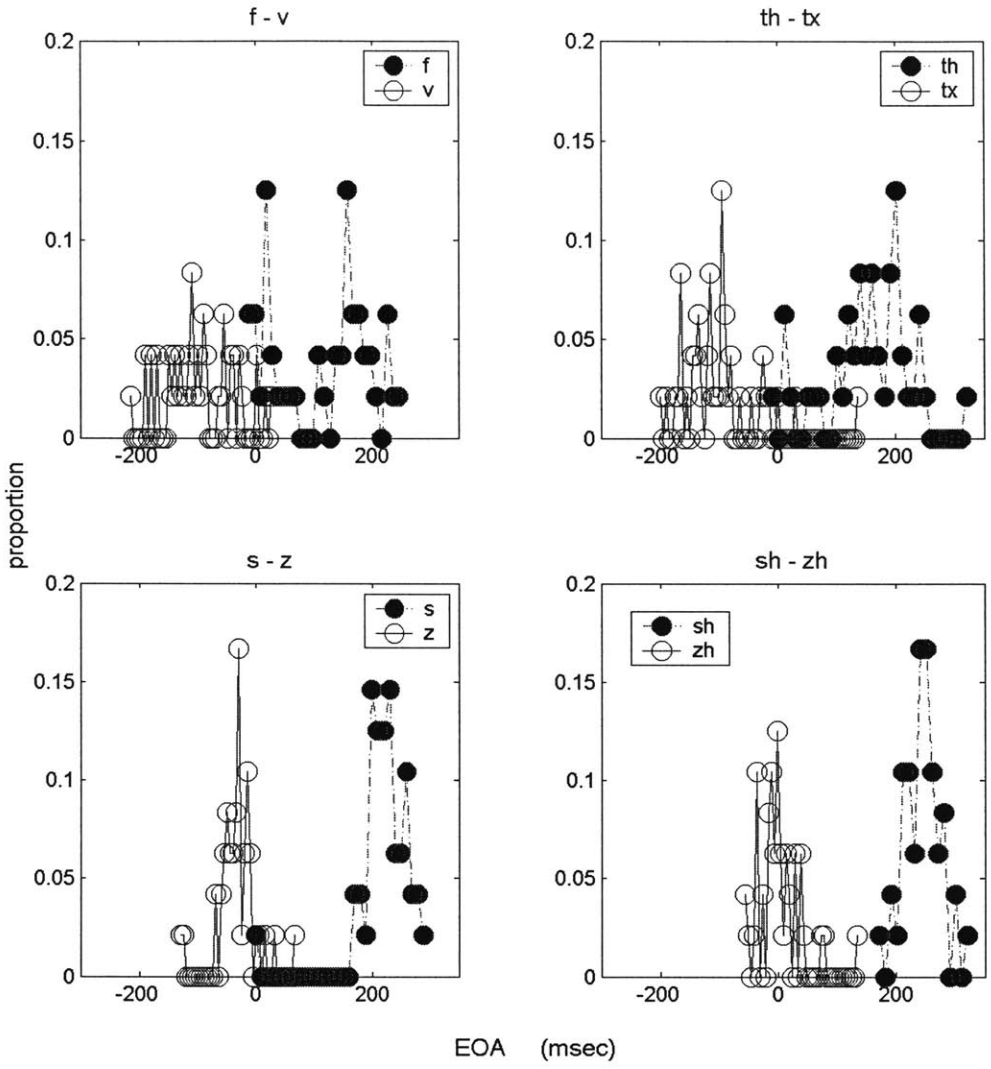

Fig. 5-6. EOA probability distributions of the fricatives in the 3-vowel stimulus set.

5.3.2 16-Vowel Stimulus Set

The distributions of EOA for the 8 pairs of voiced-voiceless contrasts were also determined for productions of each consonant with 16 vowels by two speakers. These distributions, based on 64 tokens of each initial consonant, are shown in Fig. 5-7 (for stop and affricate contrasts) and in Fig. 5-8 (for fricative contrasts). These distributions for

utterances with 16 vowels are generally similar to those shown previously in Figs. 5-5 and 5-6 for three vowel contexts, suggesting that the timing cue examined here is not heavily dependent on vowel context.

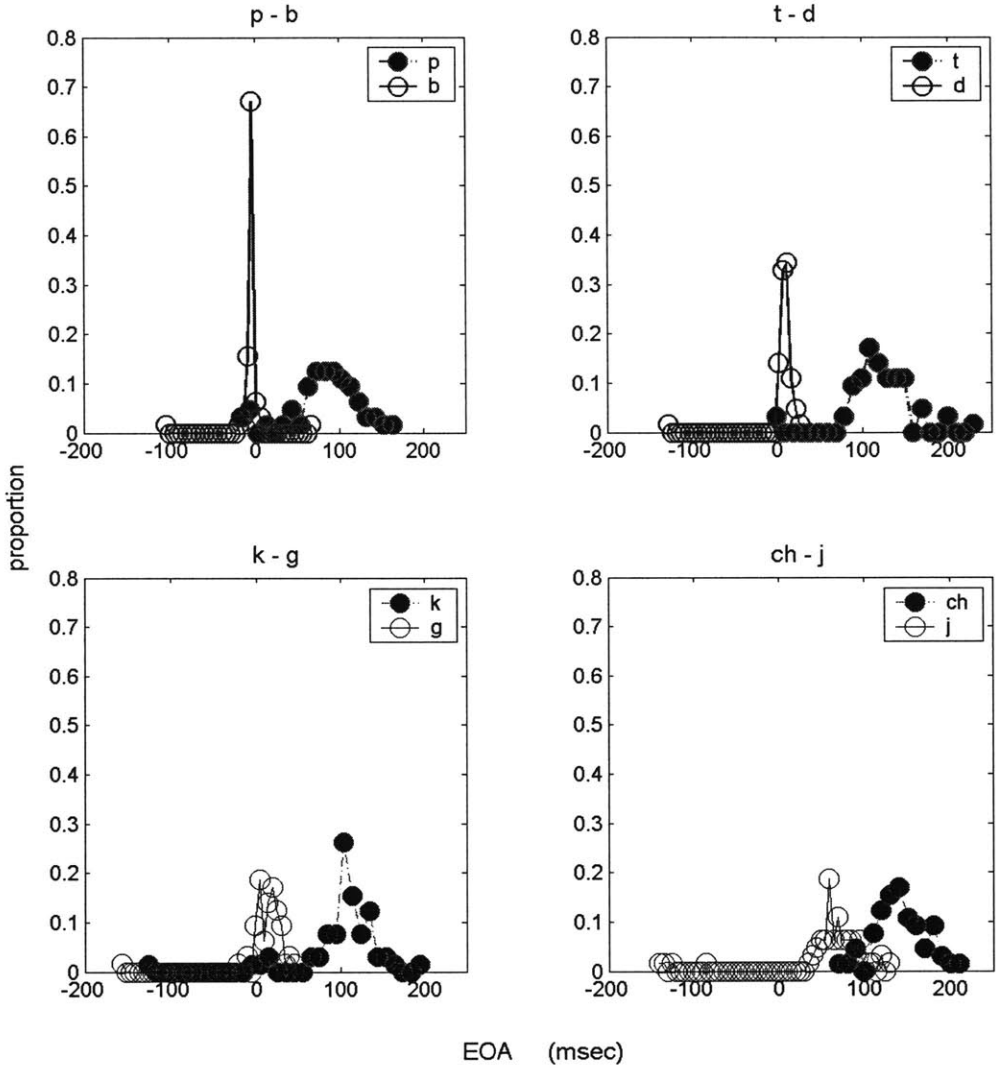

Fig. 5-7. EOA probability distributions of the stops and affricates in the 16-vowel stimulus set.

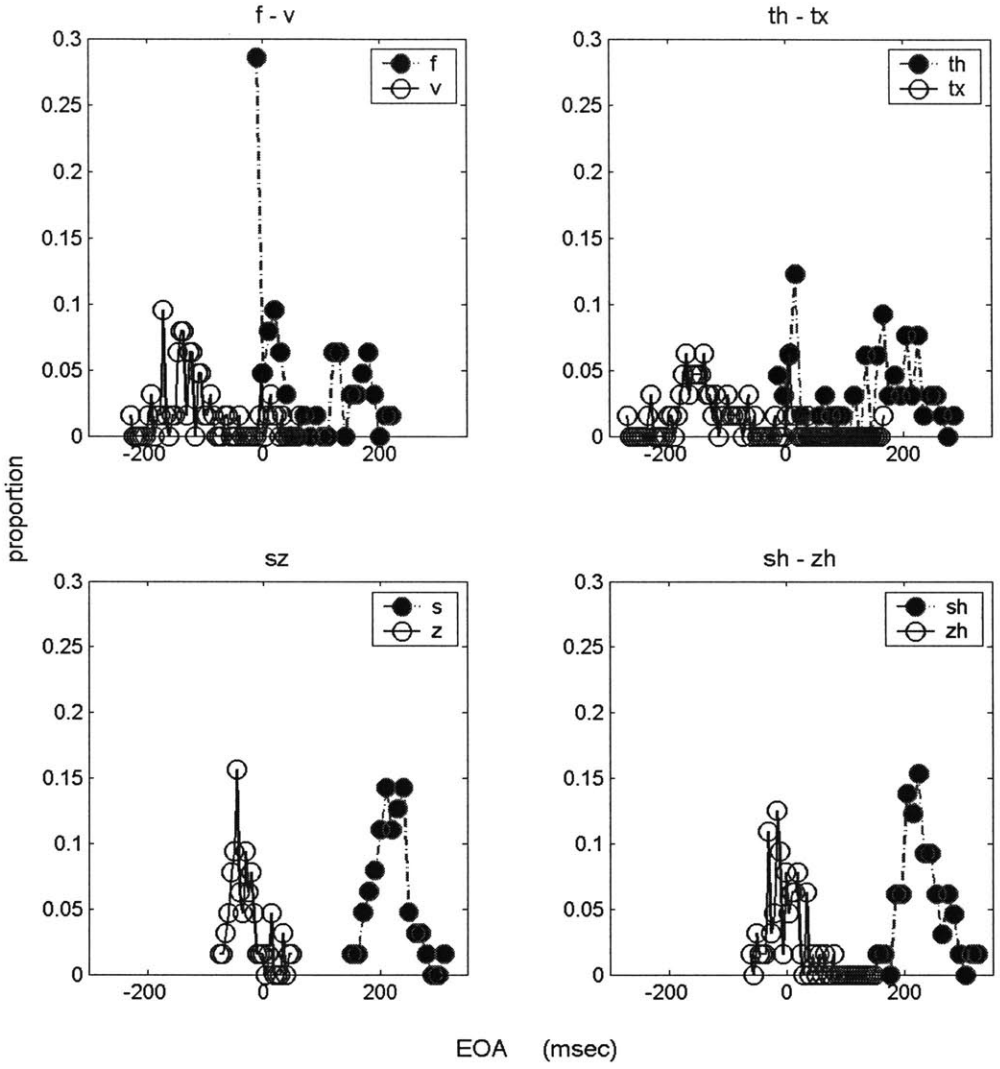

Fig. 5-8. EOA probability distributions of the fricatives in the 16-vowel stimulus set.

The distribution of all EOA values for two major categories of "voiced" and "voiceless" is shown in Fig. 5-9. The eight voiced consonants are grouped together for the "voiced" category, and the eight voiceless consonants are grouped together for the "voiceless" category. All 16 vowels and both speakers are represented, leading to a total of 512 ($16 \times 4 \times 8$) tokens in each category. The data plotted in Fig. 5-9 indicate that

EOA patterns are stable across different manners and places of consonant production and in a variety of vowel contexts.

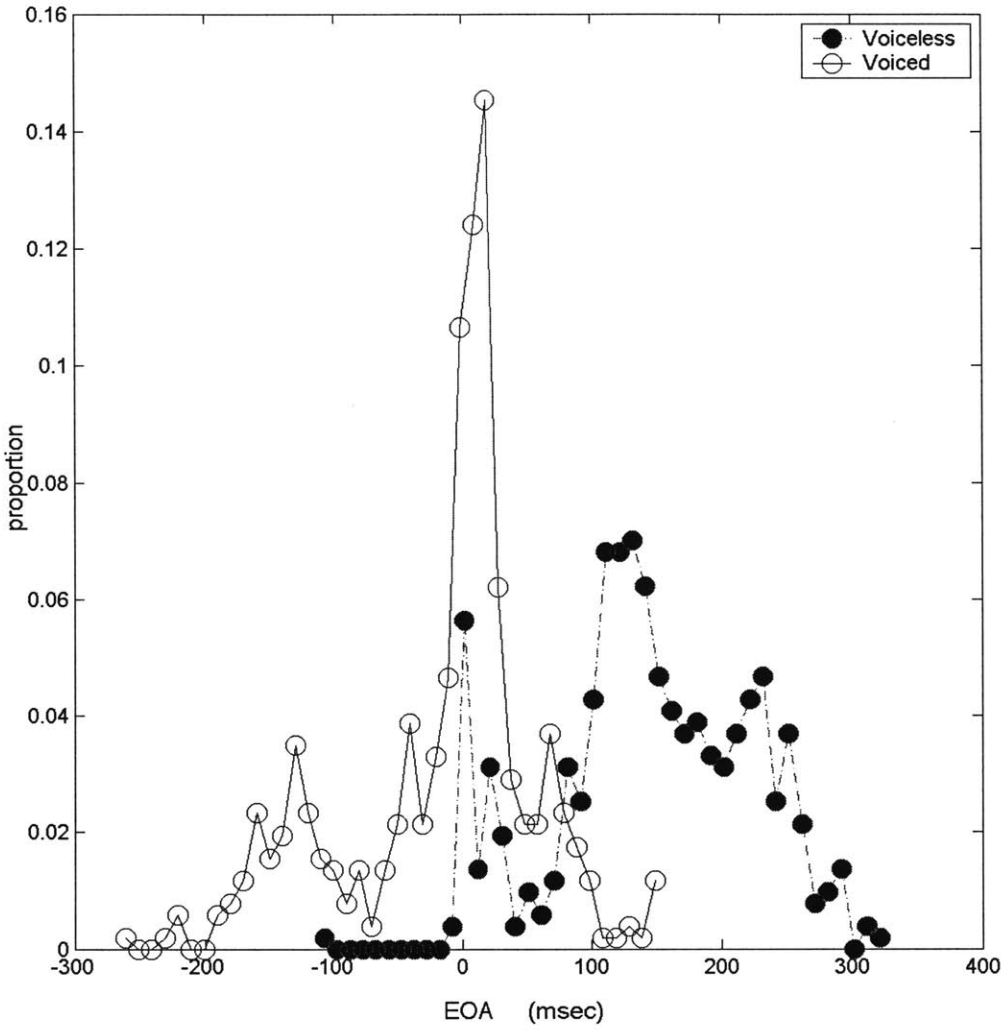

Fig. 5-9. EOA distributions of all voiceless consonants versus all voiced consonants in the 16-vowel stimulus set.

5.4 Cumulative Distributions of EOA Values

5.4.1 3-Vowel Stimulus Set

The cumulative distributions of EOA were derived from the EOA distributions by summing the proportions below a given criterion. The cumulative distribution functions of the 8 pairs (3 vowels, 2 speakers) are shown in Figs. 5-10 and 5-11.

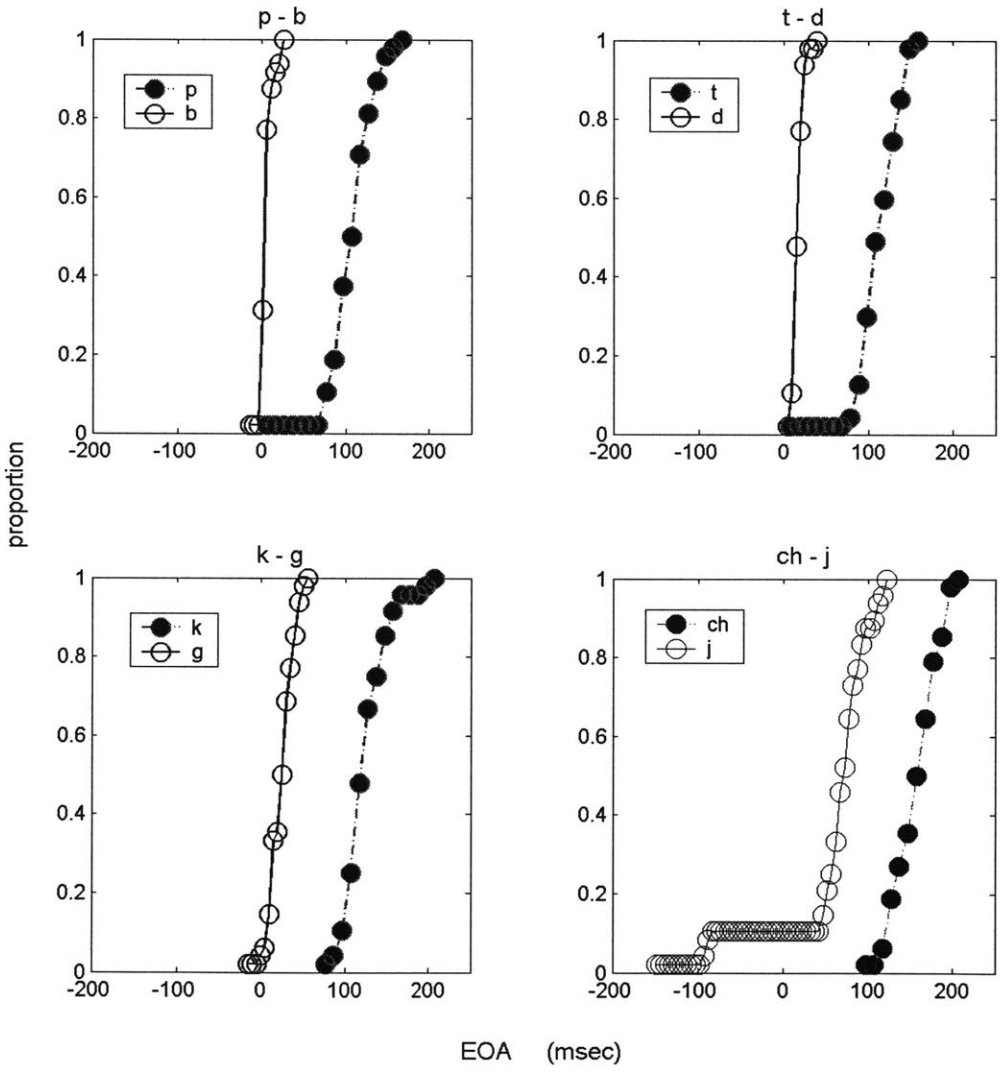

Fig. 5-10. EOA cumulative distribution functions of the stops and affricates in the 3-vowel stimulus set.

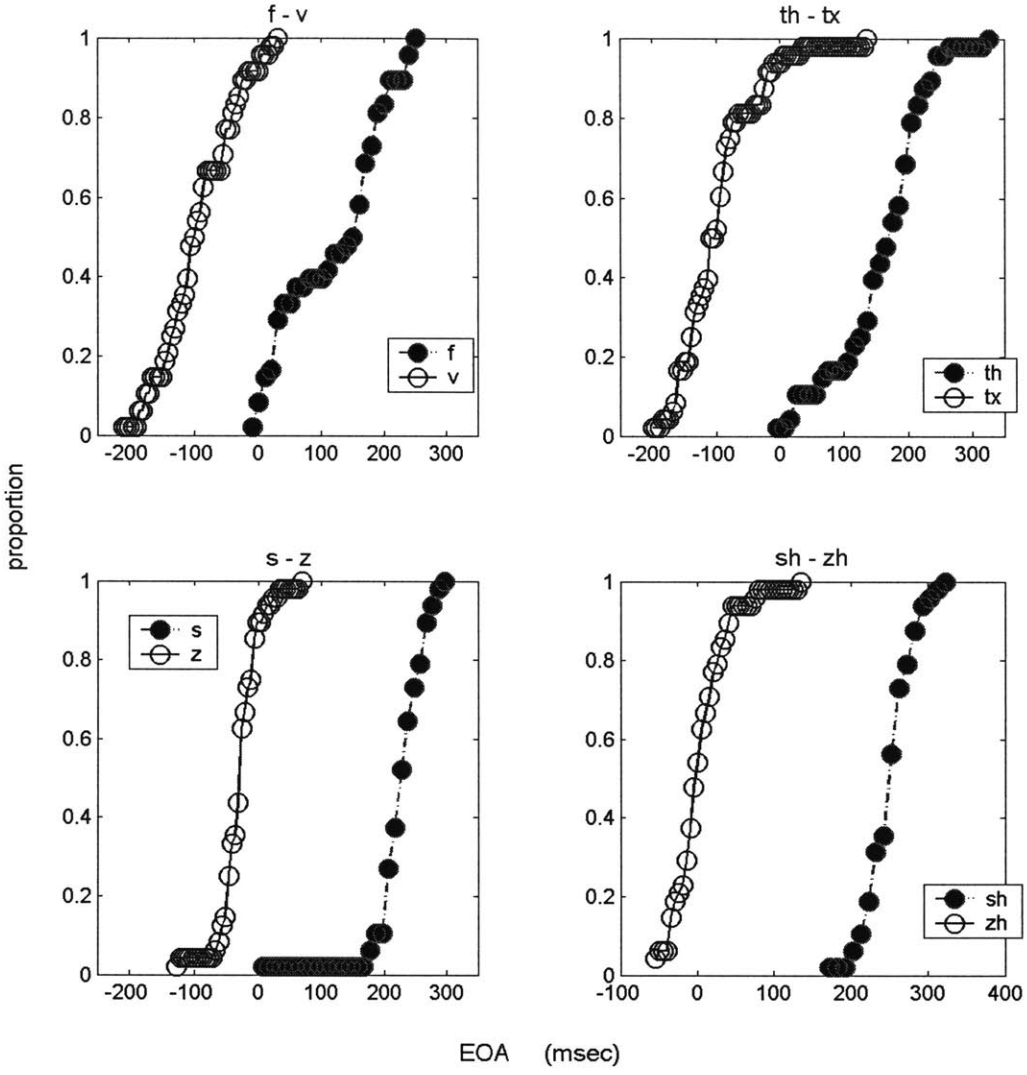

Fig. 5-11. EOA cumulative distribution functions of the fricatives in the 3-vowel stimulus set.

5.4.2 16-Vowel Stimulus Set

The cumulative distribution functions of the 8 pairs for the 16-vowel stimulus set for both speakers are shown in Figs. 5-12 and 5-13.

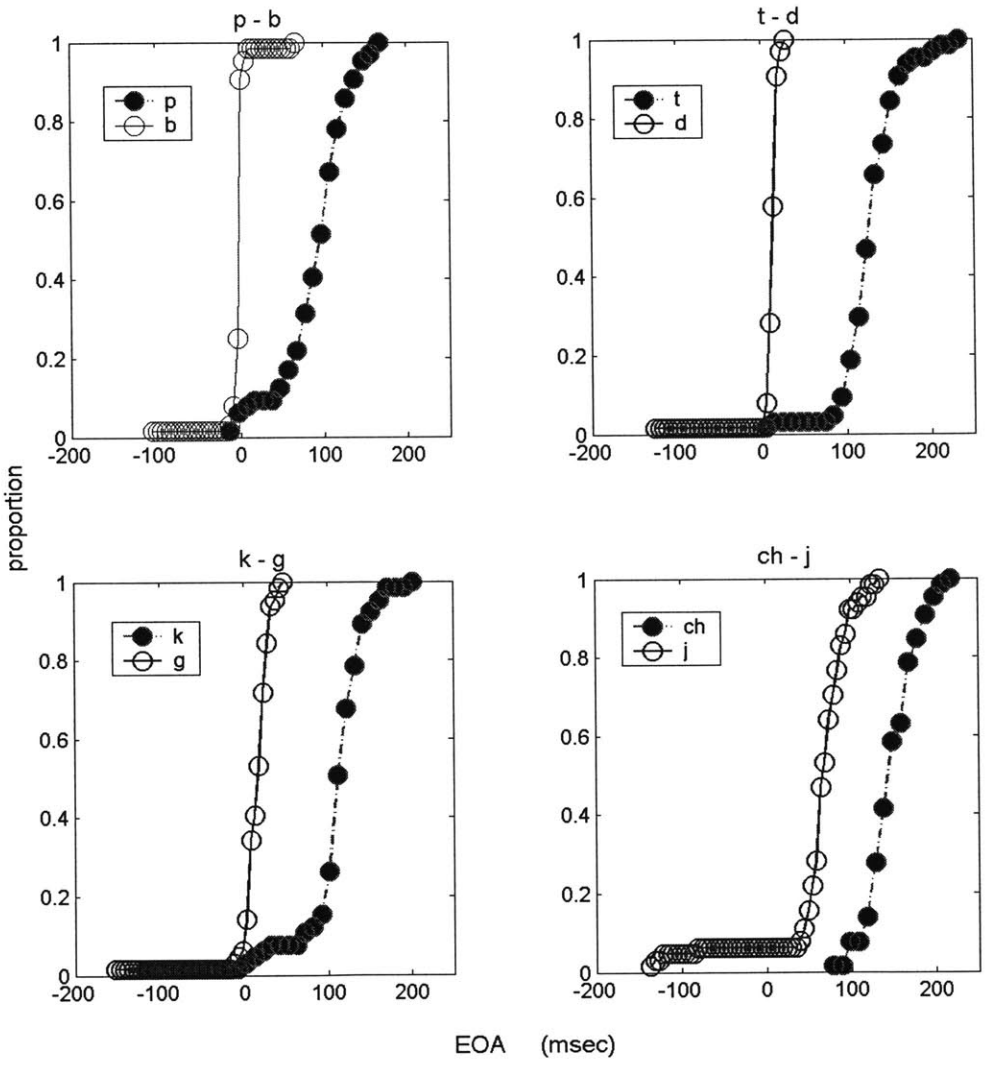

Fig. 5-12. EOA cumulative distribution functions of the stops and affricatives in the 16-vowel stimulus set.

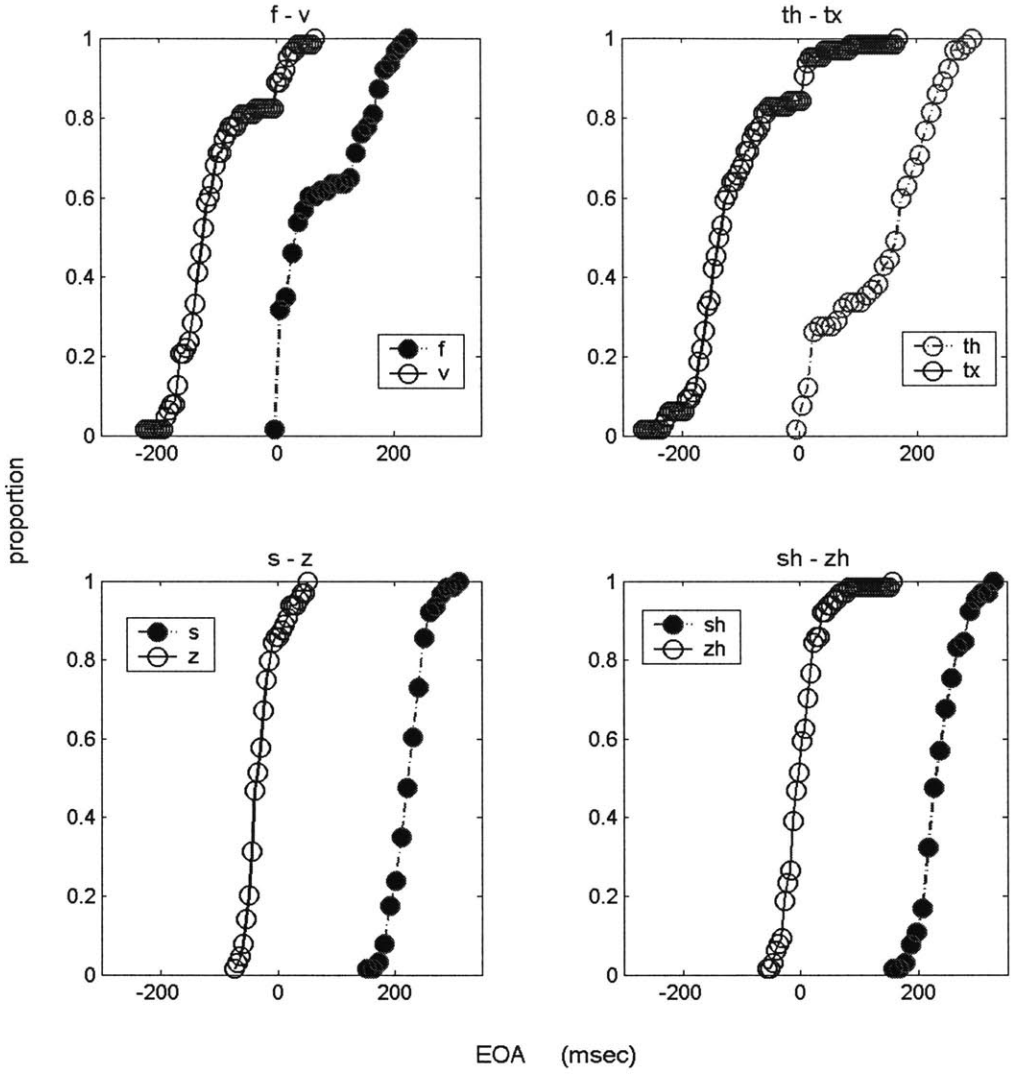

Fig. 5-13. EOA cumulative distribution functions of the fricatives in the 16-vowel stimulus set.

The cumulative distribution functions (cdf) for all voiceless versus voiced tokens from the 16-vowel stimulus set for both speakers are shown in Fig. 5-14. The cumulative distribution functions, both for individual pairs of voicing contrasts and pooled across all consonants, are quite separable for voiced compared to voiceless consonants.

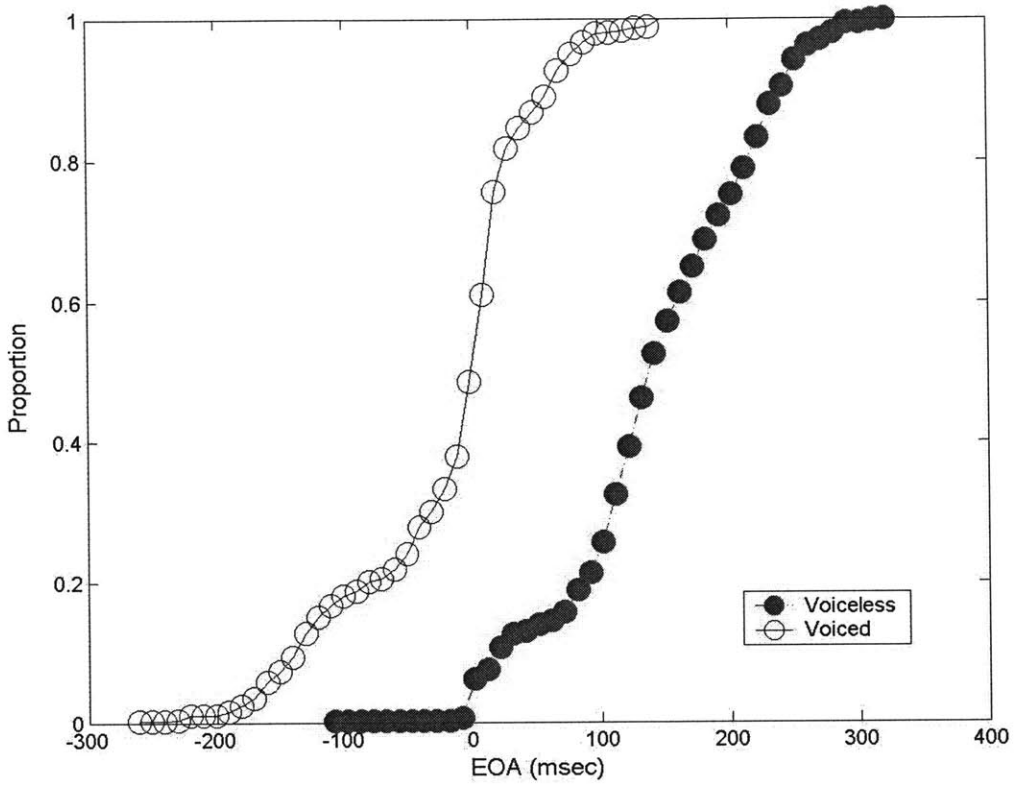

Fig. 5-14. EOA cumulative distribution functions of all tokens in the 16-vowel stimulus set.

5.5 Performance of an Ideal Observer

Using the EOA as a perceptual-distance measurement, we can calculate the performance of ideal observers in making the voiced-voiceless distinction with sensitivity (d′) and bias (β), two measures widely used in Signal Detection Theory (see Green and Swets, 1966; Durlach, 1968; Macmillan, 1990). Signal detection theory analyzes decision-making in the presence of uncertainty. Its application in sensory experiments provides a way to separate the sensory process of the stimuli from the subjective decision process. It offers measures of performance that are not specific to procedure and that are

independent of motivation. The measure d′ is basically determined by the underlying sensory process, and the measure β is basically determined by the subject's use of the response categories. The calculation of d′ and β in a one-interval two-alternative forced-choice paradigm derived from the stimulus-response confusion matrix shown in Table 5-1, is given by:

$$d' = z(H) - z(F)$$

$$\beta = 0.5 \times (z(H) + z(F))$$

where the hit rate: $H = N(Hits)/(N(Hits) + N(Misses))$, is the proportion of presentations of stimulus S1 to which the subject responds S1, the false-alarm rate: $F = N(False\ alarms)/(N(False\ alarms) + N(Correct\ rejections))$, is the proportion of presentations of stimulus S2 to which the subject responds S1, and z is the inverse of the normal distribution function that converts a hit or false-alarm rate to a z-score.

Table 5-1. Confusion matrix derived from one-interval two-alternative paradigm experiment.

		Response	
		R1	**R2**
Stimulus	**S1**	*N(Hits)*	*N(Misses)*
	S2	*N(False alarms)*	*N(Correct rejections)*

The calculation of d′ and β above inherently assumes that the decision variables (EOA) have Gaussian distributions of equal variances or variables that can be transformed into Gaussian distributions of equal variances. Thus, the first step towards computing d′ requires fitting the cumulative distributions by Gaussian distributions.

5.5.1 Gaussian Fitting of the Cumulative Distributions of EOA Values

5.5.1.1 Gaussian Fitting Procedure

The objective of Gaussian fitting is to find the best mean and standard deviation to approximate a Gaussian distribution using the cumulative distributions at the sampling points. The fitting procedure was implemented using the following steps:

1. Observing the cumulative distributions of EOA and selecting a range of values over which the mean and standard deviation are likely to occur.
2. Changing the mean and standard deviation by an amount Δ (the minimum increment of mean and standard deviation is 0.1 msec).
3. Calculating the error that is defined as the sum of squares of the mismatch between the observed cdf and the expected cdf at the sample points, for each pair of values of mean and standard deviation.
4. Choosing the Gaussian distribution that minimizes the error.

5.5.1.2 3-Vowel Stimulus Set

The Gaussian fittings for 8 pairs with 3 vowels are shown in Figs. 5-15 and 5-16. Below each figure, a table lists the mean and standard deviation of the best Gaussian fit, as well as the error (%) of the fit for each consonant (Tables 5-2 and 5-3). Error (%) is calculated as square root of the ratio between the sum of the squares of the mismatch and the sum of the squares of the cdf.

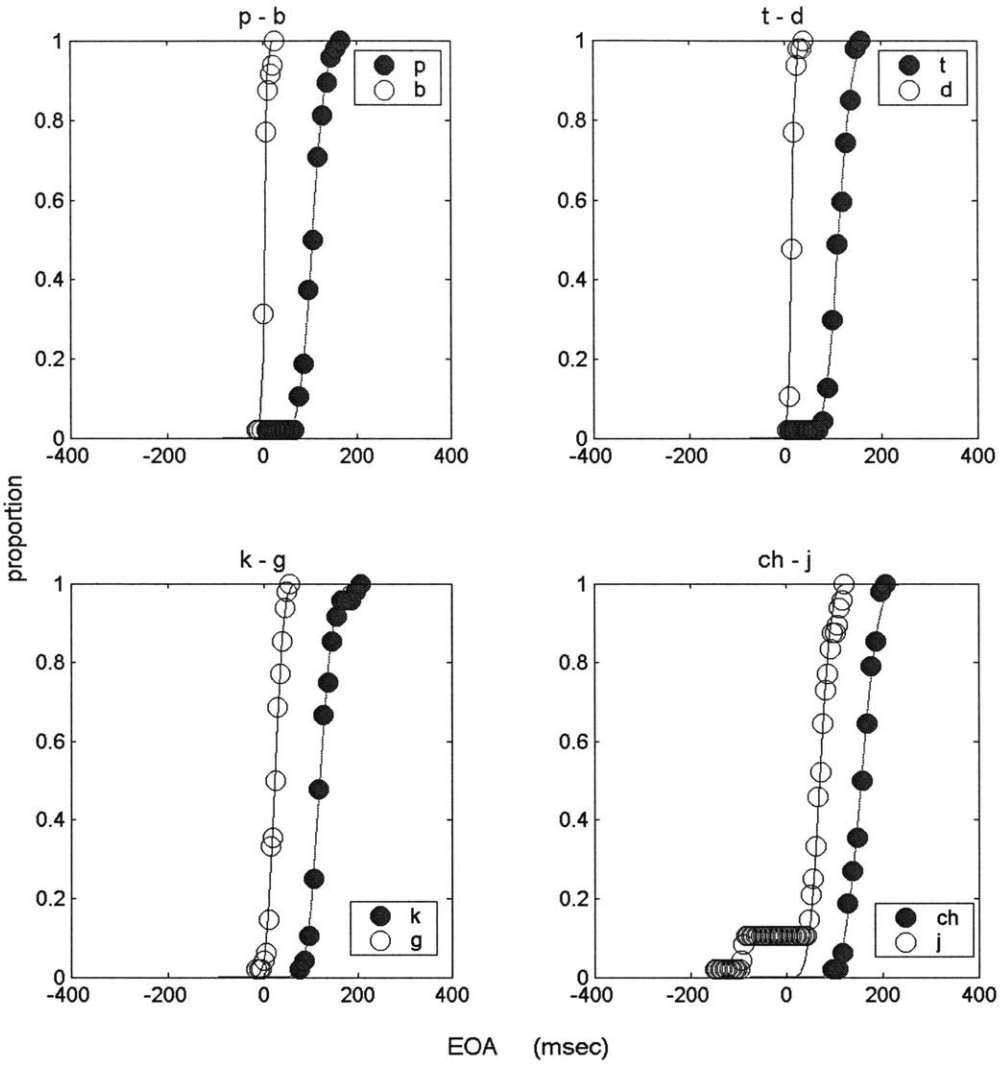

Fig. 5-15. Gaussian fitting of the EOA cdf of the stops and affricates in the 3-vowel stimulus set.

Table 5-2. Gaussian approximation of the EOA cdf of stops and affricates in the 3-vowel stimulus set.

	p	b	t	d	k	g	ch	j
Mean (msec)	105.8	3.7	111.8	15.5	120.5	24.4	156.1	70
s.d. (msec)	23	4.6	22.9	5.4	22.1	14.2	27.3	25.5
Error (%)	3.02	7.05	4.93	2.93	4.21	4.36	4.12	17.57

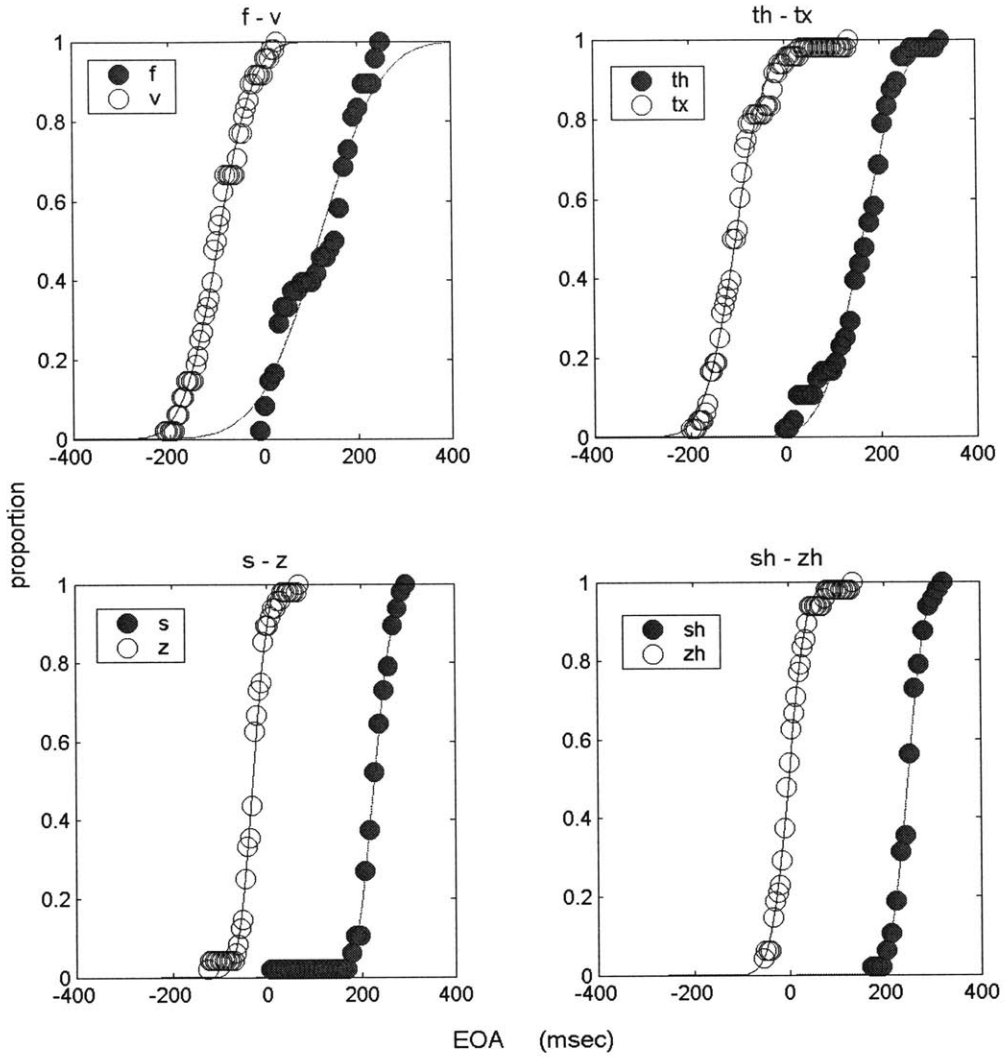

Fig. 5-16. Gaussian fitting of the EOA cdf of the fricatives in the 3-vowel stimulus set.

Table 5-3. Gaussian approximation of the EOA cdf of fricatives in the 3-vowel stimulus set.

	f	v	th	tx	s	z	sh	zh
Mean (msec)	117.8	-95.5	161.8	-103.6	227.9	-29.8	248.9	-0.6
s.d. (msec)	102.6	59.6	66.6	52.8	31.2	24.2	28.7	30.8
Error(%)	11.71	3.88	6.6	4.58	4.82	4.48	3.38	3.21

5.5.1.3 16-Vowel Stimulus Set

The Gaussian fittings for 8 pairs with 16 vowels are shown in Figs. 5-17 and 5-18. The mean and standard deviation of the best Gaussian fit, as well as the error (%) of the fit for each consonant are listed in Table 5-4 to 5-5.

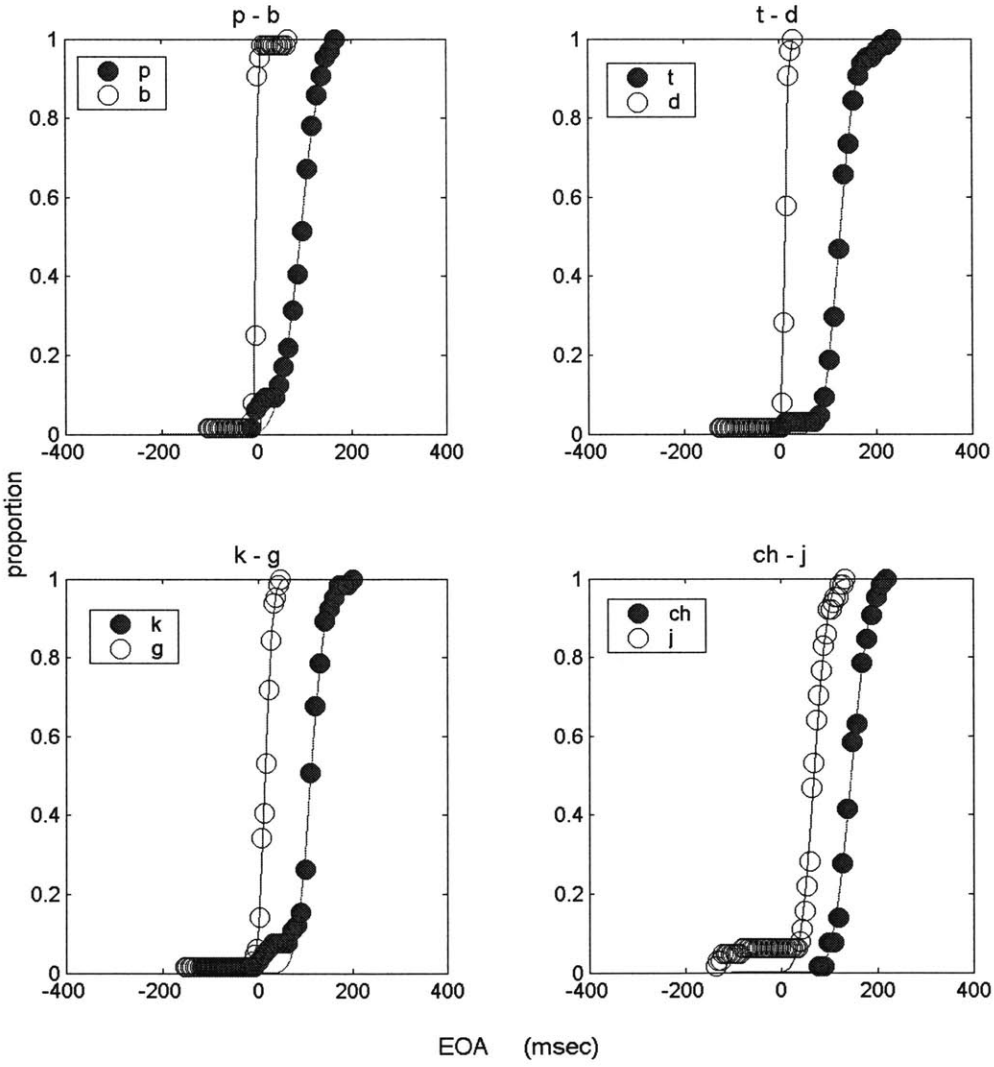

Fig. 5-17. Gaussian fitting of the cdf of the stops and affricates in the 16 vowel stimulus set.

Table 5-4. Gaussian approximation of the EOA of stops and affricates in the 16-vowel stimulus set.

	p	b	t	d	k	g	ch	j
Mean (msec)	91.9	-2	126.1	12.8	112.8	15.5	146	68.6
s.d. (msec)	36.2	2.6	26.8	5.7	24.9	12.9	29.6	21.9
Error (%)	6.41	3.41	3.77	5.08	7.34	5.23	3.5	10.26

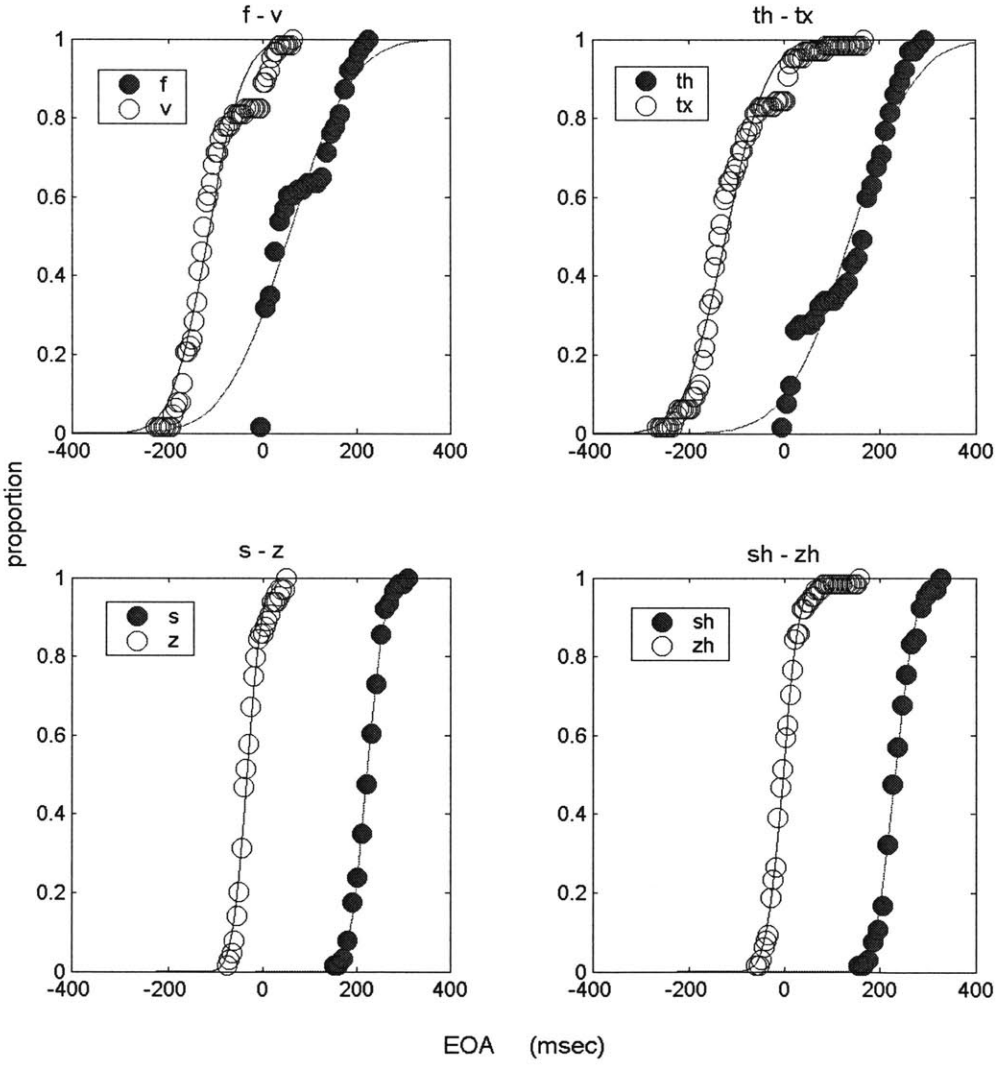

Fig. 5-18. Gaussian fitting of the EOA cdf of the fricatives in the 16-vowel stimulus set.

Table 5-5. Gaussian approximation of the EOA cdf of fricatives in the 16-vowel stimulus set.

	f	v	th	tx	s	z	sh	zh
Mean (msec)	54.6	-116.8	140	-122.4	222.4	-33.3	233.5	-2.6
s.d.(msec)	107.5	60.6	106.8	69.8	29.8	23.7	33.8	27.7
Error(%)	11.44	9.87	11.14	7.02	1.8	6.3	4.01	2.75

From these Figures and Tables, it is obvious that the variances for each pair are not equal. For example, in Table 5-2, the standard deviations associated with the voiceless stops /p, t, k/ are similar (22-23 msec) and substantially greater than that of their voiced counterparts /b, d, g/ whose standard deviations ranged from 4.6-14.2 msec. Except for the pair /f-v/, differences in standard deviations are not as pronounced for the fricatives and affricates as for the stops (see Tables 5-3 and 5-5).

5.5.1.4 Goodness-of-Fit Testing

The goodness-of-fit of each normal distribution to the empirical distribution of each consonant was evaluated using the Kolmogorov-Smirnov (KS) test (DeGroot and Schervish, 2002). The steps for KS testing are summarized as follows: 1) calculate the cumulative frequency (normalized by the sample size) of the observations as a function of class (i.e., for each individual phoneme), 2) calculate the cumulative frequency for the distribution under the null hypothesis (normal distribution with specified mean and standard deviation estimated from the empirical data using least square error), 3) calculate the D-statistic which is the maximum difference between the observed and expected cumulative frequencies, and 4) compare this D-statistic against the criterion for that sample size at a certain level of significance. If the calculated D-statistic is greater

than the criterion, then reject the null hypothesis that the distribution is of the expected form.

The D-statistic values for each consonant are shown in Table 5-6 for the 3-vowel stimulus set and in Table 5-7 for the 16-vowel stimulus set.

Table 5-6. D-statistic values for each consonant in the 3-vowel stimulus set.

3V	p	b	t	d	k	g	ch	j
D-statistic	0.027	0.0806	0.0474	0.0376	0.0521	0.0727	0.0443	0.1042
3V	f	v	th	tx	s	z	sh	zh
D-statistic	0.1257	0.0634	0.0835	0.0892	0.0614	0.0684	0.0627	0.0521

Table 5-7. D-statistic values for each consonant in the 16-vowel stimulus set.

16V	p	b	t	d	k	g	ch	j
D-statistic	0.0748	0.0728	0.0487	0.0307	0.0763	0.066	0.0515	0.0625
16V	f	v	th	tx	s	z	sh	zh
D-statistic	0.2738	0.1421	0.1225	0.1212	0.0275	0.0768	0.0552	0.0449

For a significance level of 0.01, the model will not be rejected if the D-statistic is less than $\frac{1.63}{\sqrt{N}}$, where N is the sample size. Thus, in order not to reject the model, the D-statistic must satisfy $D < \frac{1.63}{\sqrt{48}} = 0.2353$ for the 3-vowel stimulus set with sample size 48, and satisfy $D < \frac{1.63}{\sqrt{64}} = 0.2037$ for the 16-vowel stimulus set with sample size 64. When these two criteria are compared with the D-statistics shown in Tables 5-6 and 5-7, it can be seen that the normal distributions with mean and standard deviation estimated using least-mean-square error are not rejected at a significance level of 0.01 for all cases except the distribution for /f/ in the 16-vowel stimulus set.

The Gaussian fittings for all consonants in the 16-vowel stimulus set are shown in Fig. 5-19 and summarized in Table 5-8. The mean EOA for voiceless consonants is 142.5 msec compared to -12.4 msec for voiced consonants and standard deviations are similar. The function for voiced consonants is less well fit than that for voiceless consonants.

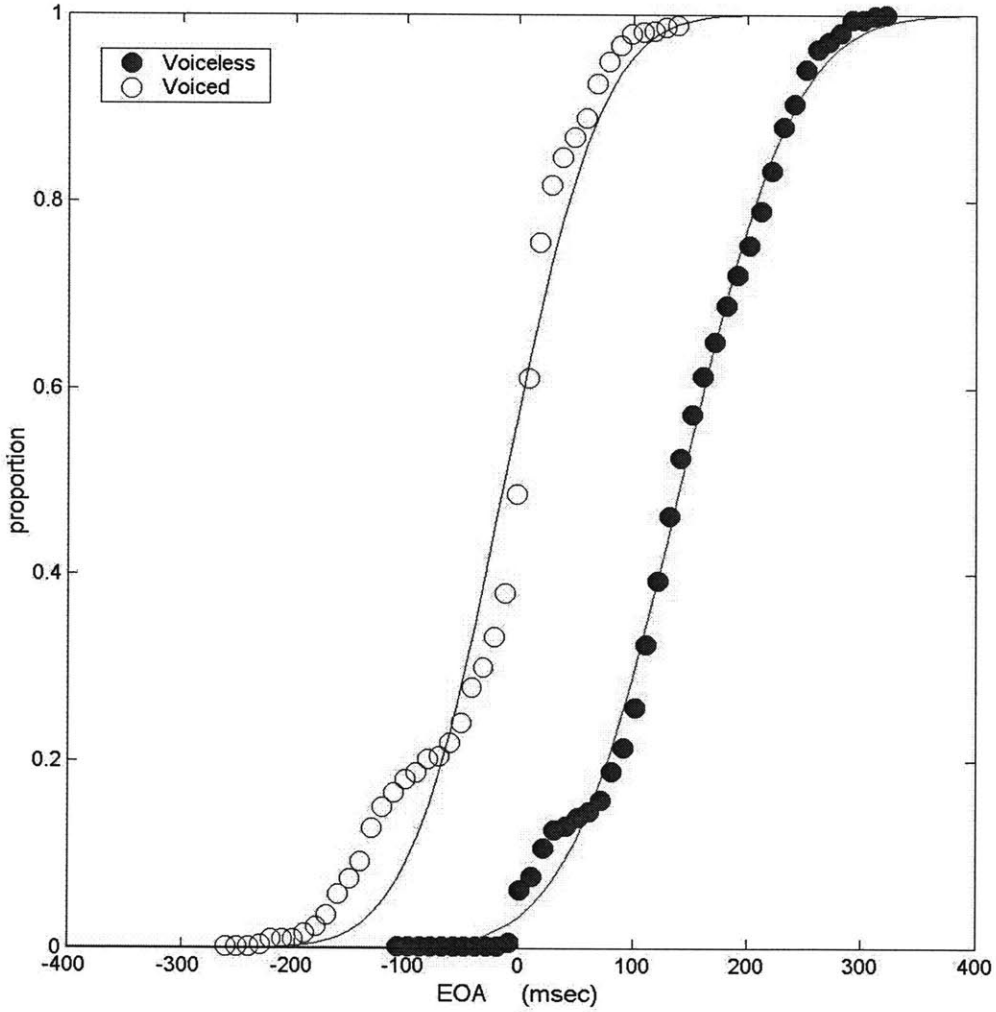

Fig. 5-19. Gaussian fitting of the EOA cdf of all voiceless consonants and voiced consonants in the 16-vowel stimulus set.

Table 5-8. Gaussian approximation of the EOA cdf of all voiceless consonants and voiced consonants in the 16-vowel stimulus set.

	Mean (msec)	s.d. (msec)	Error rate (%)
Voiceless	142.5	77.2	3.79
Voiced	-12.4	66.5	9.78

5.5.2 Computation of d′ for One-Interval, Two-Alternative Forced-Choice (1I, 2AFC) Experiment

From the above plots, it is found that the two distributions have different variances. To find d′ from an ROC that is linear in z-coordinates, it is easiest to first estimate the vertical intercept (d'_2), the horizontal intercept (d'_1), and the slope of the ROC, $s = d'_2 / d'_1$. And finally, d′ can be calculated by $d' = d'_2 / (0.5 * (1 + s^2))^{0.5}$ (Macmillan, 1990, p. 70).

The ROC plots for the pair /t-d/ in three-vowel context and in 16-vowel context are shown in Fig. 5-20. The d′ calculations for these two cases are 5.94 and 5.62, respectively, indicating high discriminability by the "ideal" observer.

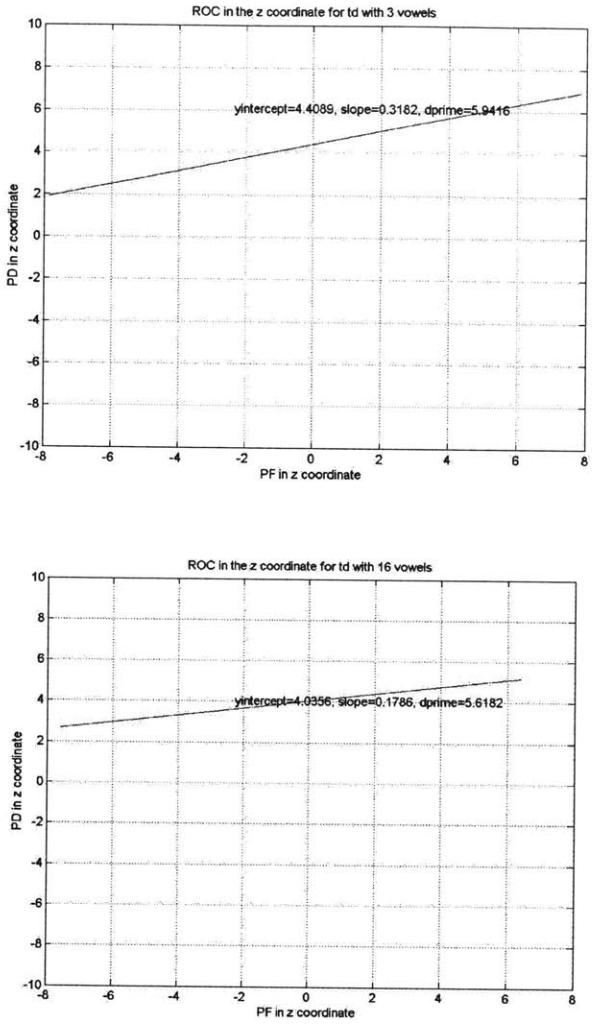

Fig. 5-20. ROC plot for the pair /t-d/ in the 3-vowel stimulus set (upper panel) and 16-vowel stimulus set (lower panel).

5.5.3 Computation of d′ for Two-Interval Two-Alternative Forced-Choice (2I, 2 AFC) Experiment

The decision variable in a 2I, 2AFC experiment is formed from the subtraction of EOA of first interval from that of the second interval. By randomly selecting one voiced token /d/, and randomly selecting one voiceless token /t/, we can calculate the difference between the two EOAs for the pair /t-d/. Repeating this for many trials, we can obtain the

distribution of the subtraction of the EOAs of /d/ and /t/, as well as the cumulative distributions which are shown in Figs. 5-21 and 5-22 (for three-vowel contexts and for 16-vowel contexts, respectively). The Gaussian fitting is the same as described above. The best fitting mean is –0.096, and variance is 0.024. In the two-interval experiment, the decision variable for the two possible intervals is symmetric with the origin. Therefore, $d' = 2 * mean / std$. For this case, d′ for the "ideal" observer is calculated to be 8.0 (for 3 vowels) and 8.1 (for 16 vowels).

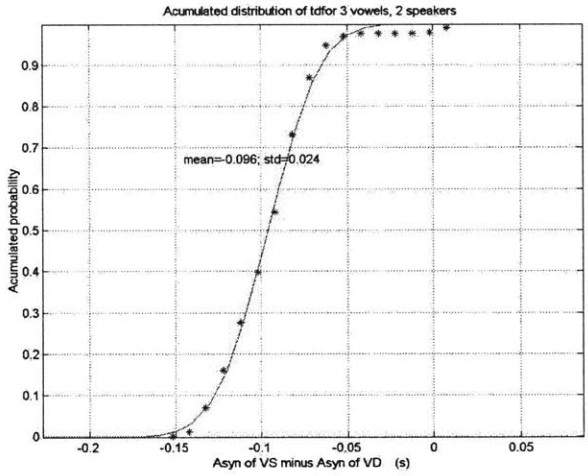

Fig. 5-21. Cumulative distribution of EOA for /d/ minus EOA for /t/ in the 3-vowel stimulus set.

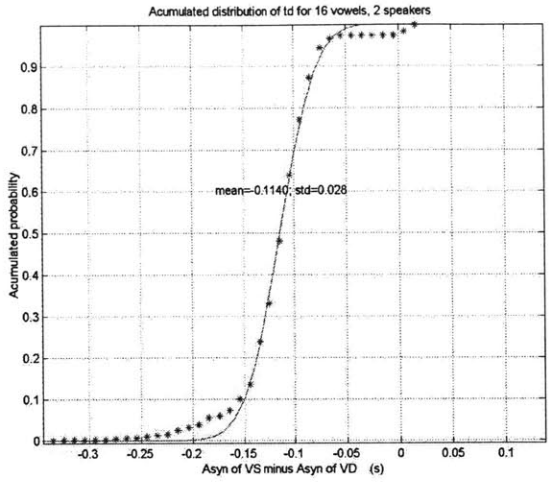

Fig. 5-22. Cumulative distribution of EOA for /d/ minus EOA for /t/ in the 16-vowel stimulus set.

The d′ values computed for each of the eight pairs in 2I, 2AFC for each of the two stimulus sets (3 vowels or 16 vowels) are provided in Table 5-9.

Table 5-9. Values of d′ for eight pairs in 2I, 2AFC experiment in the 3-vowel stimulus set and 16-vowel stimulus set.

d′	/p-b/	/t-d/	/k-g/	/ch-j/	/f-v/	/th-tx/	/s-z/	/sh-zh/
2I,2AFC (3vowels)	9.37	8.03	7.65	4.07	3.99	6.31	12.87	11.96
2I,2AFC (16vowels)	5.08	8.08	6.84	3.89	3.48	4.00	12.99	10.6

An alternative method of calculating the performance of the ideal observer using ROC curves is shown in Appendix A.

Chapter 6

Methods

6.1 Tactual Stimulating Device

6.1.1 General Description of Tan's (1996) System

The tactual stimulating device used in the experiments (referred to as the Tactuator) was initially developed by Tan (1996) for research with multidimensional tactual stimulation. The device is a three-finger display capable of presenting a broad range of tactual movement to the passive human fingers. It consists of three mutually perpendicular rods that interface with the thumb, index finger and middle finger in a manner that allows for a natural hand configuration (see Fig. 6-1). Each rod is driven independently by an actuator that is a head-positioning motor from a Maxtor hard-disk drive. The position of the rod is controlled by an external voltage source to the actuator. The real position of the rod is sensed by an angular position sensor that is attached to the moving part of each of the three motor assemblies. The rods are capable of moving the fingers in an outward (extension) and inward (flexion) direction.

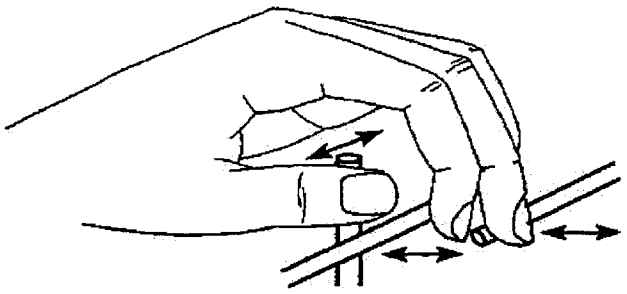

Fig. 6-1. Schematic drawing illustrating finger placement on the Tactuator. (Taken from Tan, 1996).

The overall performance of the device is well-suited for psychophysical studies of the tactual sensory system. First, the device is capable of delivering frequencies along a continuum from dc to 300 Hz, allowing for stimulation in the kinesthetic (low-frequency) and cutaneous (high-frequency regions), as well as in the mid-frequency range. Second, the range of motion provided by the display for each digit is roughly 26 mm. This range allows delivery of stimulation at levels from threshold to approximately 50 dB SL throughout the frequency range from dc to 300 Hz. Third, each channel is highly linear, with low harmonic distortion, and negligible inter-channel crosstalk. Fourth, loading (resulting from resting a finger lightly on the actuator's moving bar) does not have a significant effect on the magnitude of the stimulation.

A complete discussion of this system is provided by Tan (1996) and Tan & Rabinowitz (1996), which includes a detailed description of the hardware components, controller components, and performance characteristics of the device.

To improve the performance capabilities of the Tactuator (e.g., to allow multimedia signal control and real-time processing of speech signals), the system has been upgraded with a new computer, DSP system, and electronic control system. A complete description of the changes to Tan's (1996) system is described below. Performance characteristics of the upgraded system are described and compared to those obtained with the original system.

6.1.2 Description of the Upgraded System

A block diagram of the upgraded system is provided in Fig. 6-2.

The hardware components of the current system remain unchanged from those of the original system. These components include the motor assemblies, the fingerpad interface, the angular-position sensor (Schaevitz, R30A), the power amplifier (Crown D-150A), and the supporting structures.

The changes to the controller components of the device begin with a new host computer, a Dell Pentium III running Windows 2000 (replacing the Trident 486), which provides a more powerful platform for multimedia operations. The new host PC contains the Bittware Hammerhead PCI DSP card (dual ADSP-21160 processors, 32/40-bit floating point, 960M FLOPS) replacing the Texas Instruments DSP card (TMS320C31 processor, 32 –bit floating point, 50M FLOPS). This new DSP card allows for real time processing of the speech signals. The Bittware AudioPMC+ audio I/O card (24-bit, 96 kHz audio converter, +/- 2V maximum) provides eight input and eight output channels replacing the two Burr-Brown daughter modules with a total of four input and four output channels which use 16-bit converters, with +/- 3 volt maximum for both the input and

output. Finally, an electronic analog PID controller is used instead of the original digital PID controller since the AudioPMC+ audio I/O card does not carry dc signals (-3 dB points at 2 Hz and 20 kHz), and as such cannot be used for control applications. The electronic PID controller is an analog circuit performing second order control functions (a detailed description of the circuit is available in Brughera, 2002).

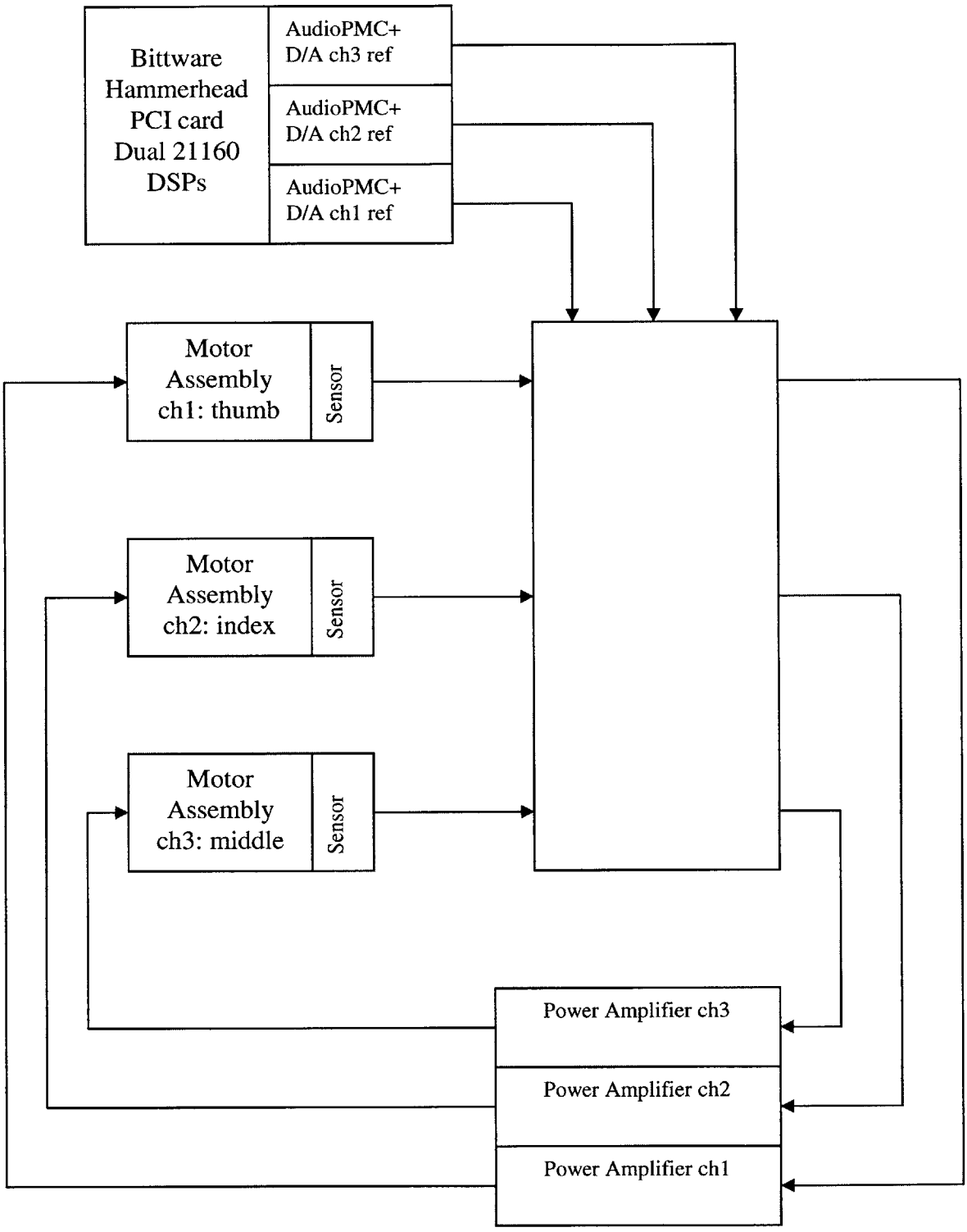

Fig. 6-2. Tactuator system with analog PID controller and reference input from Bittware Hammerhead DSP and AudioPMC+. (Taken from Brughera, 2002).

6.1.3 Performance of the Current System

The performance of the current system is generally similar to that of Tan's (1996) system. Measurements of the performance of the upgraded system, obtained with a Hewlett-Packard dynamic signal analyzer (35660 A), are summarized below. The analyzer measures signals in terms of dB re 1 V rms (dB V rms). A value of 0 dB V rms is roughly equivalent to 76 dB μm peak. [Note that a full sensor output range of +/- 3 volt corresponds to the full range of motion of 25.4 mm]. Thus 0 dB μm peak is equivalent to -76 dB Vrms: $\left(\frac{1}{\sqrt{2}}\times 10^{-3}\times\frac{6}{25.4}\right) V\ rms = 1.67\times 10^{-4}\ V\ rms, or, -75.54\ dB\ V\ rms$.

Frequency Response: The frequency response was measured as the ratio of the spectrum of the sampled sensor reading and the spectrum of the reference signal. The level of the reference was set at –20 dB V rms. The magnitude response and group delay of channel 1 are shown in Fig. 6-3 without the finger contacting the rod. The results are similar to those reported by Tan (1996) with the exception of a slight shift in the resonance frequencies (16.5 to 24.5 Hz in the present system vs. roughly 28.5 to 30.5 Hz in the original system) and the largest group delay (8 msec vs. 14 msec).

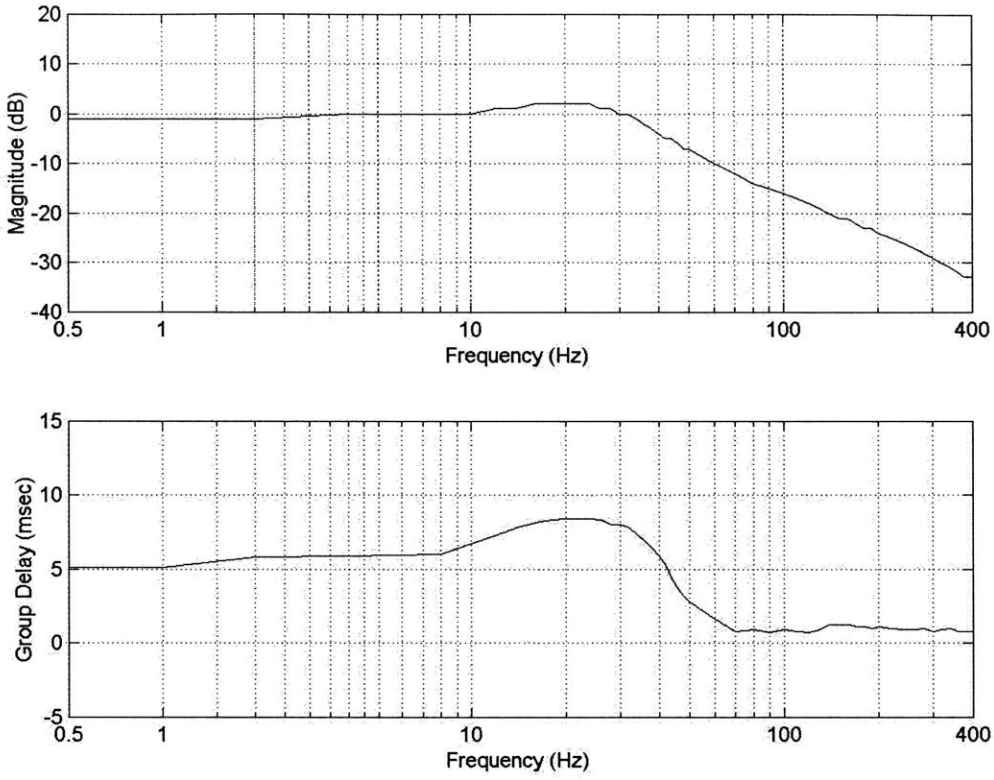

Fig. 6-3. Close-loop frequency response of channel 1 measured with a noise input. The upper panel is the magnitude response, and the lower panel is the group delay.

System Linearity: Linearity was measured for pure-tone inputs by the relationship between the sensor output levels and a wide range of reference input levels. Measurements were taken on channel 1 for one frequency within each of three distinct regions of tactual stimulation: kinesthetic (2 Hz), cutaneous (200 Hz), and mid-frequency region (20 Hz) for both "loaded" and "unloaded" conditions. In addition, further measurements were obtained for the specific channel and frequency parameters employed in the current psychophysical and speech experiments. Specifically, loaded and unloaded measurements were taken at 50 Hz on channel 1 (thumb) and 250 Hz on channel 2 (index

finger). The results are shown in Fig. 6-4. The left panel of Fig. 6-4 shows the linearity at each of the three frequencies (2, 20, 200 Hz) on channel 1. Since the data points at 2 Hz are close to those at 20 Hz, the 2-Hz results are offset by 20 dB. The right panel of Fig. 6-4 shows the linearity for 50 Hz on channel 1 and 250 Hz on channel 2. The best-fitting (using the least-square-error method) straight lines with unit-slope are also shown in the figure for each frequency under each condition.

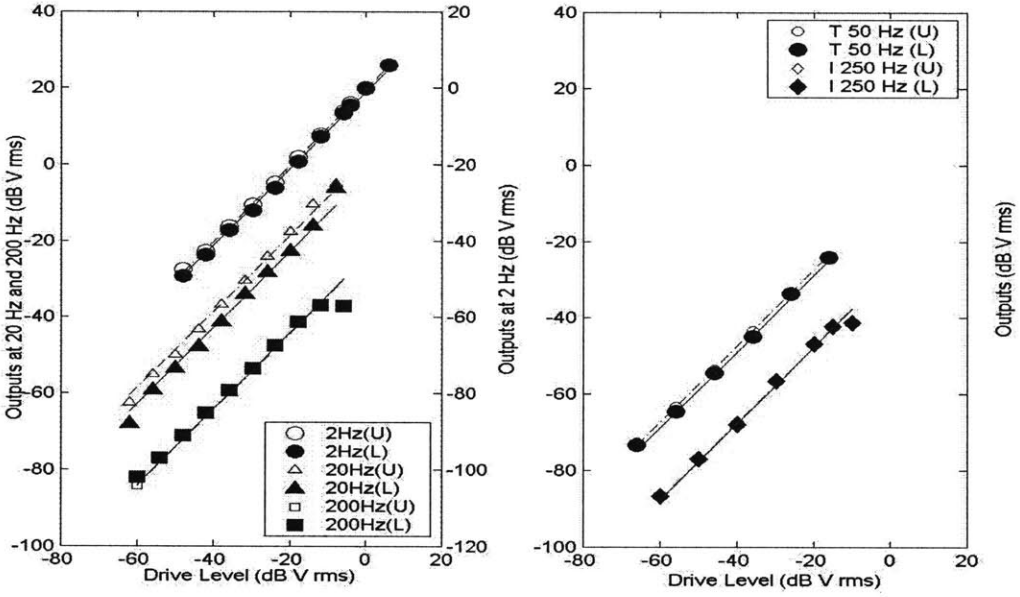

Fig. 6-4. Input-output relationship with best-fitting unit-slope lines.

The updated system demonstrates high linearity at each frequency, channel, and loading condition. The effect of loading appears to be greater at 20 Hz than at the other frequencies, indicating a reduction in motion of roughly 4 dB for loaded relative to unloaded movements. The reductions in motion under the loaded condition of the

upgraded system are compared with those of Tan's original system in Table 6-1, and are found to be generally quite similar.

Table 6-1. Comparison of the reductions in motion under loaded condition for upgraded system with those of Tan's original system (as a function of frequency and channel). T = channel 1 (Thumb); I = channel 2 (Index Finger). 2T = 2 Hz at the thumb, 20T et al.

	2T	20T	200T	50T	250I
Tan's (dB)	2	3	0	NA	NA
Upgraded (dB)	1	4	0	1	0

Noise Characteristics: The system noise includes the mechanical noise, electrical noise, and power-line noise. The noise characteristics of the current system were measured at the sensor outputs with the reference signals of all three channels set to zero. The sensor output from channel 1 is shown in Fig. 6-5. The most prominent components of noise are those related to the power-line frequency of 60 Hz and its third harmonic at 180 Hz. The noise level at 60 Hz is roughly -85 dB V rms for channel 1 and -90 dB V rms for channel 2, the noise level at 180 Hz is roughly –80 dB V rms for channel 1 and -85 dB V rms for channel 2, and the noise level at 300 Hz, (which is close to the noise floor), is roughly –95 dB V rms for channel 1 and -99 dB V rms for channel 2. These measurements are generally comparable to those made by Tan. The noise in the lower frequency region (<100 Hz) is generally below the detection threshold. The noise in the higher frequency region (>100 Hz) is greater than the detection threshold at certain frequencies (e.g., 180 Hz). In general, however, the subjects did not report feeling the electrical noise in the system at rest.

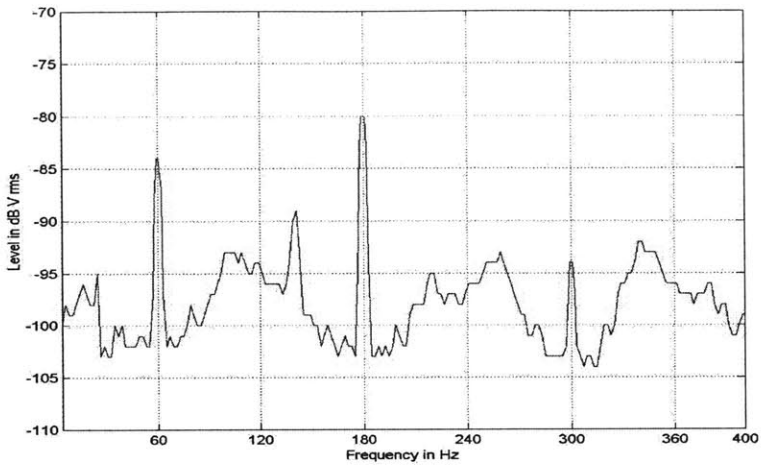

Fig. 6-5. Noise spectrum of channel 1.

Harmonic Distortion: The harmonic distortion was measured using single-tone inputs, with frequencies ranging from 1 Hz to 300 Hz. The reference drive amplitude was adjusted for each frequency so that the sensor output at that frequency was roughly 56 dB SL (relative to the thresholds from Bolanowski et al., 1988 and from Lamore et al., 1986). (Note that 56 dB SL is roughly the highest level of the Tactuator's operating range, and that the levels of the stimuli employed in the current study are roughly in the range of 25 dB to 45 dB SL.) The levels at the fundamental frequency and at the 2^{nd} through to 6^{th} harmonics at the sensor output were measured through the spectrum analyzer. The results of channel 1 are presented in Figs. 6-6 (for the unloaded condition) and 6-7 (for the loaded condition). In each of the two Figures, the output of the reference amplitude at each frequency is represented by circles. Each harmonic from 2^{nd} through 6^{th} is represented by a different symbol. Thus, for example, the level of the 2^{nd} harmonic of each input frequency is plotted with squares. In general, harmonics under the loaded condition were larger than those measured under the unloaded condition.

The absolute-detection curve (from Bolanowski et al., 1988) is plotted in each of the Figures as a solid line. For the unloaded condition, the level of harmonics lies in the range of ±10 dB relative to threshold. For the loaded condition, the level of harmonics nearly always exceeds threshold values by roughly 5 to 20 dB. The stimulus levels employed in the current study, however, are roughly 10 to 30 dB lower than the reference amplitude employed for these measurements. Thus, the effects of harmonics should be negligible in the current study.

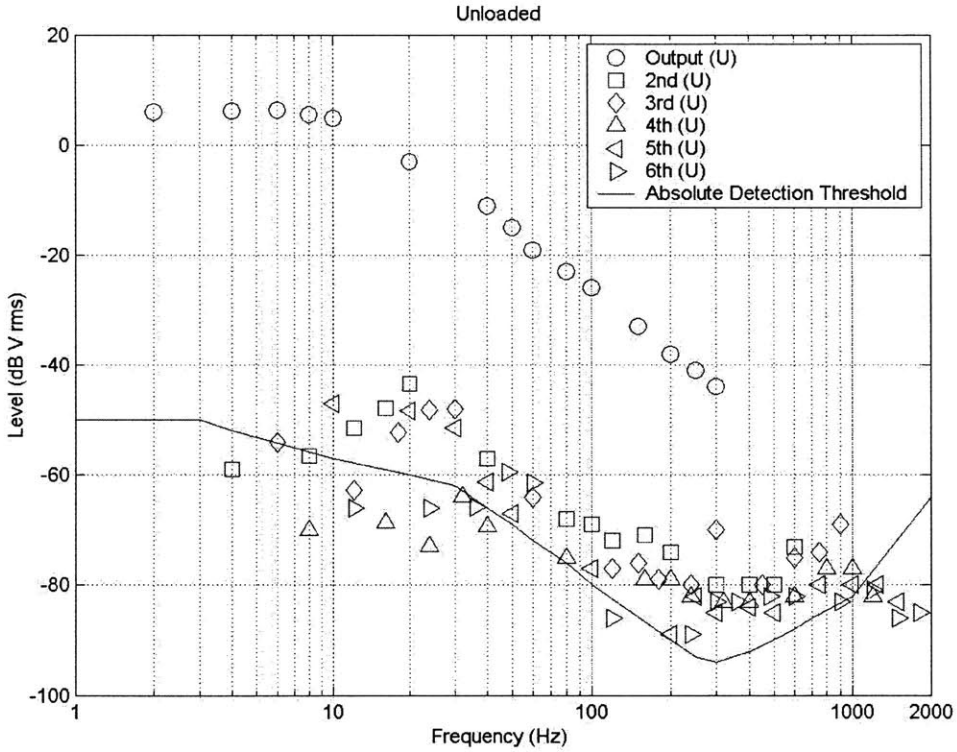

Fig. 6-6. Harmonic distortion under "unloaded" condition of channel 1.

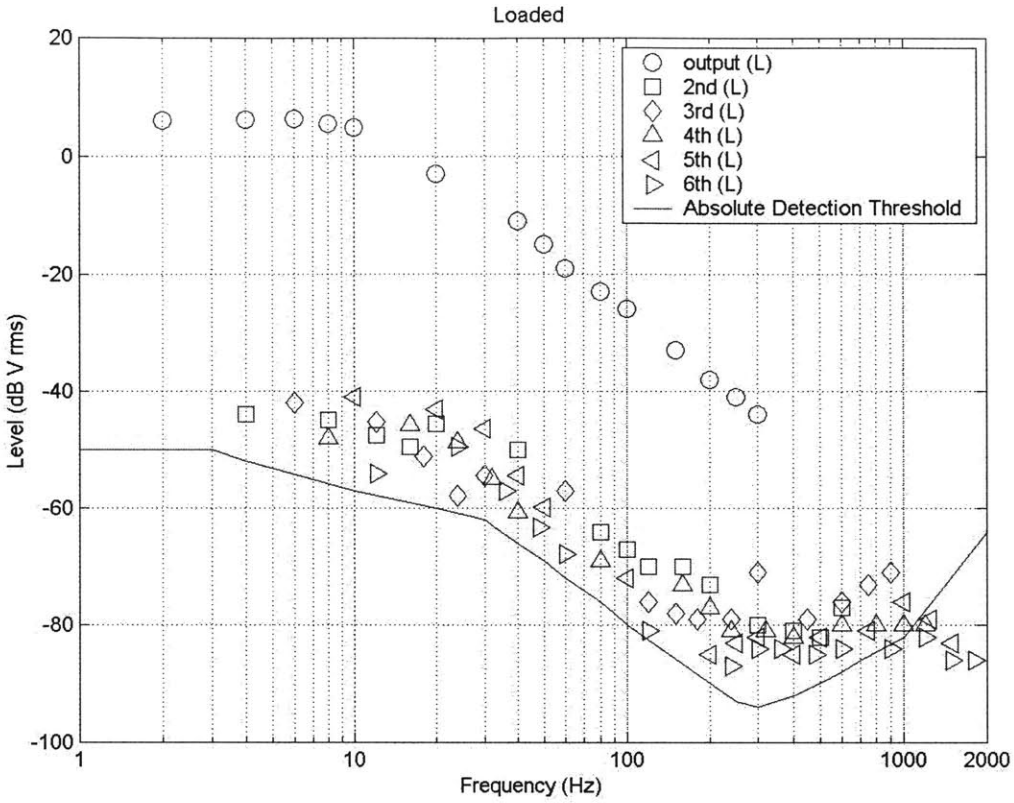

Fig. 6-7. Harmonic distortion under "loaded" condition of channel 1.

Cross Talk: Measurements of crosstalk among the three channels at four frequencies (2, 20, 50, 200 Hz) under both moderate and high stimulus levels are shown in Table 6-2a for channel 1. The crosstalk for a signal of 250 Hz on channel 2 was also measured (see Table 6-2b). Crosstalk for input frequencies on channel 1 in the range of 2 to 200 Hz ranges from -35 to -81 dB re L1 for channel 2, and from -36 to -78 dB re L1 on channel 3. For an input of 250 Hz on channel 2, crosstalk ranges from –26 to –35 dB on channel 1 and 3. Isolation between the channels diminishes as frequency increases. The maximum level of the stimuli used in the current study was roughly 45 dB SL; thus, the effect of crosstalk is small.

Table 6-2a. Crosstalk measurements for channel 1. L1 (dB Vrms) is the input voltage on channel 1 in dB Vrms, L1 (dB SL) is the corresponding vibration amplitude of channel 1 in dB SL, L2 (dB re L1) is the crosstalk between channel 2 and channel 1, and L3 (dB re L1) is the crosstalk between channel 3 and channel 1.

Freq	L1 (dB Vrms)	L1 (dB SL)	L2 (dB re L1)	L3 (dB re L1)
2	6	56	-73	-63
2	-6	44	-81	-72
20	-3	57	-69	-71
20	-21	39	-69	-78
50	-21	53	-44	-46
50	-42	40	-52	-46
200	-37	48	-35	-38
200	-50	27	-35	-36

Table 6-2b. Crosstalk measurements for channel 2. L2 (dB Vrms) is the input voltage on channel 2 in dB Vrms, L2 (dB SL) is the corresponding vibration amplitude of channel 2 in dB SL, L1 (dB re L2) is the crosstalk between channel 1 and channel 2, and L3 (dB re L2) is the crosstalk between channel 3 and channel 2.

Freq (Hz)	L2 (dB Vrms)	L2 (dB SL)	L1(dB re L2)	L3 (dB re L2)
250	-42	51	-30	-35
250	-61	32	-26	-29

The modified system is more powerful than the original one: more complicated algorithms can be processed in real-time with the fast processing power of the ADSP-21160 SHARC processor and a 64-bit, 66 MHz PCI interface. The expanded input and output channels provided by the audio PMC+ IO card enables eight channels of input and output. The new computer and operating system allow for processing and presentation of multimedia stimuli, such as the audio-visual speech materials employed in the present study.

6.2 Description of Speech Materials

The speech stimuli used in this study include nonsense syllables and sentences.

Nonsense Syllables. The nonsense C_1VC_2 syllables (see a detailed description in Chapter 5.1.1) used in the present study were chosen from the two sets of recordings that were balanced for C_1. Only those syllables with C_1 = /p, t, k, b, d, g, f, th, s, sh, v, tx, z, zh, ch, j/ and V = /i/, /a/ or /u/ were selected for the current study. Across both sets of lists, a total of 10 tokens of each C_1V combination (8 tokens from the 3-vowel context set and 2 tokens from the 16-vowel context set) were available for each of the two talkers. Thus, the total stimulus set consisted of 960 syllables (16 consonants × 3 vowels × 2 speakers × 10 repetitions).

The tokens were subdivided into two different sets for use in the experiments: a "training" set and a "testing" set (see Table 6-3). The training set consisted of 576 tokens (6 tokens per talker for each of the 48 C_1V combinations). The testing set consisted of the remaining 384 tokens (4 tokens per talker for each of the 48 C_1V combinations).

Table 6-3. Number of tokens for each C_1V combination.

	Training		Testing	
Vowel	**Speaker 1**	**Speaker 2**	**Speaker 1**	**Speaker 2**
/i/	6	6	4	4
/a/	6	6	4	4
/u/	6	6	4	4

CUNY sentences. The CUNY sentences were recorded by a female talker onto laser videodiscs (Boothroyd, Hanin, & Hnath, 1985) and consist of 60 lists of 12 sentences each. The length of the sentences ranges from 3 to 14 words and each list

contains 102 words (all of which are used in scoring). The CUNY sentences are considered to be of easy-to-medium difficulty because of their conversational style. Although the sentences in each list are arranged by topic, these topics were not made known to the subjects in this study.

6.3 Signal Processing

6.3.1 Procedures for Digitizing Speech Materials

All speech materials were digitized for real-time speech signal processing and for rapid-access control in the experiments. The details of the digitization of the C_1VC_2 nonsense syllables were described in section 5.1.2. The CUNY sentences were originally available on laser videodiscs. These analog recordings were digitized in the same way as that for the nonsense C_1VC_2 syllables. Each file contained one sentence for CUNY sentences. The number of frames for each sentence averaged roughly 140 (or a duration of roughly 4.67 sec), and the average size of the files was around 15 Mbytes. To reduce their size, the segmented files were compressed using the Sorenson Video compressor and converted to QuickTime format. The video was cut to 540×470 from 720×480, audio set to 22050 samples/sec, 16-bit mono, and the keyframe set to every 150 frames. Following compression, the size of the CUNY sentences averaged roughly 1.6 Mbytes.

6.3.2 Description of Envelope Processing for Tactual Display

The digitized acoustic speech signal was processed for presentation through the tactual display. This processing consisted of the computation of the amplitude envelope

signal from two different filtered bands of speech.

A flowchart of the processing algorithm is shown in Fig. 6-8. The multimedia file of the token was played through QuickTime for Java. The speech sound from the audio channel of the host computer was routed to the input of the AudioPMC+ audio IO card. The acoustic input was sampled at 48000 samples/second by an A/D converter. The speech samples were then processed through two parallel branches of the DSP board.

In branch 1 (the left branch), speech samples were passed through a discrete second-order Butterworth low-pass filter with a cutoff frequency of 350 Hz. This band was selected to monitor the presence or absence of low-frequency energy that accompanies glottal vibration. The low-pass filtered speech samples were then rectified and smoothed by the same discrete second-order Butterworth low-pass filter with a cutoff frequency of 25 Hz. A threshold (T1) was established for the level of the smoothed amplitude envelope (A1) in order to eliminate envelope signals arising primarily from random noise fluctuations in the passband, but yet sufficient for passing signals driven by the acoustic speech waveform. Values of amplitude-envelope below this threshold were set to zero. Values of the amplitude envelope samples above this threshold were added by the average threshold of 50 Hz of the left thumb across the three subjects (S1, S2 and S3, see Table 7-1a). The resulting amplitude-envelope signals are supra-threshold and thus can be detected by the subjects. Finally, the smoothed amplitude envelope modulated a 50-Hz sinusoid and was converted to an analog signal through a D/A converter. This modulated low-frequency was then routed to the reference signal of the PID controller to drive the thumb.

In branch 2 (the right branch), speech samples were passed through a discrete second-order Butterworth high-pass filter with a cutoff frequency of 3000 Hz. This band was selected to monitor the presence or absence of high-frequency energy that accompanies aspiration, frication, and burst characteristics of consonants. The high-pass filtered speech samples were then rectified and smoothed by a discrete second-order Butterworth low-pass filter with a cutoff frequency of 25 Hz. A threshold (T2) was established for the level of the smoothed amplitude envelope (A2) to eliminate envelope signals arising from random-noise fluctuations in the passband, but yet sufficiently low to pass signals arising from speech energy in the passband. Values of the amplitude envelope below this threshold were set to zero. Values of the amplitude envelope above this threshold were shifted by the 250-Hz absolute-detection threshold (averaged across the three subjects S1 to S3 at the left index finger, see Table 7-1a). The resulting amplitude envelope is thus always presented at a supra-threshold level. The smoothed envelope signal modulated a 250-Hz sinusoid and passed through a D/A converter. The resulting signal was then routed to the reference signal of the PID controller to drive the index finger.

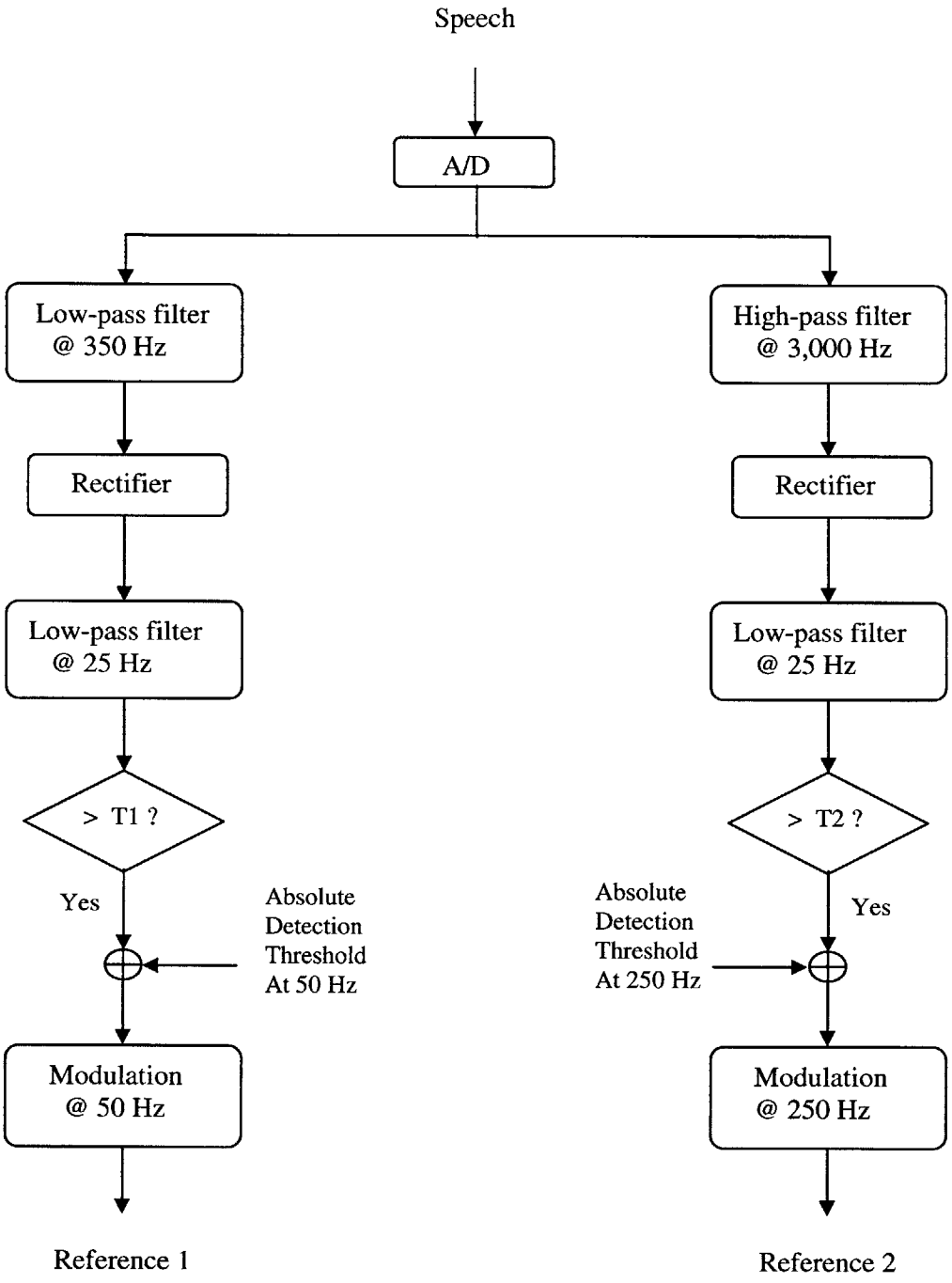

Fig. 6-8. A flowchart of the envelope-extraction algorithm.

The selection of the two modulating frequencies of 50 and 250 Hz for the lowpass and highpass bands, respectively, were based on the following criteria: (1) The two modulating envelopes have the same order as the two frequency regions from which the envelopes are extracted, i.e., the carrier with higher frequency was modulated by the amplitude envelope from the higher frequency band, while the carrier with lower frequency was modulated by the amplitude envelope from the lower frequency band; (2) The two frequencies are perceptually distinct (see Tan, 1996); and (3) In general, the duration of the amplitude envelope signals exceeds 20 msec, thus, causing each modulating envelope to contain at least one cycle.

6.4 Subjects

Four individuals (three male and one female) served as subjects in this study. One additional subject began the experiments but quit after several days. Subjects were screened for absolute-threshold detection and for temporal-onset order discrimination before participating in the experiments. Hearing testing was conducted through the Audiology Department at MIT Medical to provide a baseline for each subject's hearing level prior to exposure to masking noise. All subjects were students or alumni of MIT and ranged in age from 21 to 32 years. None of the subjects had previous experience in experiments concerned with lipreading or tactual perception. Subjects S1, S2, and S3 require corrective lenses for normal vision and wore either glasses or contact lenses for the lipreading experiments. Subjects were paid for their participation in the study and received periodic incentive bonuses throughout the course of the study.

6.5 Psychophysical Experiments

6.5.1 Absolute Threshold Measurements

Experimental Paradigm: Thresholds were measured for the detection of tactual signals as a function of frequency in the range 2 to 300 Hz. Experiments were conducted using a two-interval, two-alternative, forced-choice adaptive procedure with trial-by-trial correct-answer feedback. On each trial, a sinusoid of amplitude A dB, smoothed by a Hanning window with a 5-msec rise-fall time, was presented in one of the two intervals with equal probability. Each interval was 500 msec in duration and the two intervals were separated by 500 msec. The subject's task was to determine on which of the two intervals the signal was presented on each trial. The intervals were cued through text presented on a computer monitor. Correct-answer feedback was also provided through text presented on the monitor.

The starting amplitude A of the signal was chosen to be roughly 20 dB (re 1 μm peak) above the thresholds reported by Tan (1996) at each frequency. The amplitude was changed adaptively using the two-down one-up procedure described by Levitt (1971): two correct responses lead to a decrease in signal level, and one incorrect response leads to an increase in signal level. This adaptation rule converges on the signal level corresponding to approximately 70.7% correct. A run was terminated following ten reversals in the direction of the signal amplitude. The arithmetic average of signal amplitude across the last six reversals was defined as the threshold. A step size of 4 dB used at the start of the experiment was reduced to 1 dB following the first two reversals.

Thresholds were obtained at eight frequencies in the range 2 Hz to 300 Hz: 2, 10,

25, 50, 100, 200, 250, 300 Hz. Thresholds were measured at both the thumb and the index finger of the left hand for each of the four subjects. Measurements were always taken on the index finger prior to the thumb. For the index finger, the subject conducted two consecutive runs at each frequency. Then, three additional runs were collected for each of the two frequencies 50 Hz and 250 Hz (which were the two carrier frequencies modulating the amplitude envelopes in the speech experiments). For the thumb, only one run was collected at each frequency, and two additional runs were collected at the two frequencies 50 Hz and 250 Hz. For each subject and each finger the order in which the eight frequencies were tested was selected randomly. The subjects were instructed to press "1" if they believed the signal was presented in interval 1, or press "2" if they believed the signal was presented in interval 2. They were instructed to guess when they were unsure. Each run took around 8 minutes. Each subject took on average 4 hours on two different days to complete this experiment.

During the experiment, the subject sat approximately 0.8 meters from the video monitor (DELL Trinitron), and 0.6 meters from the Tactuator. The subject placed either the index finger or the thumb of his/her left hand on the corresponding rod of the Tactuator (see Fig. 6-1). To eliminate any auditory cues created by the Tactuator, subjects wore foam earplugs that were designed to provide 30-dB attenuation and also wore earphones through which pink masking noise was delivered. The level of the noise was adjusted for each subject to be sufficiently loud to mask the auditory signals that were present for the vibration primarily at the higher frequencies.

Data Analysis: For each subject and each of the two fingers, thresholds were averaged

across the runs conducted at each frequency. Means and standard deviations of threshold levels (in dB re 1 μm peak) were examined as a function of frequency.

6.5.2 Temporal-Onset Order Discrimination

Experimental Paradigm: Temporal-onset order discrimination thresholds were measured for stimulation at two different fingers (left index and left thumb) through the Tactuator device. The experiments employed a one-interval two-alternative forced-choice procedure with trial-by-trial correct-answer feedback. The index finger was always stimulated with a 250-Hz sinusoid and the thumb was always stimulated with a 50-Hz sinusoid. Each interval included the presentation of the two stimuli [i.e., a 250-Hz tone at the index finger (250I) and a 50-Hz tone at the thumb (50T)] in one of two possible orders: 250I-50T or 50T-250I with equal probability. See Fig. 6-9 for a schematic drawing of the signals presented at each channel: the upper track represents channel 1: 50T and the lower track represents channel 2: 250I. Performance was examined as a function of stimulus-onset asynchrony (SOA). SOA is defined as the difference in time between the onsets of the signal at the index finger and the signal at the thumb [i.e., $SOA = OnsetTime_{50T} - OnsetTime_{250I}$]. The subject's task was to determine in which of the two orders the two stimuli were presented on each trial.

The experiment was conducted with trial-by-trial roving of the duration and amplitude of each of the two sinewaves. The value of duration for each of the two stimuli was selected independently from a uniform distribution of the following seven values: 50, 100, 200, 400, 500, 600, 800 msec (i.e., there are 7×7=49 duration pairs). Thus, the offset-order of the two stimuli on each trial is random, and the offsets cannot be used as a

cue for determining the temporal order. The value of amplitude was selected from a uniform distribution of five values of sensation level: 25, 30, 35, 40, and 45 dB SL relative to the average threshold levels across the three subjects (S1, S2 and S3). The ranges of duration and amplitude were selected to simulate those of the speech envelopes. It is known that a short stimulus is more difficult to detect than a long one (Verrillo, 1965). The amplitudes of the shorter stimuli (i.e., 50 msec and 100 msec) were compensated for according to the 3 dB/octave rule. Namely, the amplitude of 50-msec signals was increased by 6 dB relative to that of stimuli with duration ≥ 200 msec, and the amplitude of 100-msec signals was increased by 3 dB. The relation between sensation level and dB re 1 μm peak is shown in Table 6-4. Mean thresholds across the three subjects were roughly 1 dB re 1μm peak for 50T and -19 dB re 1μm peak for 250I.

Table 6-4. Relationship between the motions measured in dB re 1 μm peak and sensation level in dB (relative to the average thresholds across the three subjects).

	dB SL	dB re 1 μm peak
50 Hz	25	26
	30	31
	35	36
	40	41
	45	46
250 Hz	25	6
	30	11
	35	16
	40	21
	45	26

The temporal sequence of events for two representative trials is illustrated in Fig. 6-9. The upper panel represents the stimuli delivered to the thumb through channel 1, and

the lower panel represents the stimuli delivered to the index finger through channel 2. The Figure indicates sequential events for two full trials (Trial 1 and Trial 2). The onset of each of the two trials is marked by T1 and T4 respectively. The end of the trials is marked by T3 and T6. On Trial 1, T1 marks the onset of 250-Hz signal in channel 2 (index finger) and T2 marks the onset of a 50-Hz signal in channel 1 (thumb), and T3 marks the end of the signal in channel 1. The values of the duration and amplitude of each of the two signals are selected randomly and independently from the distributions described previously. Note (as in Trial 2) that it is possible for the signal with an earlier onset to terminate after the second stimulus. During the period from T3 to T4 (that is, the end of one trial to the beginning of the next trial), the following series of events occur: the computer presents a prompt for the subject to input a response, the subject inputs a response, the computer stores the response in its memory and presents the correct-answer feedback on the monitor, the computer computes the parameters for the next trial, and finally the computer presents a prompt for a new trial. The duration of this time period varies, depending primarily on the subject's response time.

SOA was defined as the time asynchrony between the onset of 250I at the index relative to the onset of 50T at the thumb, e.g., onset time of thumb minus onset time of index finger ($50T_{onset} - 250I_{onset}$). The absolute value of SOA is kept constant in each run; however, the sign of the SOA is dependent on the stimulus ordering and varies randomly from trial to trial within a run. Trial 1 represents a positive value of SOA, where the onset time of the index finger leads the onset time of the thumb (T2-T1 > 0). Trial 2 illustrates a negative value of SOA, where the onset time of the thumb leads the onset time of the index finger (T4-T5 < 0).

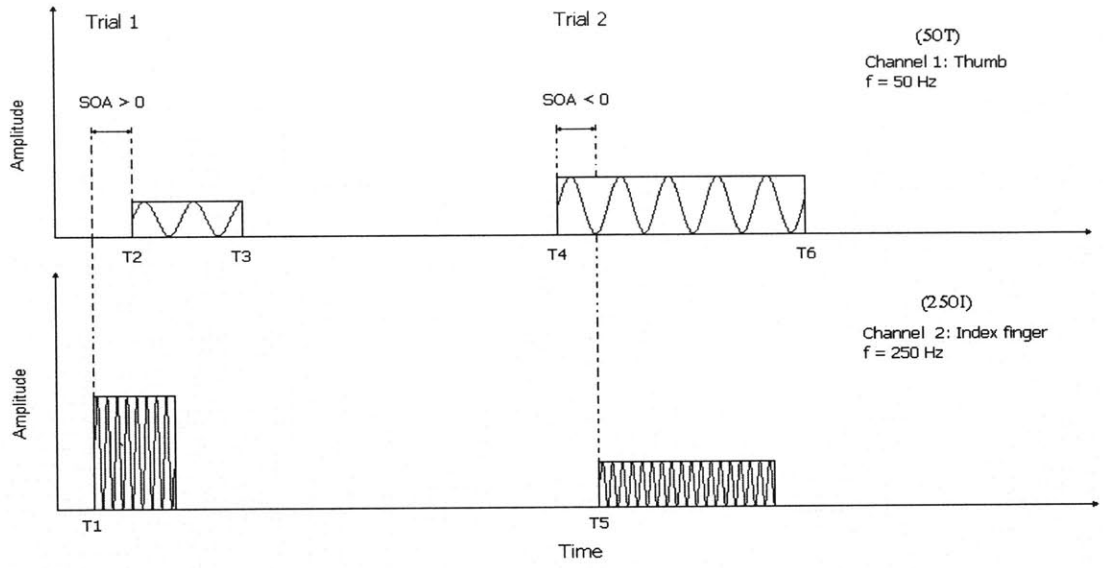

Fig. 6-9. Illustration of the time line for trials in the temporal-onset order discrimination experiment (the cycles inside the blocks are representative of the two frequencies: 50 Hz (channel 1) and 250 Hz (channel 2)).

Performance was measured as a function of stimulus-onset asynchrony (SOA). For each subject, data were collected at six different absolute values of SOA. The absolute value of SOA, which was fixed within a given run, was selected from a set of nine possible values in the range of 5 to 115 msec {5, 10, 25, 40, 55, 70, 85, 100 and 115 msec}. The particular SOA values used for each subject were selected on the basis of an initial screening process in which subjects were tested with values of SOA in decreasing order. In this initial testing, 2 consecutive 50-trial runs were collected at each SOA. Based on a given subject's performance on the initial testing, six values of SOA were selected for use in the main experiment to yield performance in the range of roughly 55-90% correct. Five experimental blocks were run using the six values of SOA for each subject.

Within each block, two consecutive 50-trial runs were conducted at each of the six values of SOA with the order of the SOA chosen at random. On average the experiment required 6 sessions on 6 different days for each subject.

During the experiment, the subject sat approximately 0.8 meters from the video monitor (DELL Trinitron) and 0.6 meters from the Tactuator, and placed the index finger and the thumb of the left hand on the two corresponding rods of the Tactuator. To eliminate any auditory cues from the vibration of the Tactuator, subjects wore foam earplugs that were designed to provide 30-dB attenuation and also wore earphones that delivered pink masking noise. The subjects were instructed to press "h" if they perceived that the 250I stimulus began first or "l" if they perceived that the 50T stimulus began first. They were instructed to guess when they were unsure of the order.

Data Analysis: For each experimental run and for each subject, the results were summarized using two percent-correct scores and the signal detection measurements of sensitivity (d′) and bias[1](β) (see Chapter 5.5, for a summary of these two measurements). The percent-correct scores were averaged over the 10 runs at each SOA and plotted as a function of SOA. The d′ and β were averaged over the 10 runs at each absolute value of SOA and plotted as a function of the absolute value of SOA. The threshold for temporal-onset order discrimination was defined as the absolute value of SOA at which d′=1. The

[1] The calculation of percent correct, d′, and β, are all based on the 2×2 confusion matrix $\begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$ obtained from each experimental run. The two percent-correct scores are the two ratios: $N_{11}/(N_{11}+N_{12})$ and $N_{22}/(N_{21}+N_{22})$ respectively. The values of d′ and β are also determined by these ratios. Thus, the two percent-correct scores of each run contain exactly the same information as d′ and β.

data were also classified into sub-ranges according to the amplitude differences and duration differences of the stimuli in each trial of the roving-discrimination paradigm, and analyzed to determine the effect of differences in amplitude and duration between the stimuli in each trial.

6.6 Speech Experiments

6.6.1 Pair-Wise Voicing Discrimination

Experimental Paradigm: The ability to discriminate consonant voicing was examined for eight pairs of consonants that contrast the feature of voicing. The eight pairs are: /p-b/, /t-d/, /k-g/, /f-v/, /th-tx/, /s-z/, /sh-zh/ and /ch-j/. Each pair was studied separately under the following three modalities: lipreading alone (L), tactual display alone (T), and lipreading combined with the tactual display (L+T). The tests were conducted using a two-interval, two-alternative, forced-choice procedure. Some conditions of the experiment employed trial-by-trial correct-answer feedback while others did not. On each trial, one of two possible stimulus orders (voiced-voiceless or voiceless-voiced) was selected at random with equal a priori probability. The subject was instructed to report the order of the presentation by clicking the icon "voiced-voiceless" if he/she perceived that the token with voiced initial consonant was presented in the first interval, or clicking the icon "voiceless-voiced" for the second alternative.

An example of a trial of the experiment is shown in Fig. 6-10. A trial is composed of two intervals, each with a fixed duration of 2 seconds. The onsets of intervals 1 and 2 are indicated by T0 and T4, respectively. One file is played in each interval. The files are

finished at T3 and T7. The times that the lips open and close are indicated by T1 and T2 for file 1, and T5 and T6 for file 2. Functionally, the stimulus in interval 1 ends at time T2 when the lips close (although the file extends for several more frames) and the stimulus in interval 2 begins at time T5 with the lip opening of the stimulus in interval 2 (although the file begins a few frames earlier). We define inter-stimulus interval (ISI) as T5-T2. ISI was a random variable with mean 530 msec, and standard deviation of 315 msec. During the period [T3, T4], a blue background was put on the screen.

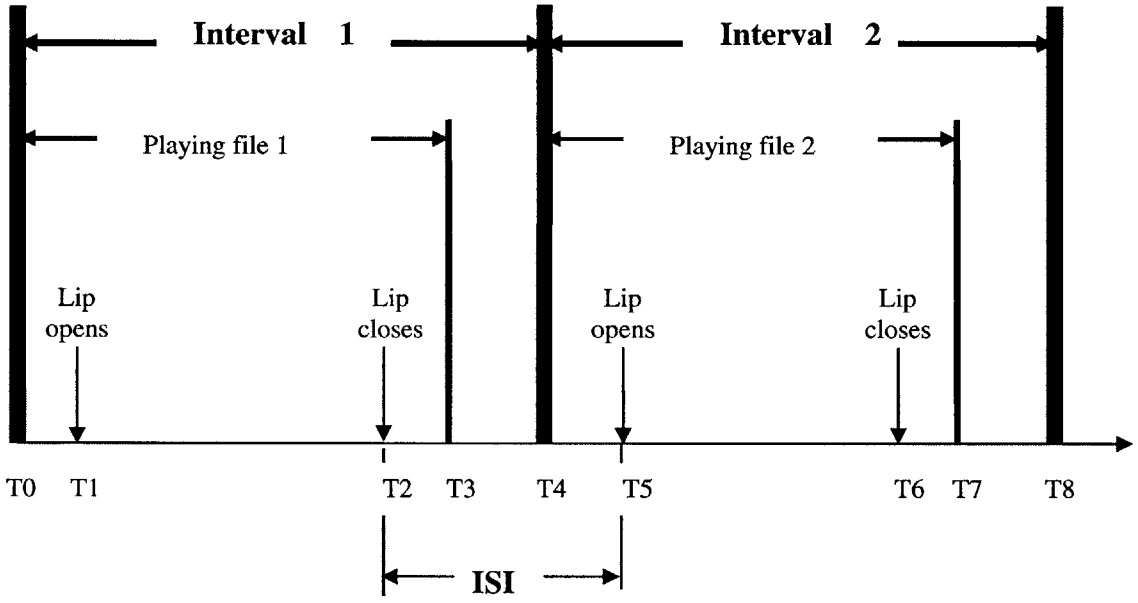

Fig. 6-10. A typical trial in the pair discrimination experiment. Details were explained in text.

Testing was conducted using the C_1VC_2 nonsense syllables described in Section 6.2. Among the 20 repetitions (2 speakers × 10 repetitions) for each C_1V combination, 6 repetitions of each speaker were used for the training, and the remaining 4 repetitions

were used for the testing. In other words, 36 tokens (2 speakers × 3 vowels × 6 repetitions) of each initial consonant were used in the "training" set, and 24 tokens (2 speakers ×3 vowels × 4 repetitions) of each initial consonant were used in the "testing" set. On each trial one of the tokens representing each of the two consonants in a given pair was selected at random from either the "training" set or the "testing" set with replacement.

Data were collected in five replications in which the eight pairs of consonants were tested separately under each of the three modalities. The order in which L and T were tested was chosen randomly for each consonant contrast; the modality L+T was always tested last. Each run consisted of 60 trials. A summary of the test parameters used in each of the five replications is shown in Table 6-5. The purpose of this design was to provide subjects with training on the task and then to test post-training performance. In the first replication, the eight stimulus pairs were tested in order of increasing difficulty based on the results of an informal pilot study (i.e., s-z, sh-zh, th-tx, p-b, t-d, k-g, f-v, ch-j). In the remaining four replications, the stimulus pairs were tested in a randomly determined order. The speech tokens from the "training" set were employed in replications 1-3. In the first two replications, training was provided in the term of trial-by-trial correct-answer feedback. In replication 3, subjects performed the task without feedback using the tokens from the training set. The final two replications of the experiment tested subjects' ability to transfer their training to a fresh set of stimuli (the "testing" tokens) without feedback. Each run took roughly 10 minutes; on average each subject required 10 two-hour sessions over 10 different days to complete the experiment.

Table 6-5. Parameters of each replication in pair-wise discrimination experiment.

Replication	Order of contrasts	Correct-answer feedback	Token set
1	Fixed	Yes	Training
2	Random	Yes	Training
3	Random	No	Training
4	Random	No	Testing
5	Random	No	Testing

For modalities of lipreading alone (L) and lipreading combined with the tactual supplement (L+T), the video image of the talker was displayed on a 19-inch color video monitor. The subject was seated roughly 0.8 meters in front of the video monitor. The tactual cue for aided lipreading in the L+T modality was presented simultaneously with the video signal. For the tactual alone (T) and L+T modalities, subjects were seated 0.5 meters from the Tactuator and placed the thumb and index finger of the left hand on the Tactuator. To eliminate any auditory cues from the vibration of the Tactuator, subjects wore foam earplugs that were designed to provide 30-dB attenuation and also wore earphones that delivered pink masking noise. The tactual cues employed in the T and L+T modalities were the two envelope signals generated using the method described in Section 6.3.2.

Data Analysis: The results of each experimental run (4 subjects × 8 consonant pairs × 3 modalities × 5 replications) were summarized in terms of a 2 × 2 stimulus-response confusion matrix (see Table 6-6 for an example). Signal-detection measures of d' and β for each matrix were computed assuming equal variance Gaussian distributions.

A three-way ANOVA analyses were conducted for each subject to test the three factors: pair, replication, and modality.

Table 6-6. 2×2 confusion matrix for pair-wise discrimination experiment. d′ and β were calculated using the formula described in Table 7-2. Vs = Voiceless, Vd = Voiced.

Stimulus\Response	Vs-Vd	Vd-Vs
Vs-Vd	N_{11}	N_{12}
Vd-Vs	N_{21}	N_{22}

6.6.2 16-Consonant Identification

Experimental Paradigm: The ability to identify the initial consonants in C_1VC_2 context was tested using a one-interval, 16- alternative forced-choice paradigm. Six lists of each speaker were assigned to the "training" set (576 tokens, 36 tokens per phoneme), and the remaining four lists of each speaker were assigned to the "testing" set (384 tokens, 24 tokens per phoneme). The experiment was conducted under three modalities: lipreading alone (L), tactual display alone (T), and lipreading combined with the tactual display (L+T). For modalities of lipreading alone (L) and lipreading combined with the tactual display (L+T), the talker's face was displayed on a 19-inch color video monitor. The tactual cues for aided lipreading in L+T modality were presented simultaneously with the video signal. For the T and L+T modalities, subjects were seated 0.6 meters from the Tactuator and placed both the thumb and index finger of the left hand on the Tactuator. To eliminate any auditory cues from the vibration of the Tactuator, subjects wore foam earplugs that were designed to provide 30-dB attenuation and also wore earphones that

delivered pink masking noise. The tactual cues employed in the T and L+T modalities were the two envelope signals generated using the method described in Section 6.3.2.

Data were collected in seven replications. In each replication, four consecutive runs of each of the three modalities were tested. The order in which L and T were presented was chosen randomly. The L+T modality was always run last. A run consisted of 80 trials. The values of feedback and token set used in each of the seven replications are shown in Table 6-7. Subjects received training in Replications 1 and 2 using the "training" set with trial-by-trial correct-answer feedback. In replication 3, subjects received practice on doing the task without feedback, using the "training" set of tokens. Testing was conducted in Replication 4 to 7 using the "testing" set of tokens. On each trial, one of the tokens was selected randomly with replacement from the appropriate set (either "training set" or "testing set") with equal probability. Each run took about 12 minutes; the entire experiment required an average of 6 two-hour sessions over 6 different days for each subject.

Table 6-7. Parameters for each replication of 16-consonant identification experiment.

Replication	Correct-answer feedback	Token set
1	Yes	Training
2	Yes	Training
3	No	Training
4	No	Testing
5	No	Testing
6	No	Testing
7	No	Testing

Data Analysis: The results from the final four replications for each subject and each condition were used to construct 16×16 confusion matrices. Mean and individual performance in percent-correct score and percent of information transfer (%-IT) were calculated from these matrices for each condition. Mean and individual performance in percent-correct score and percent of information transfer (%-IT) for the features of voicing, manner and place were also calculated from these matrices for each condition.

Given a confusion matrix $\left[\quad N_{i,j} \quad \right]$, the calculations of %-correct score (PC) and percent of information transfer (%-IT) are described below.

The %-correct score (PC) was calculated by Equation 6-1:

$$PC = \frac{\sum_i N_{i,i}}{\sum_i \sum_j N_{i,j}}, \qquad \textbf{(Equation 6-1)}$$

The information transmission in bits (IT) was calculated by (see Garner, 1962):

$$IT(x;y) = H(x) + H(y) - H(x,y).$$

where $H(x) = -\sum_i p_i \log p_i$, $H(y) = -\sum_j p_j \log p_j$, $H(x,y) = -\sum_{i,j} p_{i,j} \log p_{i,j}$,

$$p_i = \frac{\sum_j N_{i,j}}{\sum_i \sum_j N_{i,j}}, \quad p_j = \frac{\sum_i N_{i,j}}{\sum_i \sum_j N_{i,j}}, \text{ and } \quad p_{i,j} = \frac{N_{i,j}}{\sum_i \sum_j N_{i,j}}.$$

The relative information transmission %-IT was calculated by Equation 6-2:

$$\%-IT = IT(x,y) / H(x), \qquad \textbf{(Equation 6-2)}$$

6.6.3 CUNY Sentence Reception

Experimental Paradigm: Sentence reception was examined under two modalities: lipreading alone (L) and lipreading combined with tactual supplement (L+T). Data were not obtained with touch alone (T) because no significant information was expected for delivery through touch alone without extended training. Each subject viewed 3 lists/modality for training, and 27 lists/modality for testing. Two consecutive lists of sentences were presented for each modality alternatively.

For both modalities (L and L+T), the talker was displayed on a 19-inch color video monitor. Subjects sat approximately 0.8 meters from the video monitor. The tactual cue for aided lipreading in L+T modality was presented simultaneously with the video signal. In the L+T modality, subjects were seated roughly 0.6 meters from the Tactuator and placed the thumb and the index finger of the left hand on the rods. To eliminate any auditory cues from the vibration of the Tactuator, subjects wore foam earplugs that were designed to provide 30 dB of attenuation and headphones that delivered a broadband pink noise. After each sentence presentation, the subjects were given as much time as necessary to type their responses into the computer. The subjects were instructed to write down any part of the sentence that they understood. They were encouraged to guess even when they were not sure of the answer. Responses were recorded by the host computer and scored later. The response was compared to the stimulus sentence. Any words in the response that correspond exactly to words in the stimulus sentence were scored as correct. The total number of correct words was computed across the 12 sentences in each list and converted to a percent-correct score.

Data Analysis: Percent-correct scores were plotted for each subject and modality as a function of list number. In addition, scores were averaged across all lists on each modality for each subject.

Chapter 7

Results

7.1 Tactual Detection Thresholds

The results of the tactual threshold measurements are summarized in Fig. 7-1. Each of the four panels of the figure represents results for one of the four individual subjects. For each subject, data are shown for the index finger (unfilled circles) and for the thumb (filled circles). Each data point represents the mean of threshold estimates across the runs collected for each particular condition. The shape of the threshold curve was similar across subjects and displays the expected dependence of sensitivity on frequency of stimulation. Maximal sensitivities (i.e., lowest threshold values) are obtained in the region of 200-300 Hz across subjects and digits. Thresholds increase rapidly as frequency decreases below 200 Hz. Because the highest frequency tested was 300 Hz, the expected increase in threshold as a function of frequency is not evidenced in the present set of data. The standard deviation across runs at a given frequency and digit for each subject is small, ranging from approximately 0 dB to 5 dB (and averaging 2.5 dB).

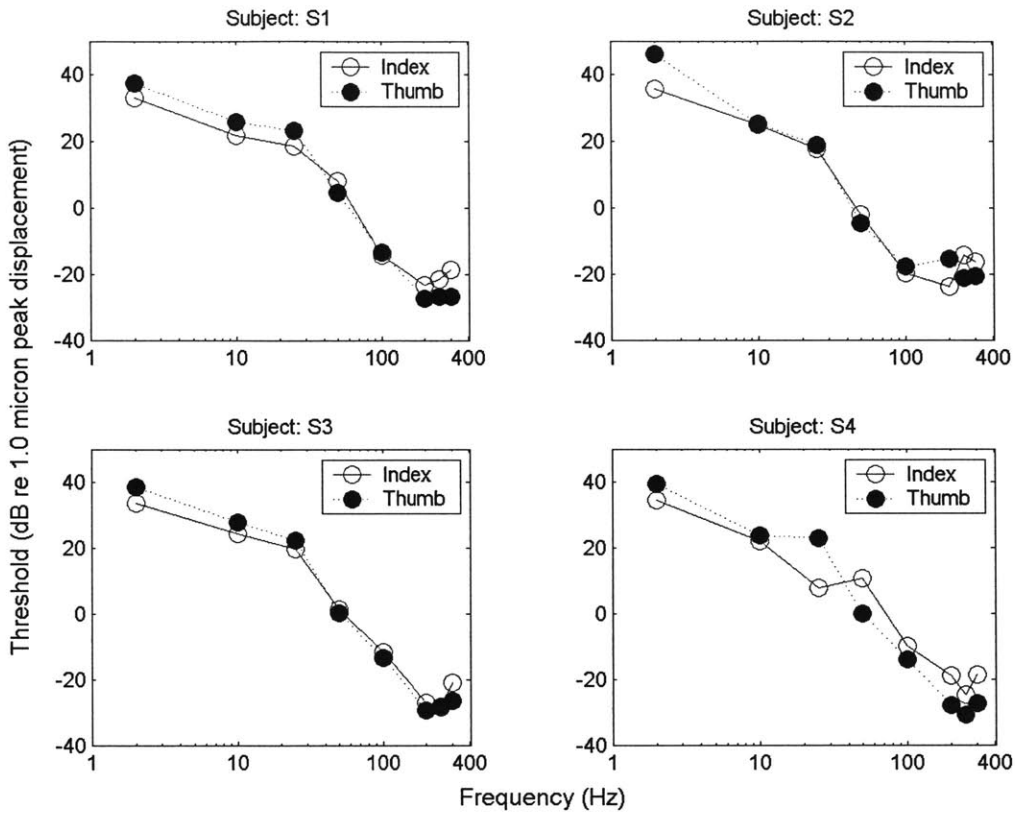

Fig. 7-1. Threshold in dB re 1.0-micron peak displacement versus frequency in Hz for each of the four subjects. Data points represent the means across runs collected at each frequency and digit.

Data averaged across the four subjects for each of the two digits are shown in Fig. 7-2. The left panel shows average results for the index finger and the right panel shows average results for the thumb. Each data point represents the mean of threshold estimates across the subjects. The shape of the threshold curve was similar for the two digits. Maximal sensitivities (i.e., lowest threshold values) are obtained in the 200-300 Hz range for each digit. Minimum threshold of -23 dB re 1.0 micron peak displacement was observed at 200 Hz for the index finger, and -27 dB re 1.0 micron peak at 250 Hz for the thumb. Thresholds increase rapidly as frequency values decrease below roughly 250 Hz.

The threshold curve below the maximal sensitivity can be approximated by two straight lines with different slopes: threshold decreases at a rate of about -5dB /octave from 2 Hz to 25 Hz and, then decreases at a rate about –12 to -14 dB/octave in the region of 25 Hz to 250 Hz. The standard deviations across subjects are small, and range from 1 dB to 6 dB across conditions. Finally, the mean and standard deviation of the thresholds at 50 Hz at the thumb and 250 Hz at the index finger are shown in Table 7-1a (across the first three subjects) and Table 7-1b (across the four subjects). These two frequency/digit combinations were used in the subsequent studies of temporal order and speech perception. The threshold means shown in Table 7-1a are used to define 0 dB SL in the remaining experiments. (The fourth subject joined the study after the other three had already begun the temporal onset-order experiment).

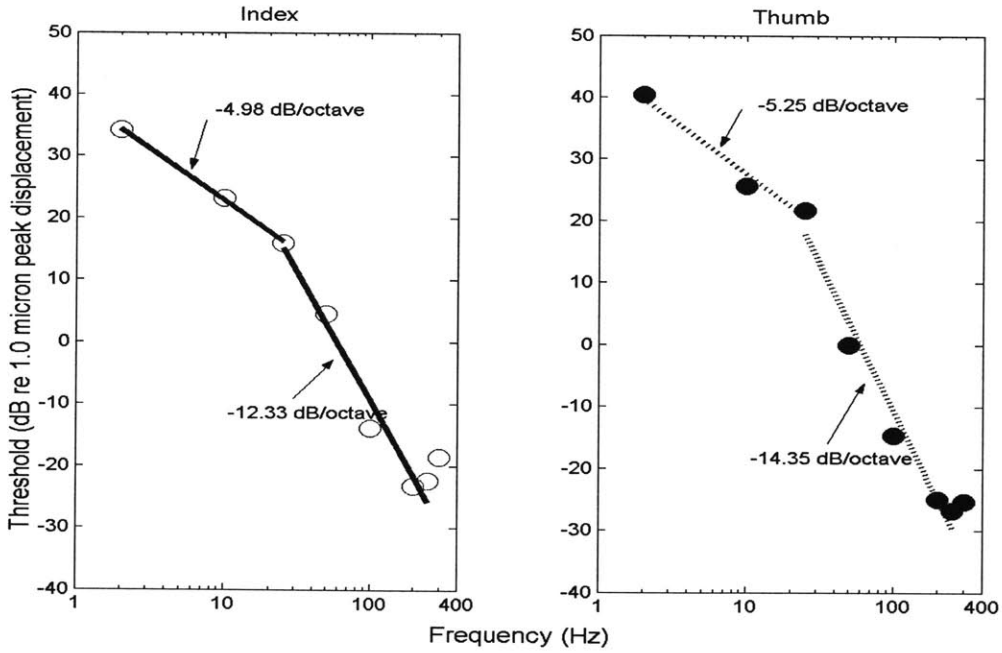

Fig. 7-2. Average threshold across the four subjects in dB re 1.0-micron peak displacement versus frequency in Hz for each digit. Data points represent the means across the four subjects.

Table 7-1a. Mean and standard deviation of the absolute threshold at 50 Hz for the thumb, and at 250 Hz for the index finger across the first three subjects.

	Mean (dB re 1.0 μm peak)	s.d. (dB re 1.0 μm peak)
50 Hz (Thumb)	1	5
250 Hz (Index)	-19	7

Table 7-1b. Mean and standard deviation of the absolute threshold at 50 Hz for the thumb, and at 250 Hz for the index finger across the four subjects.

	Mean (dB re 1.0 μm peak)	s.d. (dB re 1.0 μm peak)
50 Hz (Thumb)	0	4
250 Hz (Index)	-22	6

7.2 Tactual Temporal-Onset Order Discrimination

The results of the temporal-onset order discrimination task are presented in Fig. 7-3 for percent-correct performance as a function of stimuli onset asynchrony (SOA). In this figure, SOA is defined as the onset of the stimulus delivered to the index finger (250I) relative to the onset of the stimulus delivered to the thumb (50T). Thus, positive values of SOA represent conditions where the stimulus to the index finger preceded that to the thumb; the negative values represent cases where the thumb led the index finger. Each of the four panels of the figure represents results for one of the four individual subjects. Each point represents the mean of the percent-correct scores over the last ten runs for each subject.

The shape of the data curves is similar across subjects indicating a monotonic increase in performance with the absolute value of SOA (|SOA|) (above and below a minimum |SOA| which produced chance performance). The performance is generally

fairly symmetric relative to SOA=0 for S1, S2, and S4. For S3, performance is somewhat asymmetric: at the same absolute value of SOA, her performance is always better for trials with positive SOA than for trials with negative SOA. Individual differences are clear: the asynchrony required for 70%-correct performance ranges from 17 msec for the most sensitive subject (S4) to roughly 58 msec for the least sensitive subject (S2).

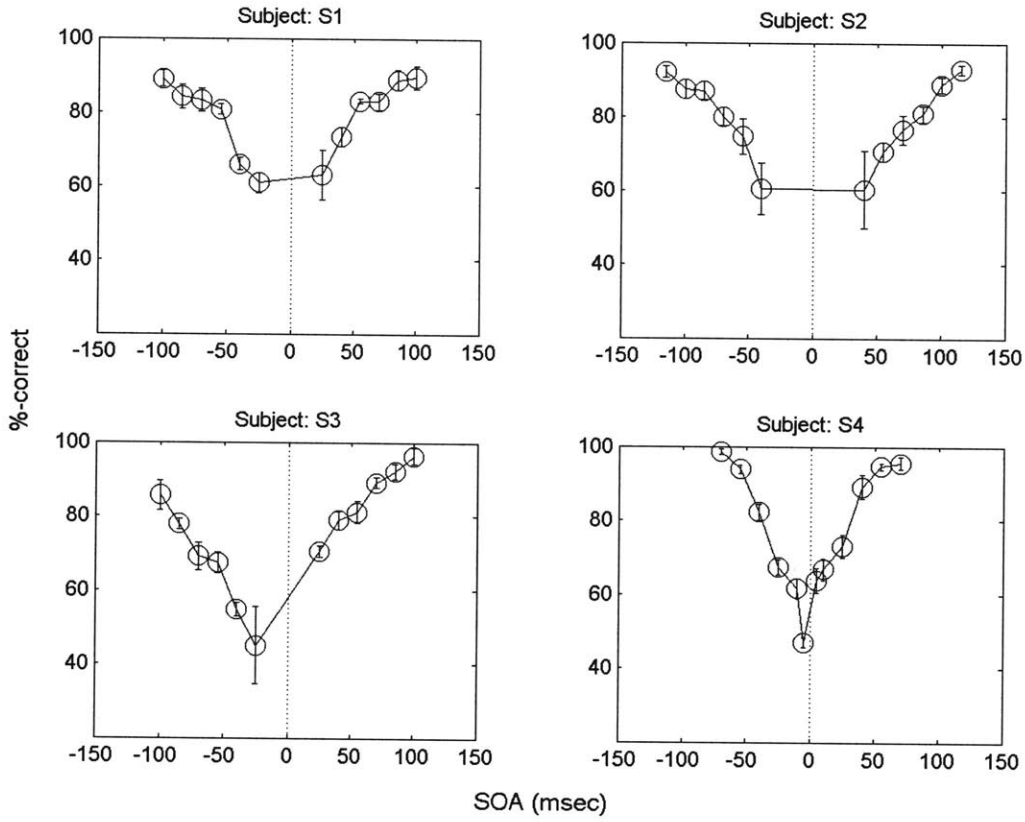

Fig. 7-3. Average score in %-correct versus stimuli onset asynchrony (SOA) for each of the four subjects.

In order to separate sensitivity from response bias, d′ and β for each subject were calculated from the collected data with the assumption that the underlying distributions of sensory events are Gaussian with equal variance. The trials in each run are grouped into a

2×2 confusion matrix. The values of d′ and β are calculated from each 2×2 confusion matrix. An illustration of the calculation of d′ and β is provided in Table 7-2.

Table 7-2. Calculation of d′ and β.

Stimulus\Response	SOA > 0	SOA < 0
SOA > 0	N_{11}	N_{12}
SOA < 0	N_{21}	N_{22}

$Pd = N_{11}/(N_{11} + N_{12}), Pf = N_{21}/(N_{21} + N_{22})$

Pd: the probability that the subject responds SOA > 0 given trials with SOA > 0

Pf: the probability that the subject responds SOA > 0 given trials with SOA < 0

Zd: Inverse of the normal cumulative distribution function with probability Pd

Zf: Inverse of the normal cumulative distribution function with probability Pf

$d' = Zd - Zf, \beta = -(Zd+Zf)/2$

The mean and standard deviation[1] of d′ are shown as a function of |SOA| in Fig. 7-4. The data of each subject are fitted by a straight line through the origin (0,0) by using the method of minimum-square-error. The data were well fit by a straight line through the origin for each of the four subjects (with correlation coefficients ranging from 0.971 to 0.996). The slopes of the lines range from 0.024 $msec^{-1}$ to 0.055 $msec^{-1}$. The thresholds (defined as the value of |SOA| at which d′=1) ranged from 18 msec to 43 msec across the four subjects. The specific values of the |SOA| at d′=1, the slope of the lines, the

[1] The standard deviation is the positive square root of the sample variance, which is the sum of the squared deviation from the mean divided by n-1. That is $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$, where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

correlation coefficient of the fit, and the root-mean-square (RMS) error of the fit are shown in Table 7-3 for each subject.

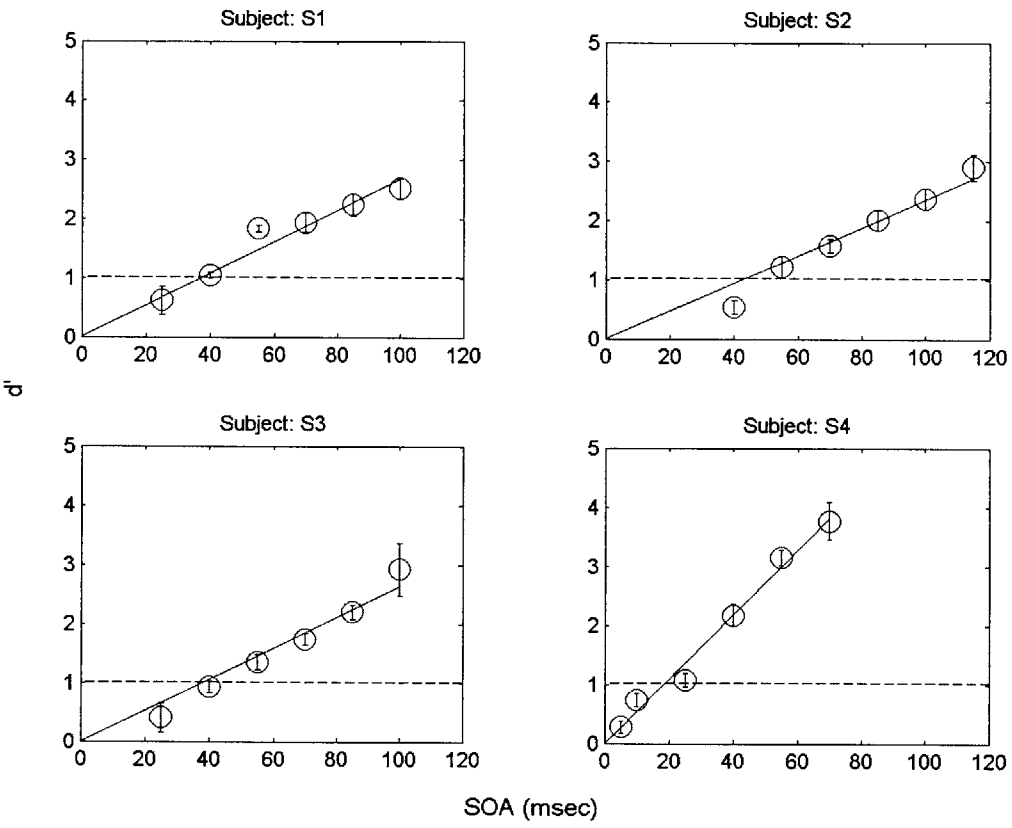

Fig. 7-4. Mean and standard deviation of d′ versus |SOA| for the four subjects.

Table 7-3. |SOA| at d′=1, slope of the fitting line, correlation coefficient of the fit, and RMS error of the fit for each of the four subjects.

Subjects	S1	S2	S3	S4
|SOA| (msec) @ d′=1	37	43	38	18
Slope of d′	0.027	0.024	0.026	0.055
Correlation coefficient	0.971	0.996	0.995	0.992
RMS error	0.166	0.188	0.177	0.158

The mean and standard deviation of β are shown as a function of |SOA| in Fig. 7-5. Positive bias indicates the subject has a greater tendency to respond SOA < 0 than to respond SOA > 0, while negative bias indicates the subject has a greater tendency to respond SOA > 0 than SOA < 0. The bias is negligible for subjects S1, S2, and S4, falling in the range -0.2 and +0.2. It is somewhat larger and negative at all |SOA| for S3.

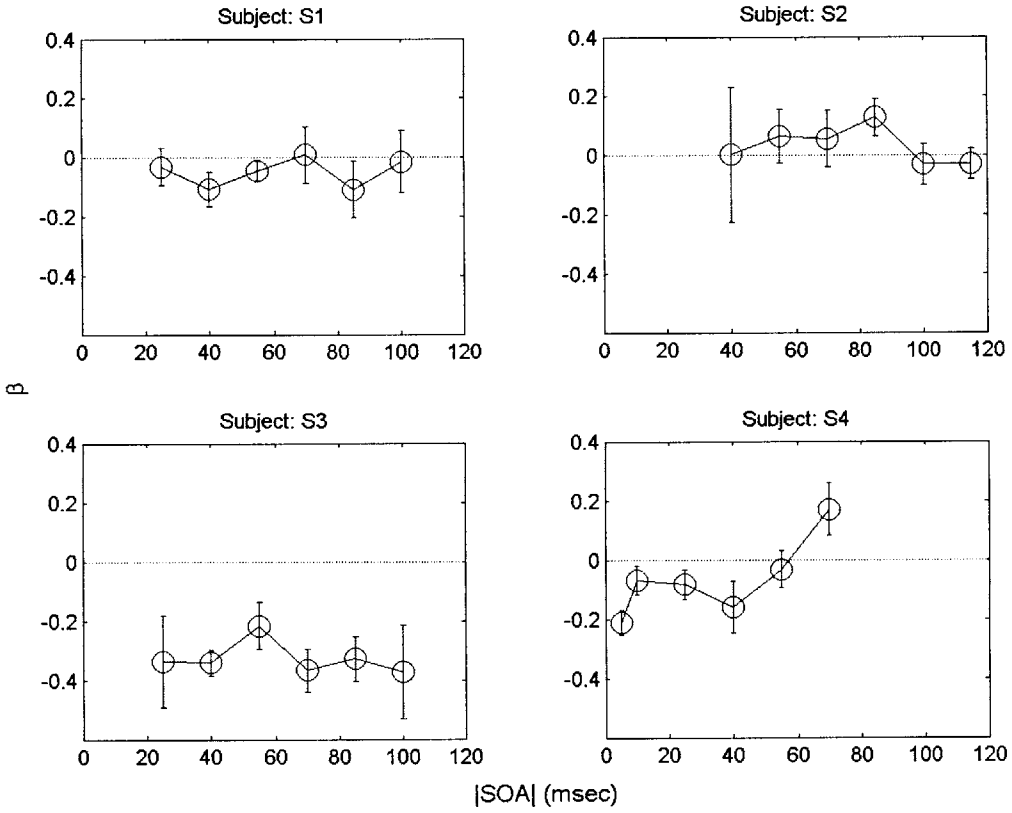

Fig. 7-5. Mean and standard deviation of β versus |SOA| for the four subjects.

The data were also classified into sub-ranges according to the amplitude differences and duration differences of the two stimuli in each trial of the roving-discrimination paradigm.

The amplitude-difference range (-20 to +20 dB SL), defined as the difference between the amplitude of the stimulus with an earlier onset (I_1) and the amplitude of the stimulus with a later onset (I_2), i.e., $I_1 - I_2$, was divided into 5 sub-ranges: ({-20, -15}, {-10, -5}, {0}, {5, 10}, {15, 20}). The amplitude difference is calculated for every possible amplitude pairing of the stimuli (see Table 7-4a). Each amplitude pairing is equally likely; thus, the distribution of each sub-range is the number of occurrences of each amplitude difference of the sub-ranges in Table 7-4a divided by 25 (the total number of all possible amplitude difference pairings of the stimuli), shown in Table 7-4b.

Table 7-4a. Mapping between the amplitude difference and the amplitude pair of the two stimuli in each trial. The rows represent the amplitudes of the first stimulus and the columns represent the amplitudes of the second stimulus (in dB SL).

$I_1 \backslash I_2$	25	30	35	40	45
25	0	-5	-10	-15	-20
30	5	0	-5	-10	-15
35	10	5	0	-5	-10
40	15	10	5	0	-5
45	20	15	10	5	0

Table 7-4b. Amplitude-difference distribution for each sub-range.

Category	**1**	**2**	**3**	**4**	**5**
Sub-range	{-20, -15}	{-10, -5}	{0}	{5, 10}	{15, 20}
Distribution	3/25	7/25	5/25	7/25	3/25

A 2×2 confusion matrix was derived for each sub-range class and each |SOA| from which the d′ was calculated. The values of d′ are shown for each sub-range in Fig. 7-6. Each of the four panels of the figure represents the results for one of the four subjects. For each subject, different symbols represent results obtained with different values of |SOA| (filled symbols represent the three largest values of |SOA|, and unfilled

symbols represent the three smallest values). The amplitude difference has a clear and consistent effect on performance for all subjects. For each subject, the data show a trend for highest levels of performance at large positive values of amplitude difference (i.e., when $I_1 >> I_2$) and for lowest levels of performance at large negative values of the amplitude difference (i.e., when $I_1 << I_2$). This trend is stronger at small values of |SOA| than at large values. For small |SOA| and negative amplitude differences, the perception of the temporal onset-order was the opposite of the order in which the stimuli were presented (i. e., d' is negative).

Individual values of β, provided in Fig. 7-7, indicate no large effects of amplitude pairing or |SOA| on bias.

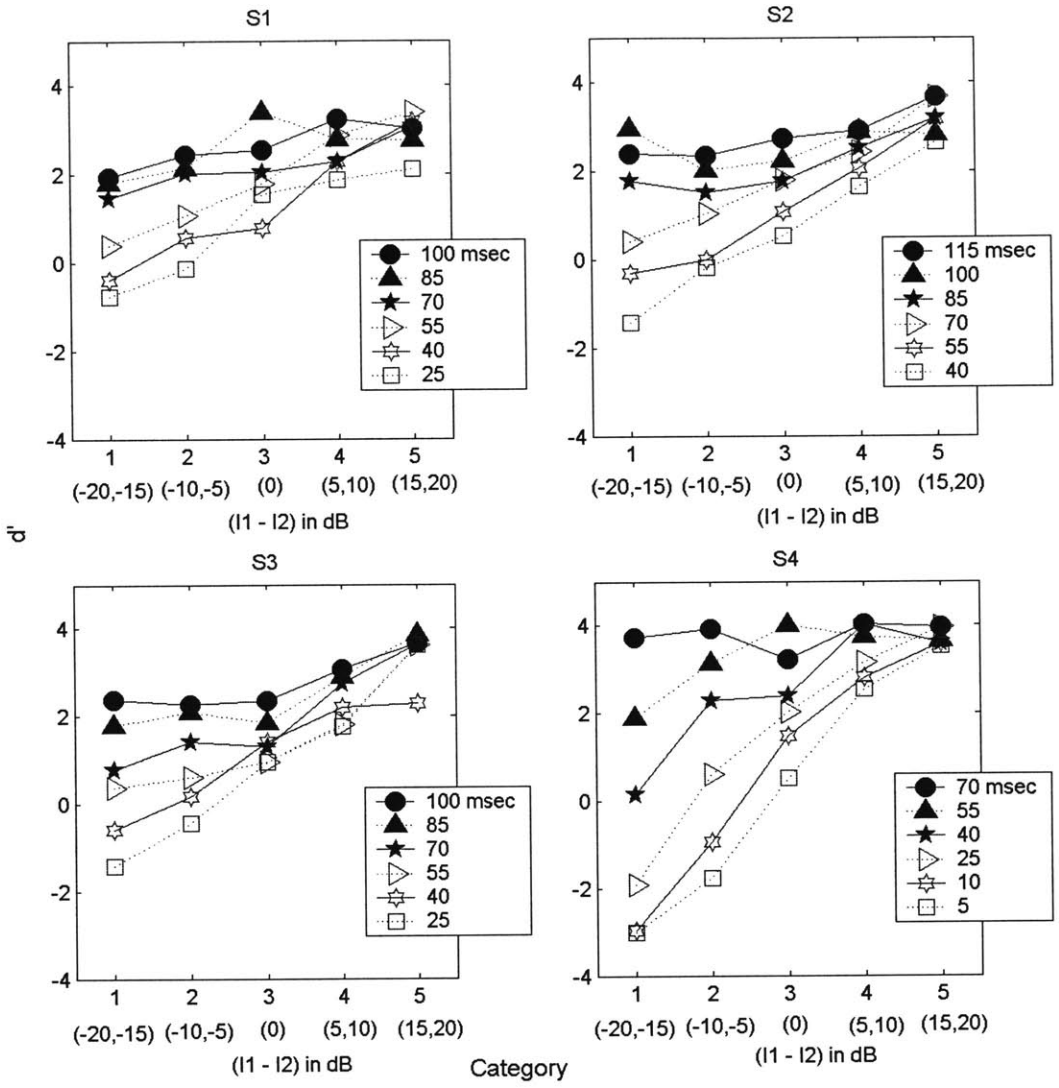

Fig. 7-6. d′ as a function of amplitude difference (I1 - I2) in dB for each |SOA| in msec for each subject. Filled symbols represent the three largest values of |SOA|, and unfilled symbols represent the three smallest values. See Table 7-4b for definition of category 1 through 5.

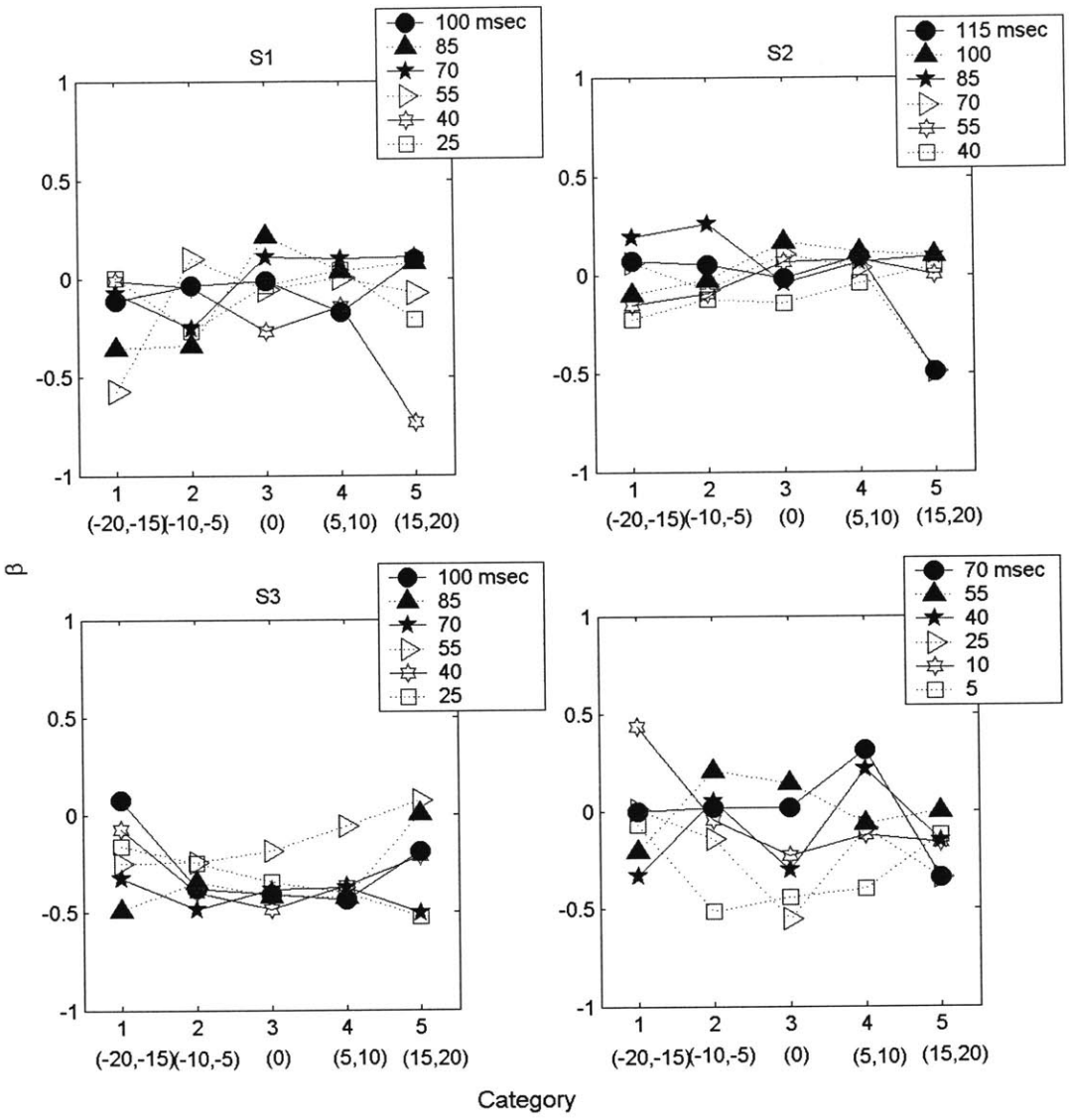

Fig. 7-7. β as a function of amplitude difference (I1 - I2) in dB for each |SOA| in msec for each subject. Filled symbols represent the three largest values of |SOA|, and unfilled symbols represent the three smallest values. See Table 7-4b for definition of category 1 through 5.

Values of d′ averaged across |SOA| are shown as a function of amplitude difference in Fig. 7-8 for each subject (represented by 4 different unfilled symbols), as well as the values of d′ averaged both across |SOA| and subjects (filled squares). This

Figure demonstrates a strong tendency for improvement in performance as the difference in amplitude ($I_1 - I_2$) increases.

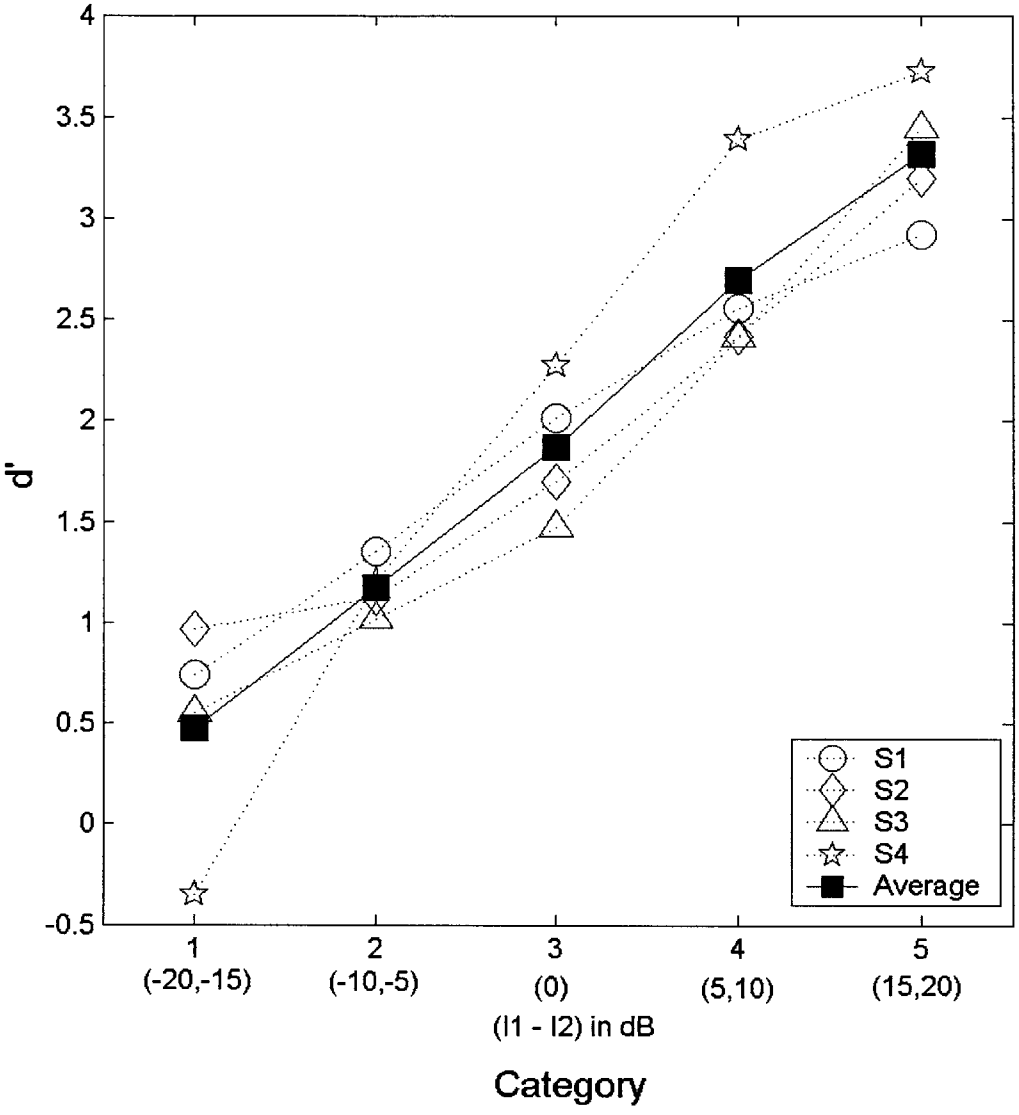

Fig. 7-8. d′ averaged across |SOA| for each subject and d′ averaged across both |SOA| and subjects as a function of amplitude difference.

Values of β averaged across |SOA| for each subject (shown in Fig. 7-9) indicate relatively flat bias for each subject as a function of amplitude pairing, a range from roughly -0.35 (S3) to 0.05 (S2), and an overall average of roughly -0.1.

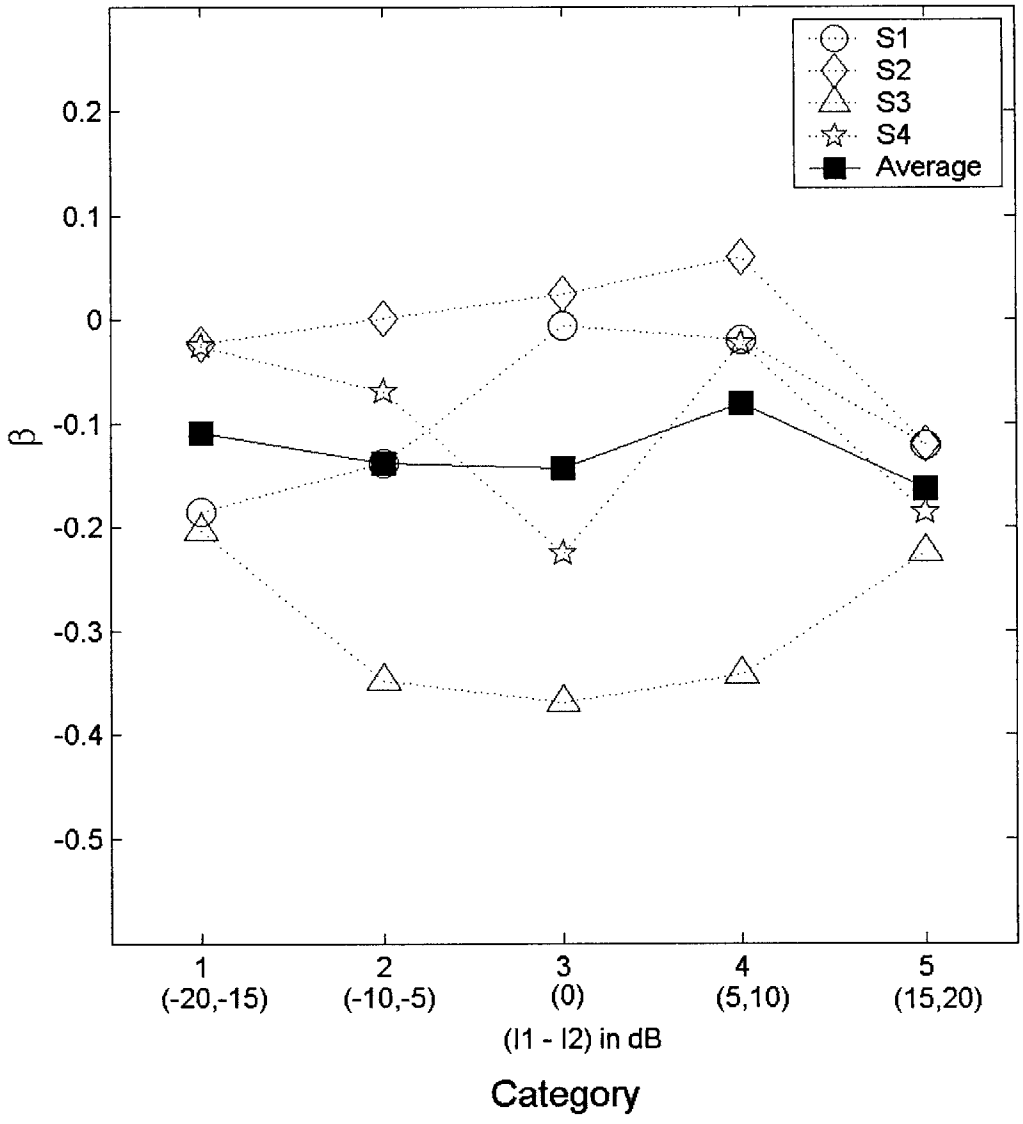

Fig. 7-9. β averaged across |SOA| for each subject and averaged across both |SOA| and subjects as a function of amplitude difference.

A two-way ANOVA was performed on the d′ values of each subject using amplitude difference and |SOA| as the two factors. The results of the ANOVA are shown in Appendix B for each subject. As expected, the factor of |SOA| was highly significant for each subject. The effect of the amplitude difference was significant at $p>0.008$ level.

The duration difference range (defined as the subtraction of the duration of the stimulus with a later onset from the duration of the stimulus with an earlier onset, D_1-D_2) was divided into 7 categories: ({-750, -700, -600, -550, -500, -450}, {-400, -350, -300}, {-200, -100, -50}, {0}, {50, 100, 200}, {300, 350, 400}, {450, 500, 550, 600, 700, 750} msec). The duration difference was calculated for every possible pairing of durations of the two stimuli in each trial (Table 7-5). The distribution of each sub-range class is shown in Table 7-6. For each sub-range class, a confusion matrix was derived and the corresponding d′ was calculated.

Table 7-5. Mapping between the duration difference and the duration pair of the two stimuli in each trial. The rows represent the durations of the first stimulus and the columns represent the durations of the second stimulus (in msec).

$D_1 \backslash D_2$	50	100	200	400	500	600	800
50	0	-50	-150	-350	-450	-550	-750
100	50	0	-100	-300	-400	-500	-700
200	150	100	0	-200	-300	-400	-600
400	350	300	200	0	-100	-200	-400
500	450	400	300	100	0	-100	-300
600	550	500	400	200	100	0	-200
800	750	700	600	400	300	200	0

Table 7-6. Duration-difference distribution for each sub-range.

Category	1	2	3	4	5	6	7
Sub-range	{-750, -700, -600, -550, -500, -450}	{-400, -350, -300}	{-200, -100, -50}	{0}	{200, 150, 100, 50}	{400, 350, 300}	({750, 700, 600, 550, 500, 450}
Distribution	6/49	7/49	8/49	7/49	8/49	7/49	6/49

The value of d′ for each sub-range is shown as a function of duration difference in Fig. 7-10. Each of the four panels of the figure represents the results for one of the four subjects. For each subject, different symbols represent different absolute values of SOA (filled symbols represent the three largest values of |SOA|, and unfilled symbols represent the three smallest values). The duration difference appears to have no clear effect on performance for S1, S2 and S4. For S3, performance decreases as the difference in duration between the two stimuli ($D_1 - D_2$) increases at all values of |SOA|.

Individual values of β, presented in Fig. 7-11, indicate a similar range from roughly -0.5 to +0.5 for subjects 1, 2, and 4 and a range of -0.5 to 0 for S3.

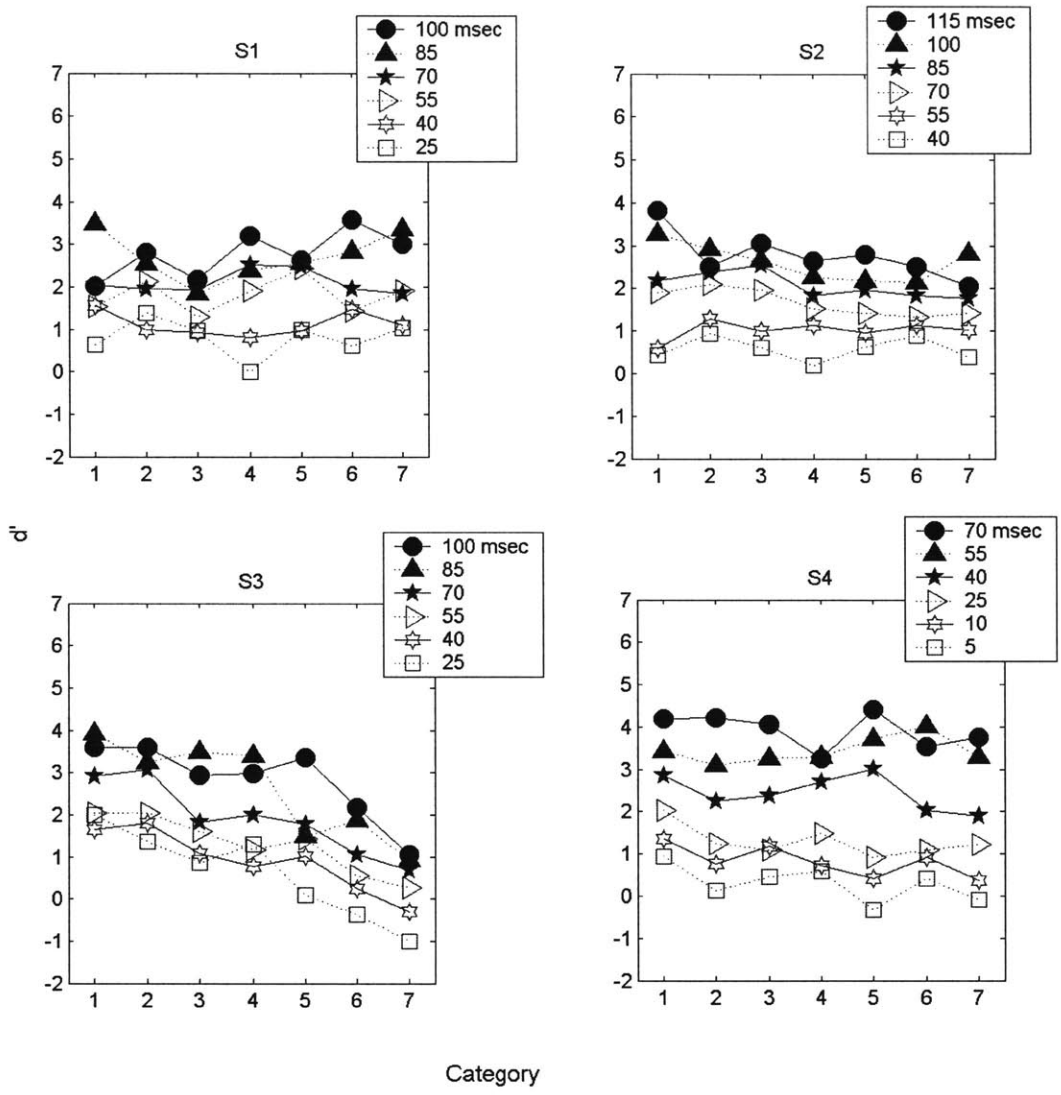

Fig. 7-10. d′ versus category in duration difference (in msec) for each subject and each |SOA| (in msec). See Table 7-6 for definition of category 1 through 7.

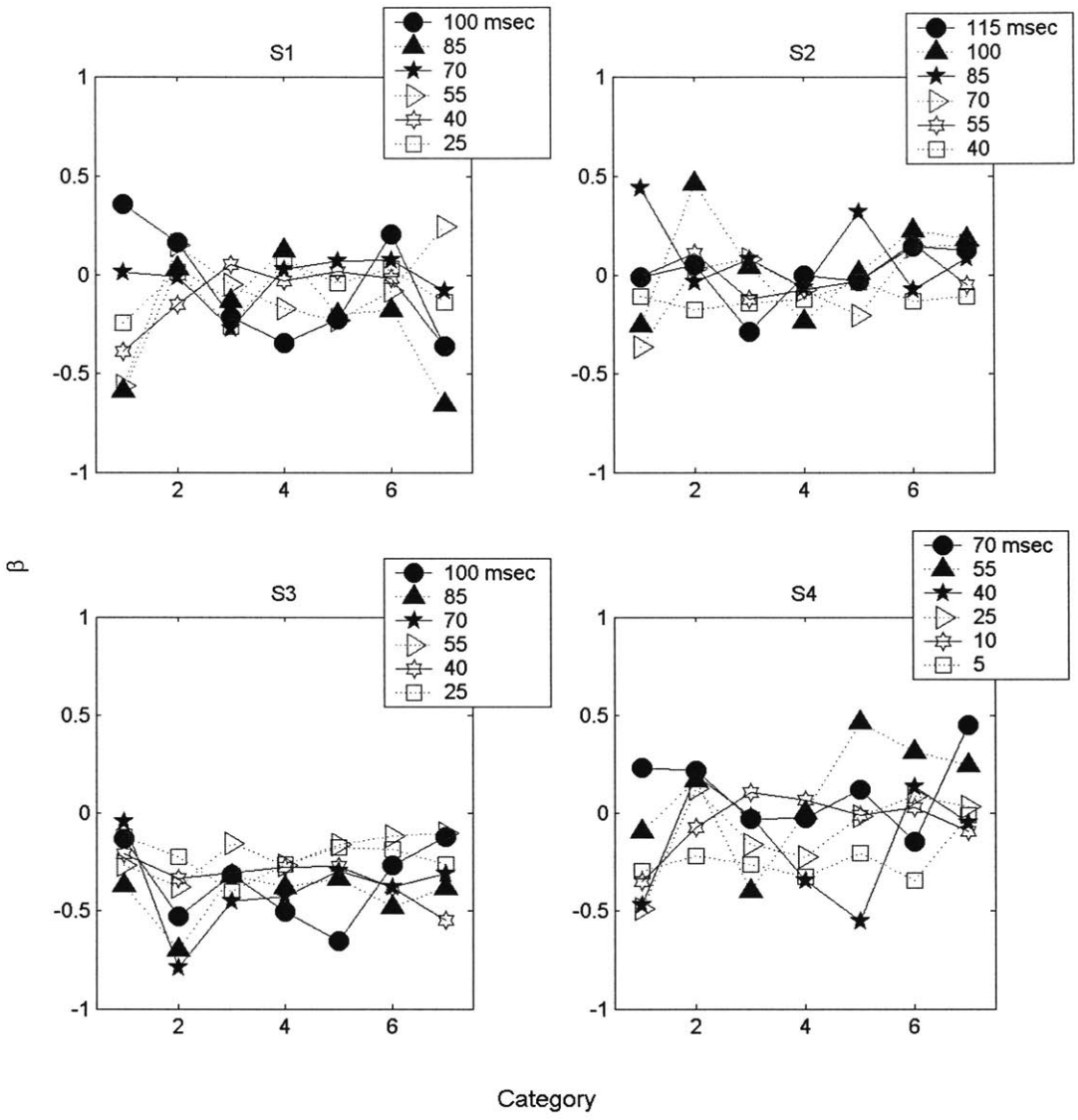

Fig. 7-11. β versus category in duration difference (in msec) for each subject and each |SOA| (in msec). See Table 7-6 for definition of category 1 through 7.

Values of d′ averaged across |SOA| are shown in Fig. 7-12 as a function of duration difference for each subject (represented by 4 different unfilled symbols), as well as the d′ averaged both across |SOA| and four subjects (filled squares). As in Fig. 7-10, the effect of duration difference is negligible for S1 and S2 and S4, but appears to have a systematic effect on the performance of S3.

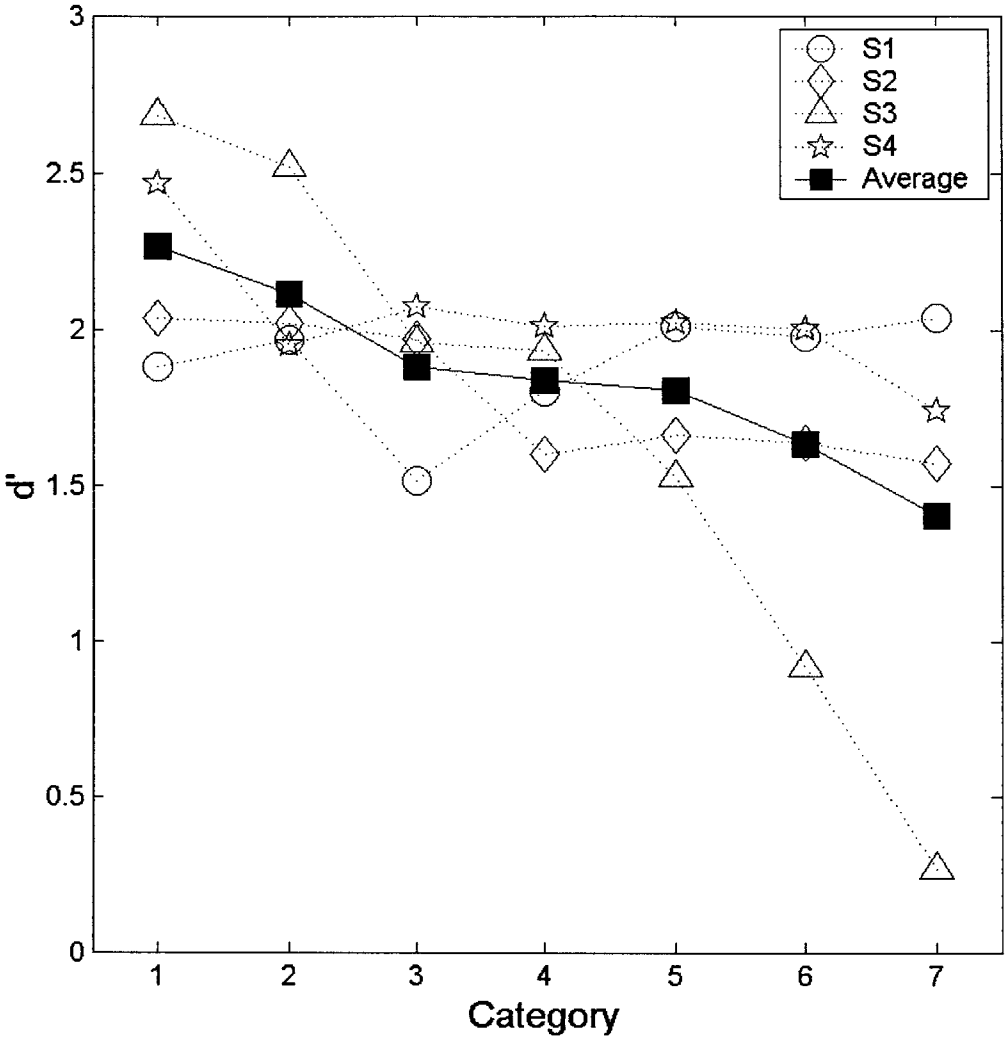

Fig. 7-12. d′ averaged across |SOA| versus category for each subject and across subjects.

Values of β averaged across |SOA| (see Fig. 7-13) indicate very little bias across the range of duration differences for S1, S2, and S4 and small negative bias for S3.

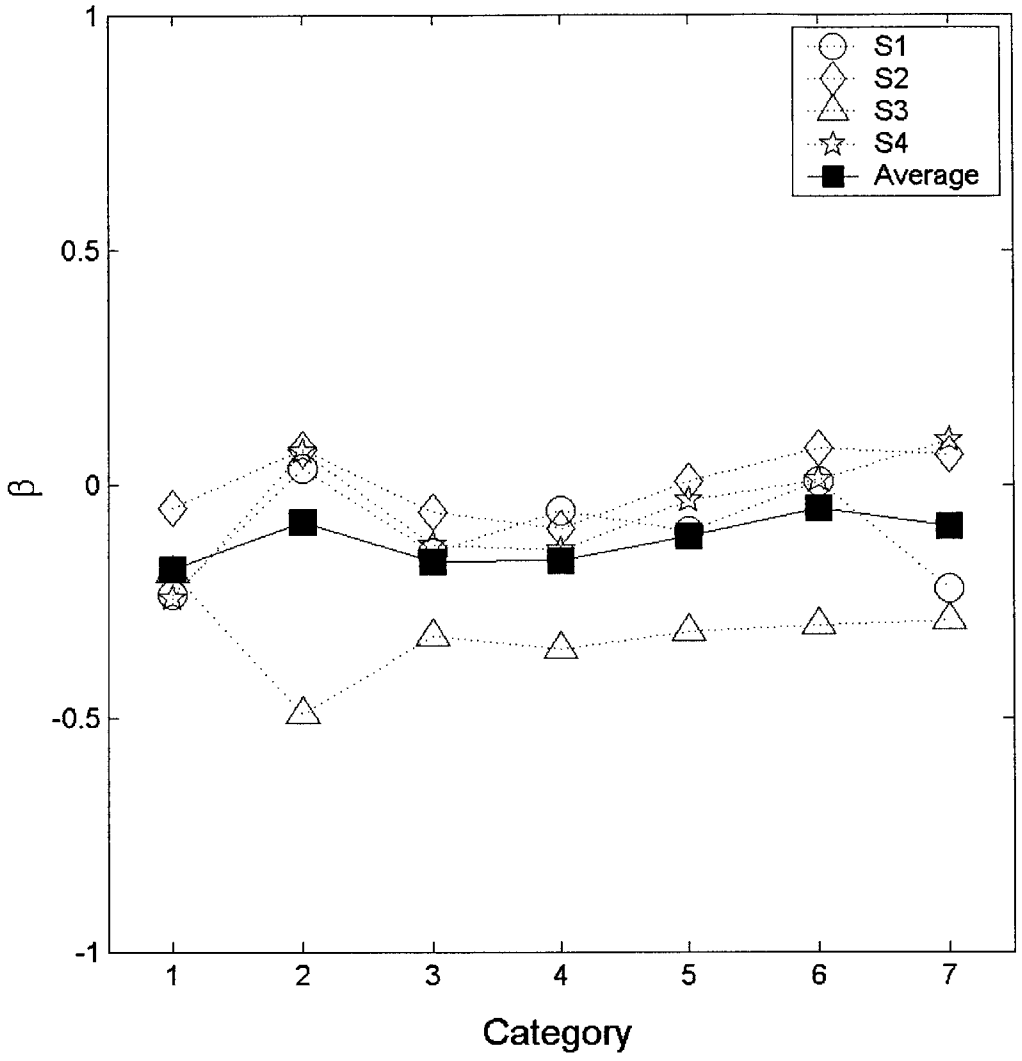

Fig. 7-13. β averaged across |SOA| versus category for each subject and across subjects.

A two-way ANOVA was performed using duration difference and |SOA| as the two factors for each of the four subjects. The results of the ANOVA are shown in Appendix C for each subject. The effect of the duration difference was significant only for S3 at p=0.003 level.

7.3 Pair-Wise Consonant Discrimination

Results were analyzed separately for each of the five replications of the experiment (see Table 6-5 for a summary of token set and feedback setting used in each replication). Under each replication, the results of each experimental run (4 subjects × 8 consonant pairs × 3 modalities) were summarized in terms of a 2×2 stimulus-response confusion matrix (see Table 6-6 for a general example). These matrices were used to compute values of d′ and β.

A summary of overall performance in each of the five replications of the experiment is provided in Figs. 7-14 (d′) and 7-15 (β) for individual subjects and for averages across subjects. For each subject, values of d′ and β were averaged across the eight consonant pairs for each of the three modalities (L, T, L+T) under each replication.

For each subject under each of the five replications, a clear and consistent effect was observed for modality. The d′ for L was near 0 for each subject under each replication indicating that performance was at chance level. This result is consistent with the well-known observation that lipreading carries little if any information about voicing. Performance under the modalities of T and L+T was similar within a given subject. Inter-subject variability was observed on T and L+T. Performance on T or L+T ranged from d′ of roughly 3.0 for S1 to roughly 1.5 for S2. The ability to discriminate voicing under modalities T and L+T is improved significantly over modality L for each subject. Averaged over subjects and replications, d′ values were roughly 0.09 for L alone, 2.4 for T alone, and 2.4 for L+T.

The effects of token set and feedback on average performance can also be examined from the data shown in Fig. 7-14. There seems to be a slight improvement in

performance in replication 3 (with "training" set and no feedback) compared to performance in replications 1 and 2 (with "training" set and feedback). The improvement due to training seems to saturate quickly. No apparent difference is found for no-feedback performance between the "test" set (replications 4 and 5 without feedback) and the "training" set (replication 3 without feedback). This result indicates that training conducted with a large set of tokens (multiple repetitions in three vowel contexts from 2 speakers) was sufficient for generalization to a "fresh" set of utterances by the same speakers in the same vowel contexts in a simple discrimination task.

For individual subjects, β was in the range of -0.3 to +0.3. S2 and S3 show some indication of different magnitudes of β dependent on modality. Averaged across subjects, β appears to be minimal and ranged from -0.15 to +0.05 across replications.

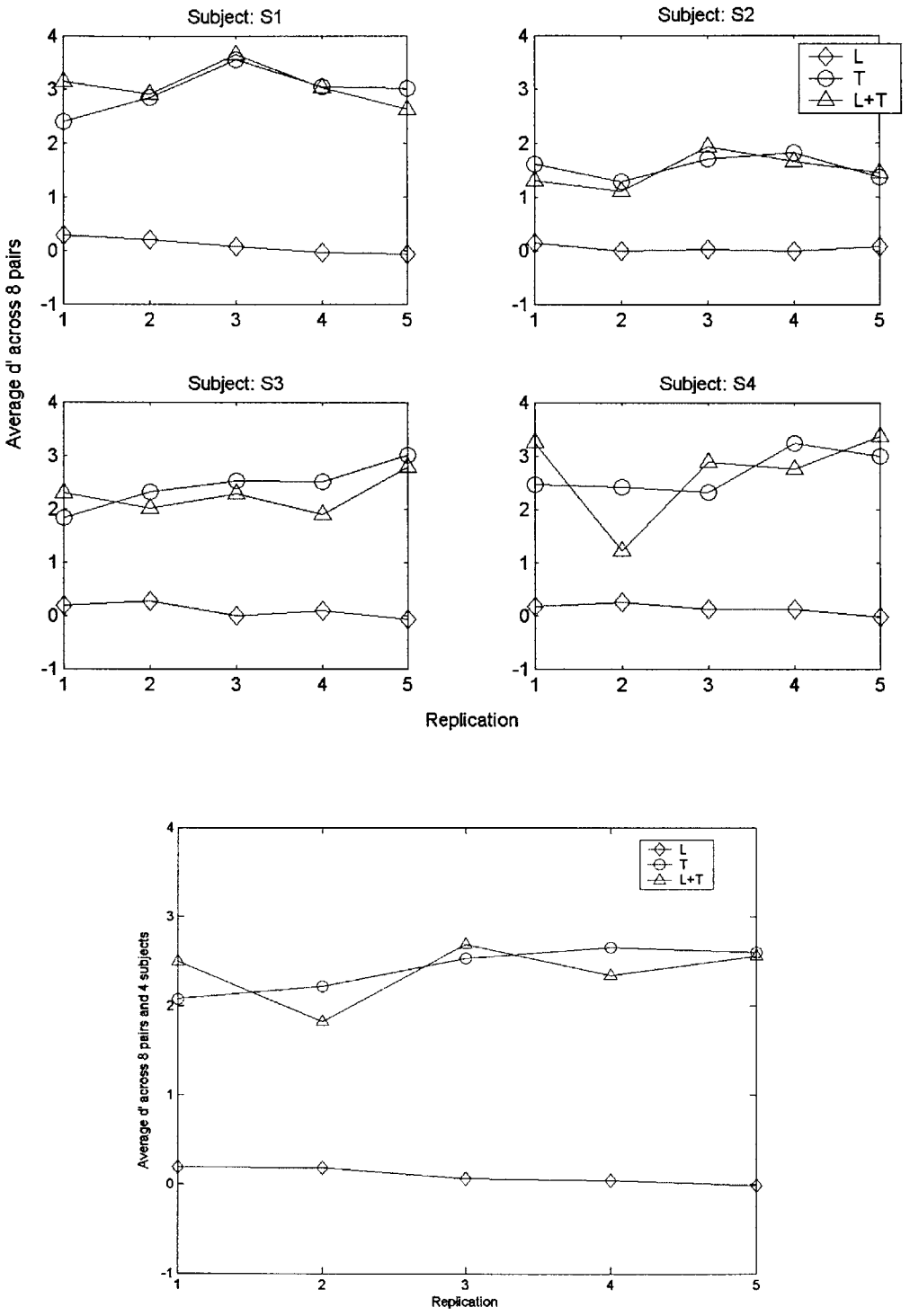

Fig. 7-14. Mean and individual d′ across eight pairs versus replications under three modalities for the four subjects.

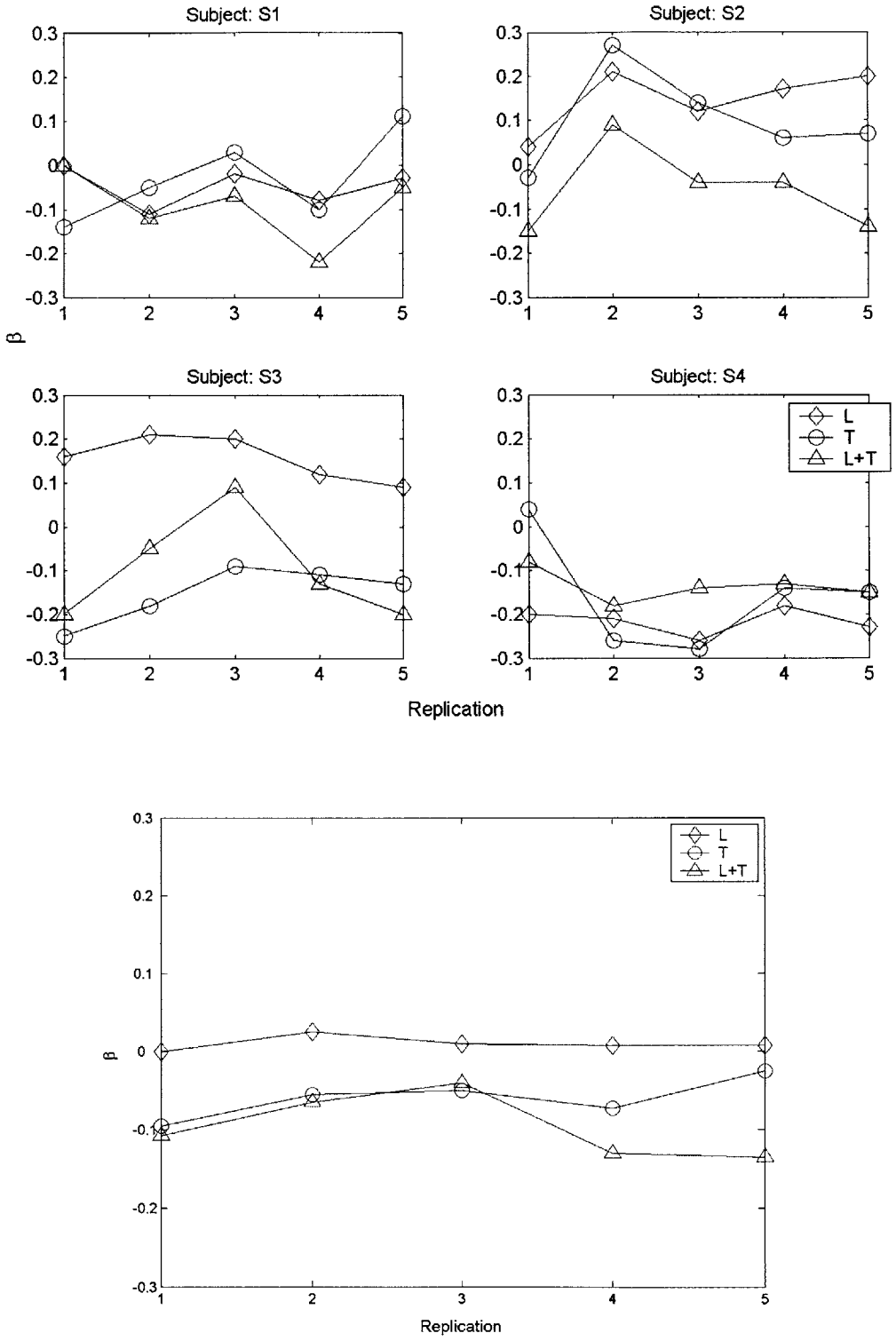

Fig. 7-15. Mean and individual β across eight pairs versus replications under three modalities for the four subjects.

Values of d′ averaged over replications 4 and 5, and across the four subjects, are shown in Fig. 7-16 for each of the eight pairs under each of the three modalities. A clear and consistent effect was observed for modality. The d′ for L was near 0 for each of the 8 pairs indicating that performance was at chance level. Performance under the modalities of T and L+T was similar within each pair: difference in d′ never exceeded 0.5. Inter-pair variability was observed on T and L+T. Performance ranged from d′ of roughly 3.3 for the pair /sh-zh/ to roughly 1.6 for the pair /ch-j/.

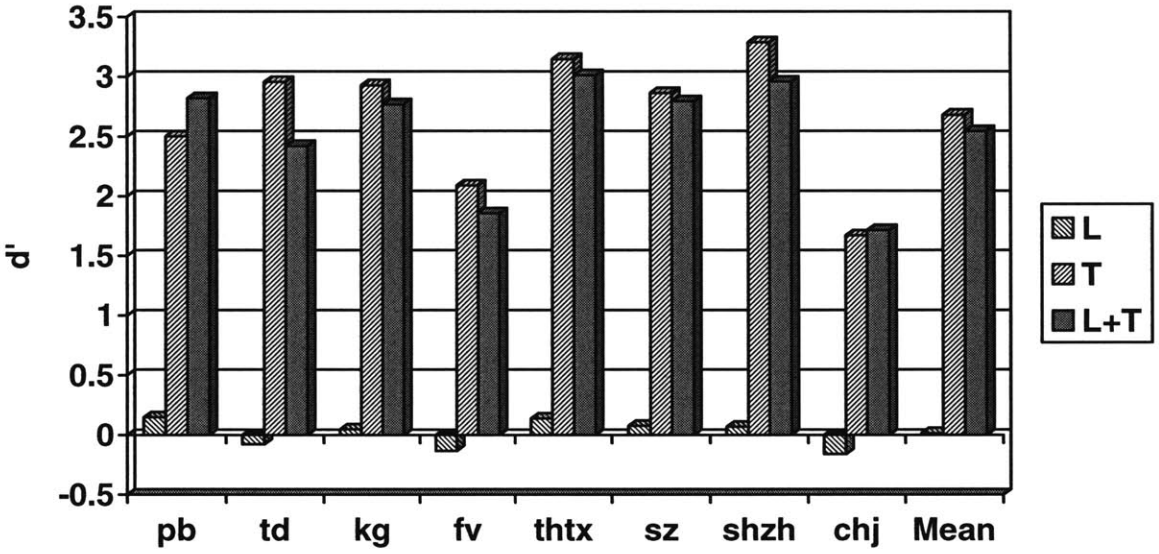

Fig. 7-16. Values of d′ averaged across the four subjects for conditions of no-feedback with the test tokens (replications 4 and 5) as a function of consonant pairs and modalities.

A three-way ANOVA was carried out on the data for each subject using the three main factors of modality, replication and pair. The full results of the ANOVA are given in Appendix D. The results indicate that each of the three main factors was significant at

a level of $p<0.05$ for each subject. The following interaction terms were significant at $p<0.05$: modality × replication for S3 and S4; modality × pair for S1 and S4; and replication × pair for S3. A post-hoc Scheffe analysis was also conducted to examine the significance of differences within pairs of means on each of the main factors. For the factor of modality, L < T = L+T for each of the four subjects. For the factor of replication, the first or second replication was found to be significantly lower than one of the later replications in S3 and S4. In addition, a significant improvement was noted between replication 4 and 5 for S3. For the factor of pairs, the only significant difference noted involved comparisons between the scores on /ch-j/ or /f-v/ with higher-scoring contrasts.

The perceptual discrimination results are compared with the acoustic measurements and the perceptual results from the temporal-onset-order experiment. The observed d' in the pair-discrimination experiment is plotted as a function of the predicted d' for an ideal observer (based on acoustic measurement of Envelope-Onset Asynchrony (EOA) in 3-vowel context, see Chap. 5.5) in Fig. 7-17. For each pair, perceptual data are averaged over the 4 subjects for the two modalities T and L+T. The observed d' is substantially lower than the d' calculated from the EOA measurements. This finding is reasonable given that the prediction of d' from the EOA measurements assumes an ideal observer, while in the perceptual experiment, sensory noise and memory noise are assumed to limit performance. The correlation of values of d' based on acoustic versus perceptual measures is relatively high (0.67 and 0.73 for T and L+T, respectively). The least two discriminable pairs (/f-v/ and /ch-j/) in the discrimination experiment are also

the two pairs with the lowest predicted value of d′ on the basis of the EOA measurements.

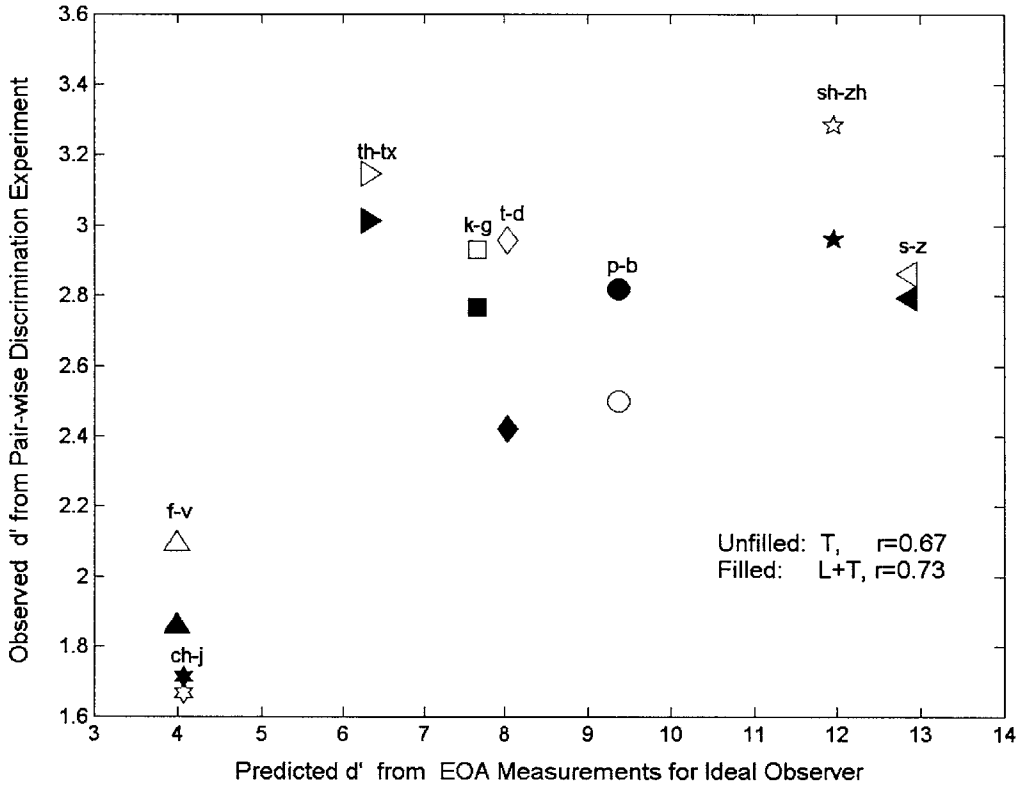

Fig. 7-17. Scatter plot of averaged d′ from pair-discrimination versus d′ predicted from EOA.

The correlation coefficient for each individual subject between predicted d′ and perceptual d′ across the eight pairs is shown in Table. 7-7. Correlations are lowest for S2 (0.04 for T and 0.32 for L+T) and highest for S4 (0.66 for T and 0.56 for L+T).

Table 7-7. Correlation coefficients between predicted d′ and perceptual d′ in pair-wise discrimination for individual subjects under two modalities T and L+T.

	S1	S2	S3	S4
T	0.45	0.04	0.59	0.66
L+T	0.49	0.32	0.50	0.56

Averaged d′ across pairs is plotted in Fig. 7-18 as a function of temporal onset-order threshold (|SOA| for d′=1) for individual subjects (with different symbols) under the two modalities T (unfilled symbols) and L+T (filled symbols). The performance of the perceptual discrimination results is consistent with that of the temporal order experiment: the most sensitive subject in temporal onset order is the subject with the best performance in discrimination experiment, and vice versa. The correlation coefficients under conditions T and L+T are -0.66 and -0.77 respectively.

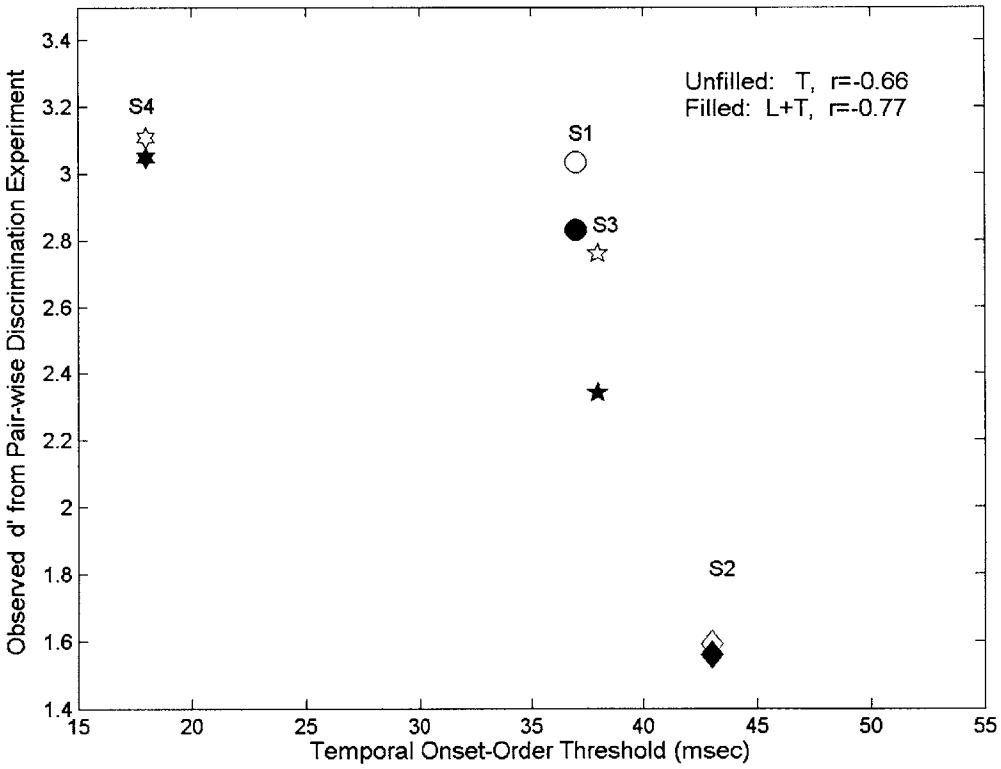

Fig. 7-18. Scatter plot of averaged d′ from pair-discrimination versus temporal onset-order threshold.

The correlation coefficient for each individual pair between observed d′ and temporal onset-order threshold (|SOA| for d′=1) is shown in Table 7-8. The pair /s-z/ yielded correlations of –0.96 (T) and –0.99 (L+T); somewhat lower correlations were observed across the remaining pairs.

Table 7-8. Correlation coefficients between SOA and perceptual d′ in pair-wise discrimination for individual pairs under two modalities T and L+T.

	p-b	t-d	k-g	f-v	th-tx	s-z	sh-zh	ch-j
T	-0.52	-0.77	-0.59	-0.42	-0.85	-0.96	-0.52	-0.65
L+T	-0.76	-0.43	-0.52	-0.85	-0.95	-0.99	-0.75	-0.75

7.4 16-Consonant Identification

Percent-correct scores were calculated separately for each of the seven replications of the experiment (see Table 6-7 for a summary of token set and feedback setting used in each replication). Each of the four panels of Fig. 7-19 represents results for one of the four subjects. For each subject, data are shown for L (diamonds), T (circles), and L+T (triangles). Each data point represents the percent-correct score of identification performance across the four runs (320 trials) collected at that particular replication. Results averaged across the four subjects are also shown in Fig. 7-20. Chance performance on this task is 6.25%.

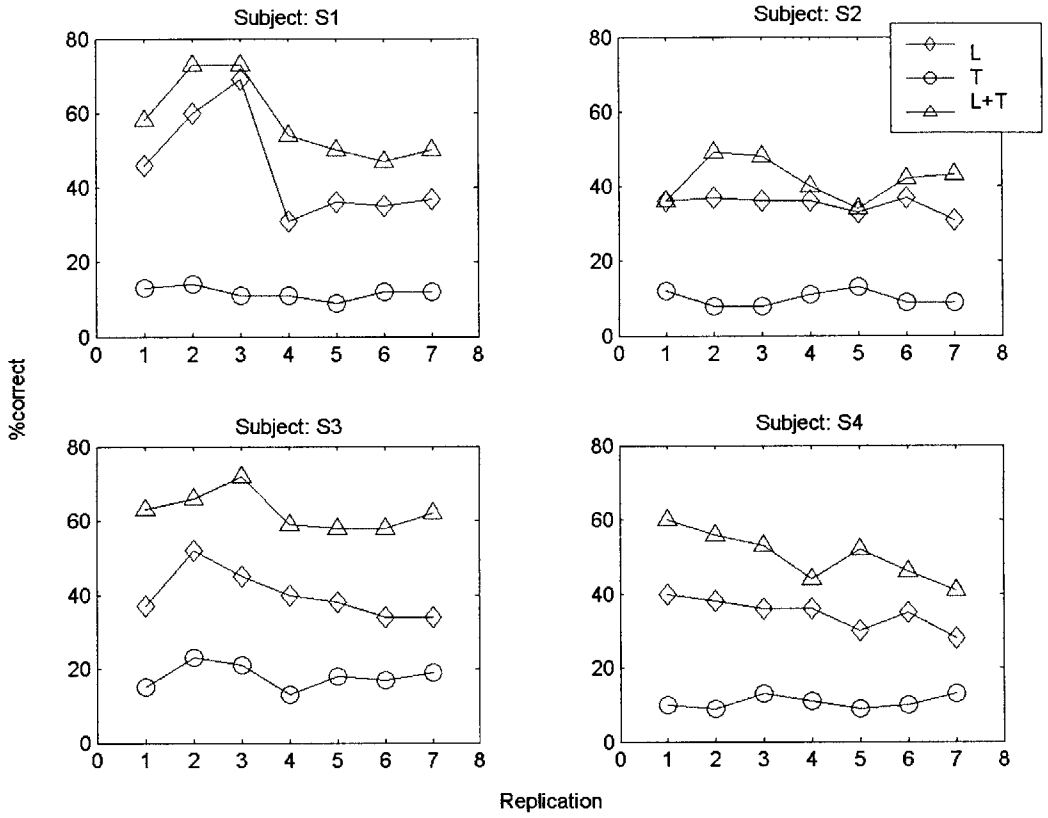

Fig. 7-19. Scores in %-correct versus replication number for individual subjects.

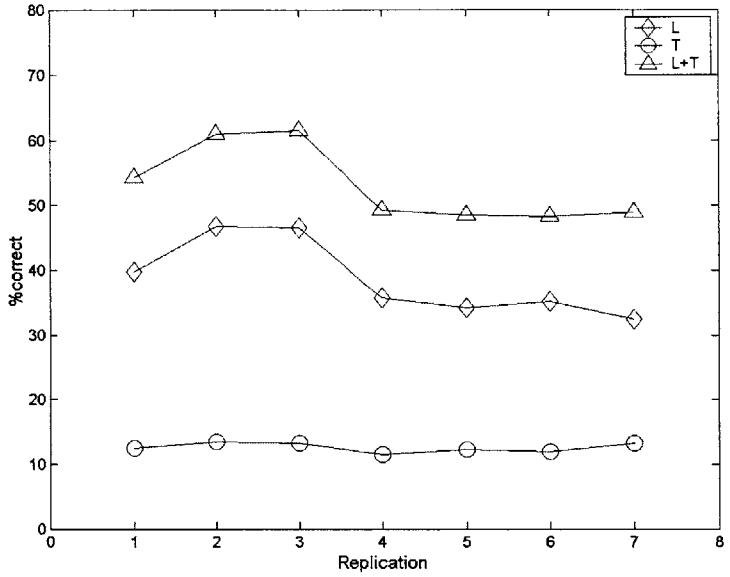

Fig. 7-20. Mean %-correct scores averaged across subjects versus replication number.

The effect of modality is clear and consistent across all subjects: the performance in percent correct under the modality T is always the lowest; the performance under modality L always lies in the middle; and the performance under the modality L+T is always the highest. All subjects demonstrated some type of learning effect from replication 1 to 2 (with feedback) for the two modalities L and L+T except S4. The improvement from replication 1 to 2 averaged roughly 7 percentage points for these two modalities. The inter-subject variance of this learning effect ranges from -2 to 15 percentage points for modality L, and -4 to 15 percentage points for modality L+T. The learning effect is negligible for modality T. No apparent effect of feedback is observed (i.e., replication 3 is similar to replication 2). The performance using the "testing" set without feedback (replications 4 through 7) decreased by roughly 12 percentage points from the performance using the "training" set without feedback (replication 3). This

result suggests that the subjects may have learned to take advantage of some artfactual types of cues following repeated exposure to the "training" set both with and without feedback. Inter-subject variance is most obvious under the modality L+T: performance varies from roughly 40% correct for S2 to roughly 60% correct for S3.

Confusion matrices for the set of 16 consonants were compiled for each subject and each condition using the no-feedback replications with the "test" tokens (i.e., replications 4 to 7). Average percent-correct scores over the final four replications are presented for each subject under each modality in Fig. 7-21. Modality has a clear and consistent effect for each subject. The %-correct score is roughly 30% under modality L, roughly 10% under modality T (slightly better than chance), and roughly 50% under the modality L+T. Inter-subject variability is negligible for L (where performance ranged from 32% to 35%) and T (where performance ranged from 11% to 17%). Inter-subject variability, however, is somewhat larger for L+T, where it ranged from roughly 40% to 60%-correct across subjects. The improvement in L+T over L alone averaged roughly 15 percentage points, and ranged from 7 to 23 percentage points across subjects.

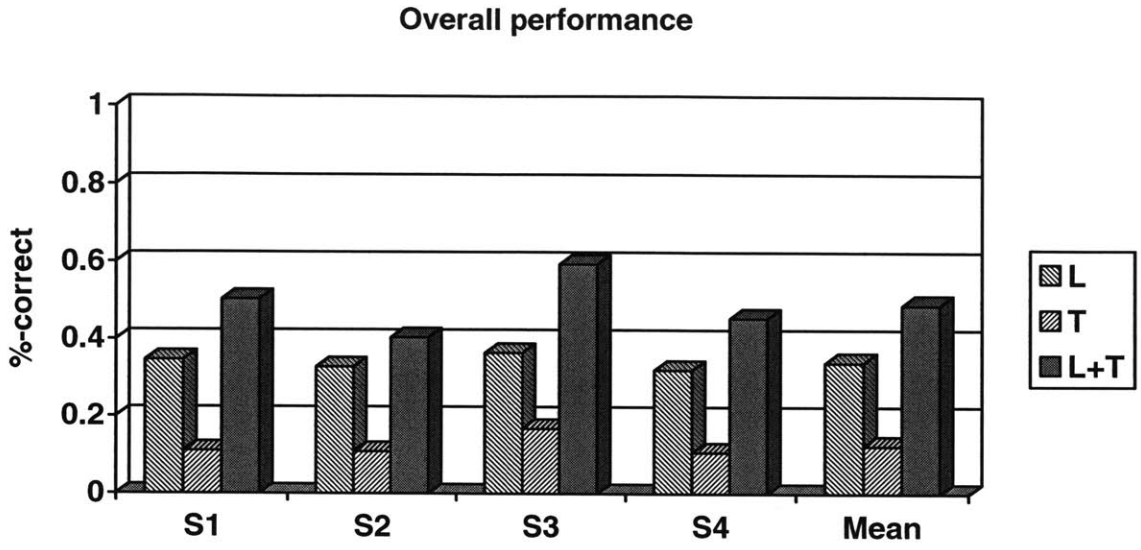

Fig. 7-21. Mean and individual performance in %-correct for 16-consonant-identification experiment under three modalities across the replications with no-feedback and with the "test" tokens (i.e., replications 4 to 7).

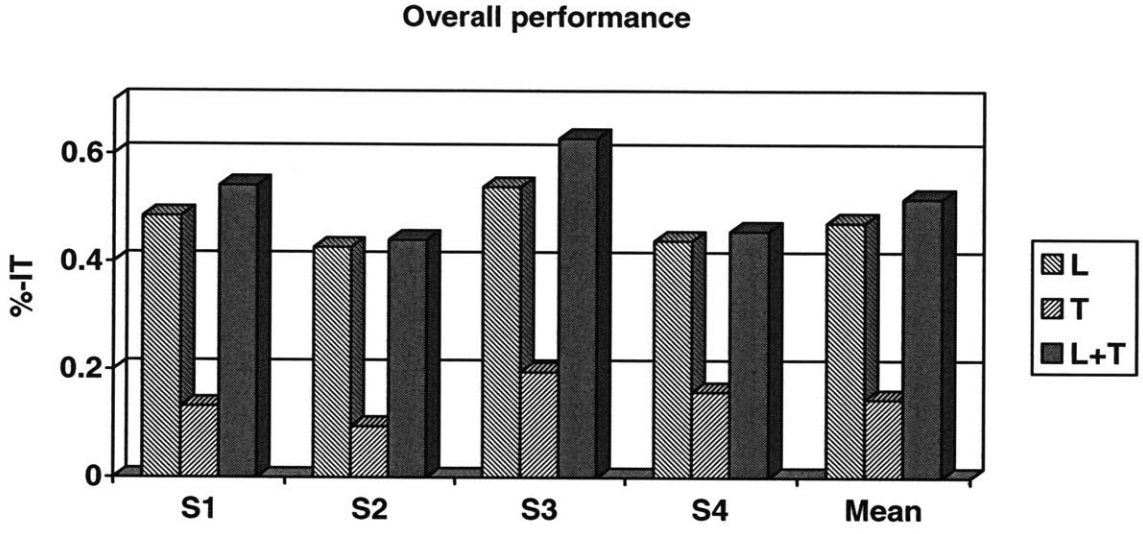

Fig. 7-22. Mean and individual performance in %-IT for 16- consonant-identification experiment under three modalities across the replications with no-feedback and with the "test" tokens (i.e., replications 4 to 7).

Measurements of percentage overall information transfer (IT) were also calculated from the confusion matrices and are shown in Fig. 7-22 for each subject and modality. The pattern seen in Fig. 7-22 differs from that seen in Fig. 7-21 primarily in that the %-IT scores for L and L+T are more similar than are the %-correct scores. The improvement in %-IT performance observed for L+T compared to L alone averaged roughly 4 percentage points (ranging from 1 to 9 percentage points across subjects).

The 16 consonants can be uniquely identified by the three articulatory features of voicing, place, and manner (see Table 7-9). The feature voicing is related to the state of the vocal folds which vibrate during the production of voiced consonants and do not vibrate when voiceless consonants are produced. The feature place refers to the location of the major constriction of the vocal tract during consonant articulation. Place is represented here by locations of constriction: labial, dental, alveolar, and velar. The feature manner refers to the type of constriction made by the articulators. Stop consonants are produced with a complete closure in the oral tract, fricatives with a close articulation that produces turbulence, and affricates are made by a stop releasing into a fricative. Performance of the identification of individual features in %-correct and %-IT are analyzed below.

Table 7-9. The features of the 16 consonants. "+" indicates that the consonant owns that feature.

		p	b	t	d	k	g	f	v	th	tx	s	z	sh	zh	ch	j
Voicing	Voiced		+		+		+		+		+		+		+		+
	Voiceless	+		+		+		+		+		+		+		+	
Place	Labial	+	+					+	+								
	Dental									+	+						
	Alveolar			+	+							+	+				
	Velar					+	+							+	+	+	+
Manner	Stops	+	+	+	+	+	+										
	Fricatives							+	+	+	+	+	+	+	+		
	Affricates															+	+

The reception of feature "voicing" was analyzed by reducing the 16×16 confusion matrix into a 2×2 confusion matrix (see Table 7-10). In the table, N_{11} is the sum of the entries in the 16×16 confusion matrix for which both the stimuli and their corresponding responses are in the voiceless set (e.g, p, t, k, f, th, s, sh and ch). N_{12} is the sum of the entries in the 16×16 confusion matrix for which the stimuli are in the voiceless set, while their corresponding responses are in the voiced set (e.g, b, d, g, v, tx, z, zh and j). The entries of N_{21} and N_{22} are derived similarly. Measures of %-correct and %-IT were derived from matrices produced for each subjects.

Table 7-10. 2×2 confusion matrix derived from 16×16 confusion matrix of the 16-consonant identification. Reception of the feature "voicing" was derived from this matrix.

Stimulus\Response	Voiceless	Voiced
Voiceless	N_{11}	N_{12}
Voiced	N_{21}	N_{22}

The mean and individual performance in %-correct (refer to Equation 6-1 for calculation of %-correct) on voicing under each of the three modalities is shown in Fig. 7-23. Modality has a clear and consistent effect on the performance of voicing discrimination. The performance under modality L is roughly 50%, i.e., chance performance. This is consistent with the performance observed in pair discrimination and with the knowledge that the feature voicing, characterized by the activities of the vocal folds, is almost invisible through lipreading. The performance under modality T averaged roughly 80%, representing an improvement of roughly 30 percentage points in the delivery of voicing relative to lipreading alone. The performance under L+T averaged roughly 73% and was 7 percentage points lower than the modality T. Inter-subject variance is negligible for the modality L (standard deviation of 1 percentage point). Greater inter-subject variance was observed for the modalities T and L+T where standard deviation across subjects was roughly 6 - 7 percentage points. For modality T, performance ranged from 70% - 85% across subjects, and for modality L+T, performance ranged from 64% - 79%. Although the overall performance in the T modality was near chance, performance on the voicing feature was well above chance level.

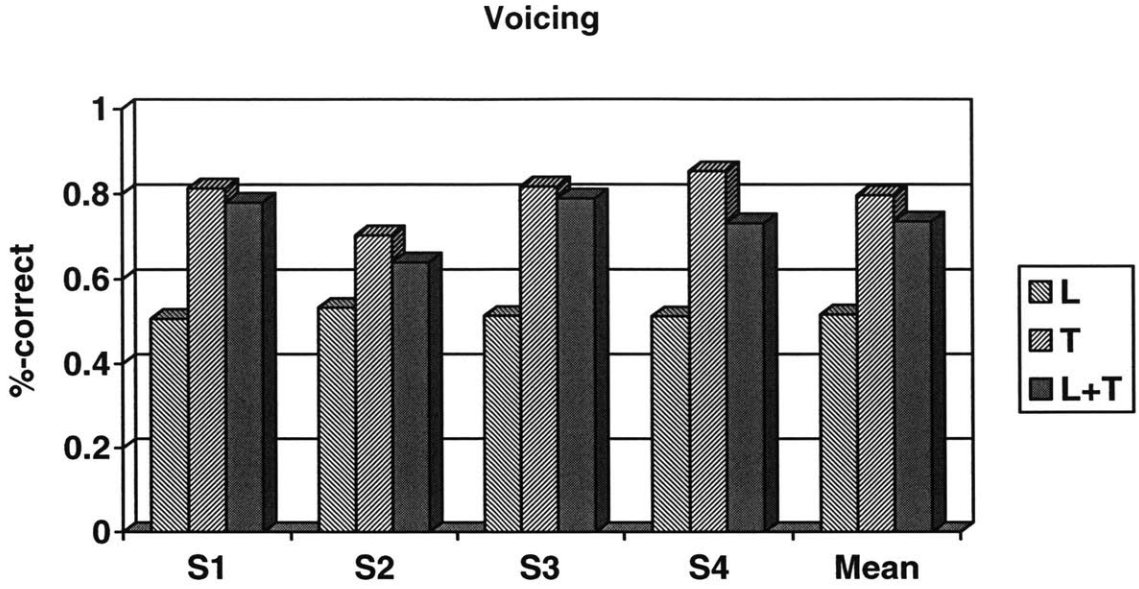

Fig. 7-23. Mean and individual performance in %-correct of voicing for 16-consonant-identification experiment under three modalities.

Measurement of percentage feature IT was also calculated from the 2×2 confusion matrix and shown in Fig. 7-24 (refer to Equation 6-2 for calculation of %-IT).

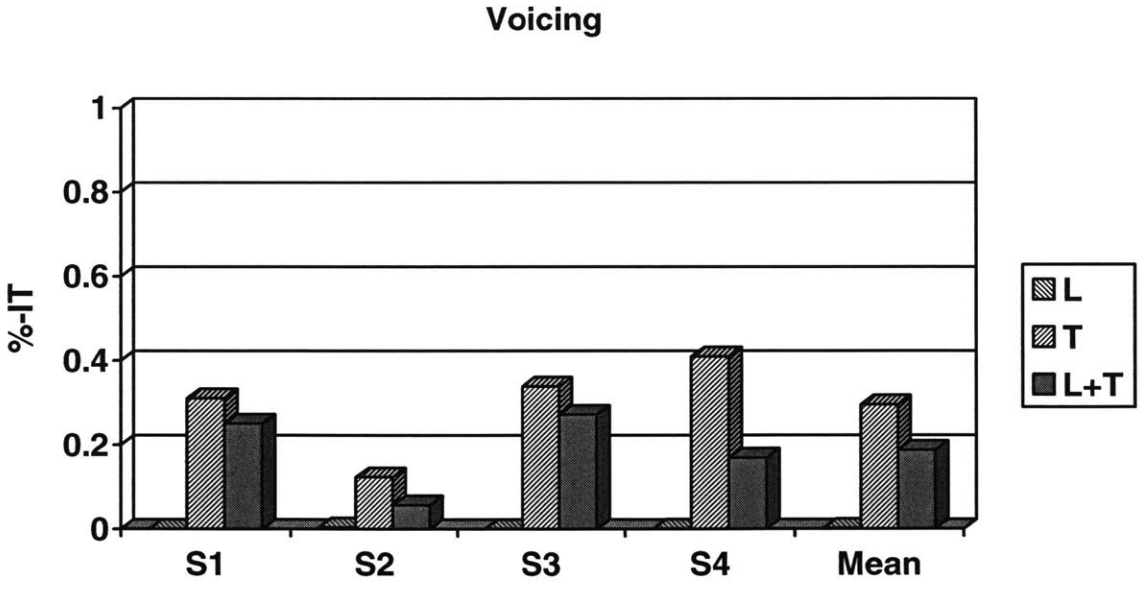

Fig. 7-24. Mean and individual performance in %-IT of voicing for 16- consonant-identification experiment under three modalities.

As in Fig. 7-23, modality has a consistent effect. The %-IT feature was close to 0 for modality L and averaged 30% and 20% for T and L+T, respectively. The pattern of inter-subject variance is similar to that observed for %-correct. The variance under L is small (standard deviation 0.1%), while variances for T and L+T are larger (standard deviation 12 and 10 percentage points, respectively). The absolute improvement of T or L+T relative to L in both measurements is similar, in the range of 20% to 30%. According to the %-correct measurement, the remaining potential-improvement for voicing communication is about 20% to 30%. However, the potential room for improvement increases to 70% - 80% based on the measurement of % feature IT. The distortion between the achieved improvement and potential improvement of these two measurements is due mainly to their nonlinear relationship (see Fig. 7-25). From this Figure, it can be seen that when the %-correct increases from 50% to 70%, the increase in %-IT is only roughly 10 percentage points (0 to 12%); that a 20% increase in %-correct in the region of 70% to 90% leads to a roughly 40 percentage-point-increase in %-IT (12% to 55%); and that a 10 percentage-point-increase in %-correct from 90% to 100% leads to more than a 40 percentage-point-increase in %-IT.

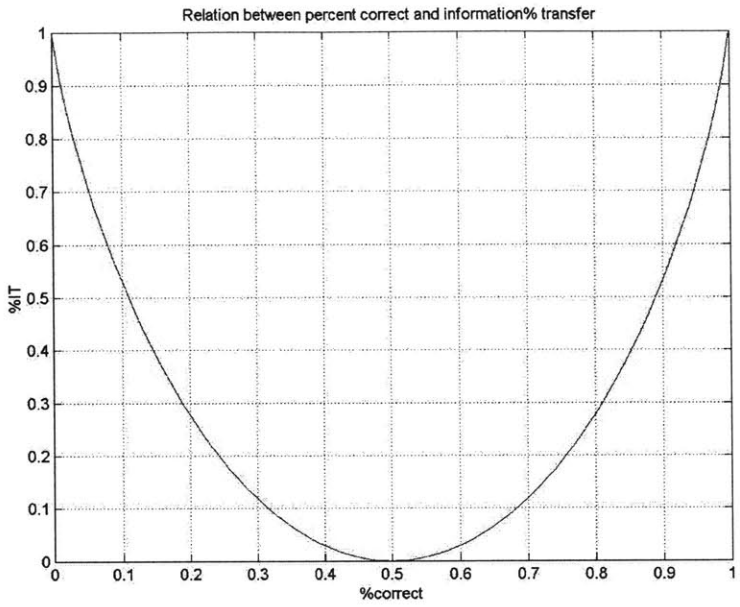

Fig. 7-25. Relationship between %-IT and %-correct for a 2×2 symmetric confusion matrix.

The feature manner classifies the 16 consonants into 3 groups: stops, fricatives and affricates. A 3×3 confusion matrix (see Table 7-11) was derived from the 16×16 confusion matrix for each subject under the three modalities. N_{ij} is the sum of the entries in the 16×16 confusion matrix for which the stimuli belong to group i, and their corresponding responses belong to group j. The %-correct (PC) was calculated using Equation 6-1. Chance performance is 33%.

Table 7-11. 3×3 confusion matrix derived from the 16×16 confusion matrix for the feature "manner".

Stimuli\Response	Stops	Fricatives	Affricates
Stops	N_{11}	N_{12}	N_{13}
Fricatives	N_{21}	N_{22}	N_{23}
Affricates	N_{31}	N_{32}	N_{33}

The mean and individual performance in %-correct of manner under each of the three modalities is shown in Fig. 7-26. Modality has a clear and consistent effect on the performance of manner. The performance under T is the lowest, roughly 48%. The performance under L is similar to that under L+T, roughly 72%. The inter-subject variance is negligible for L and L+T, with standard deviation of 2 to 3 percentage points. The inter-subject variance is greater for T, with standard deviation of 7 percentage points. Performance on T alone ranged from 42% to 57% correct across subjects.

Measurement of %-IT for the manner feature was also calculated from the 3×3 confusion matrix using Equation 6-2, and is shown in Fig. 7-27. As in Fig. 7-26, modality has a significant effect. Performance under T averaged 3%, and performance under L and L+T was similar, averaging roughly 30%. Performance ranged from 22% to 37% across subjects under modality L, from 22% to 42% for L+T, and from 0 to 8% for T. The standard deviations across subjects under the three modalities are 7, 3 and 9 percentage points for L, T and L+T respectively.

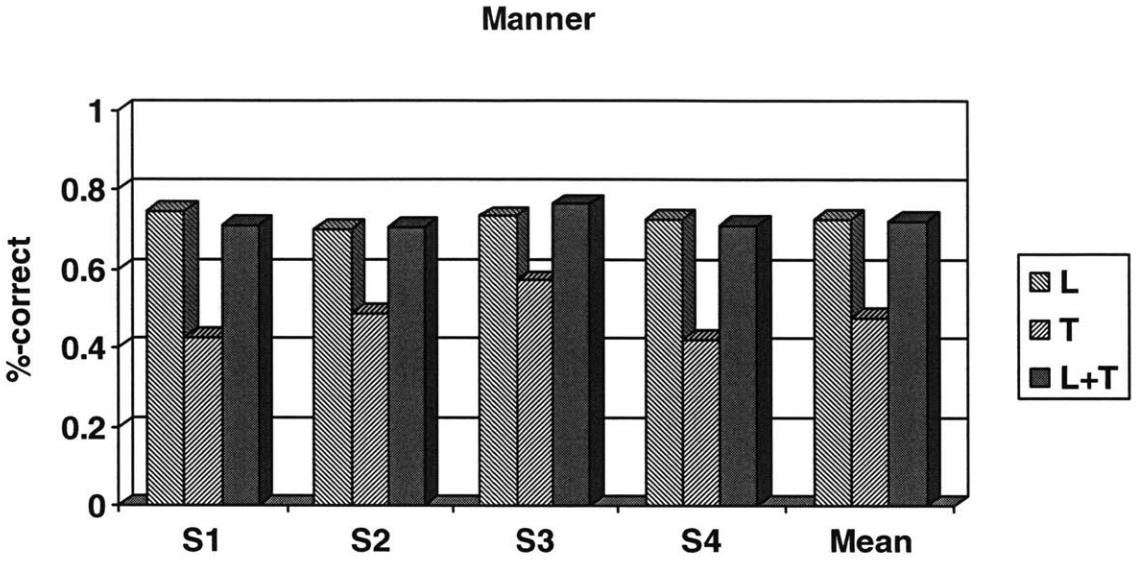

Fig. 7-26. Mean and individual performance in %-correct of manner for 16-consonant-identification experiment under three modalities.

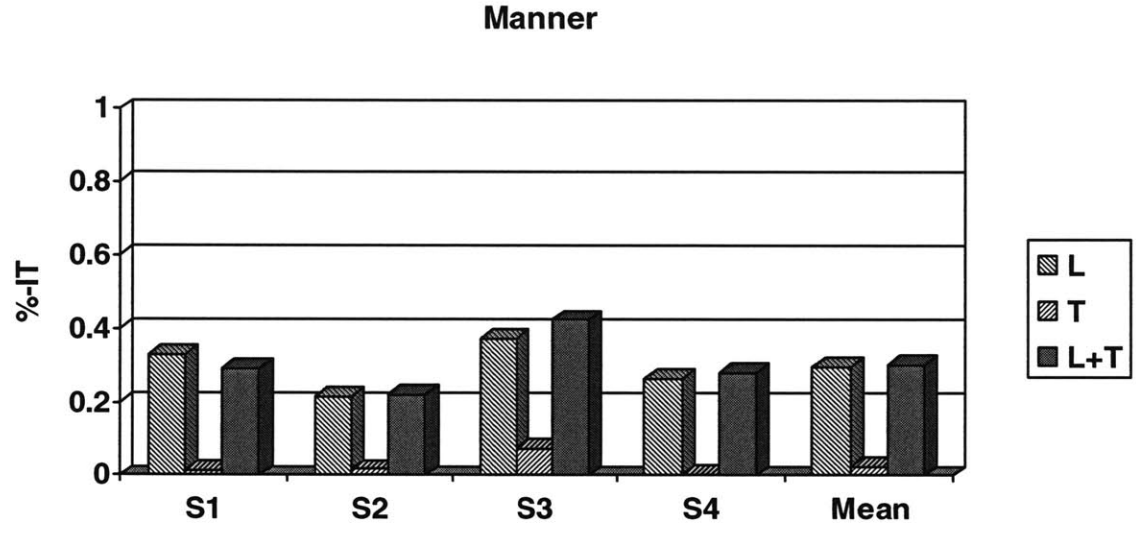

Fig. 7-27. Mean and individual performance in %-IT for manner for 16- consonant-identification experiment under three modalities.

The feature place classifies the 16 consonants into 4 groups: labial, dental, alveolar and velar. A 4×4 confusion matrix (see Table 7-12) was derived from the 16×16

confusion matrix for each subject under the three modalities. N_{ij} is the sum of the entries in the 16×16 confusion matrix for which the stimuli belong to the group i, and their corresponding responses belong to group j. Chance performance is 25%.

Table 7-12. 4×4 confusion matrix derived from the 16×16 confusion matrix for the feature "place".

Stimulus\Response	1. Labial	2. Dental	3. Alveolar	4. Velar
1. Labial	N_{11}	N_{12}	N_{13}	N_{14}
2. Dental	N_{21}	N_{22}	N_{23}	N_{24}
3. Alveolar	N_{31}	N_{32}	N_{33}	N_{34}
4. Velar	N_{41}	N_{42}	N_{43}	N_{44}

The mean and individual performance in %-correct reception of the place feature under the three modalities is shown in Fig. 7-28. Percent-correct was calculated using Equation 6-1. Modality has a clear and consistent effect on the performance of manner in the order T<L=L+T, with average scores of 28% on T (near chance performance), 81% on L, and 83% on L+T. Standard deviation over subjects was roughly 5 to 6 percentage points for L and L+T, and 2 percentage points for T. Scores for L ranged from 76% to 88%, from 77% to 90% for L+T, and from 26% to 30% for T.

Measurement of %-IT for the feature of place was also calculated using Equation 6-2 from the 4×4 confusion matrix and shown in Fig. 7-29. Performance under L and L+T was similar, averaging roughly 68%. Performance across subjects ranged from 61% to 80% for L (with standard deviation of 9 percentage points), and from 56% to 81% for L+T (with standard deviation of 11 percentage points). Basically no information was transmitted on the place feature for T alone.

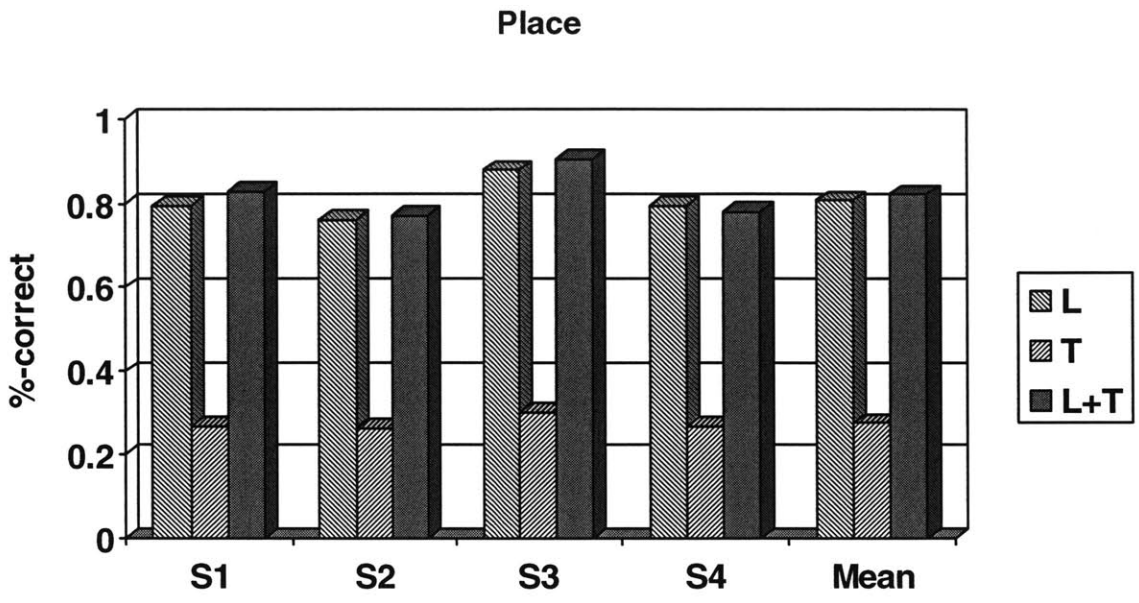

Fig. 7-28. Mean and individual performance in %-correct of place for 16-consonant identification experiment under three modalities.

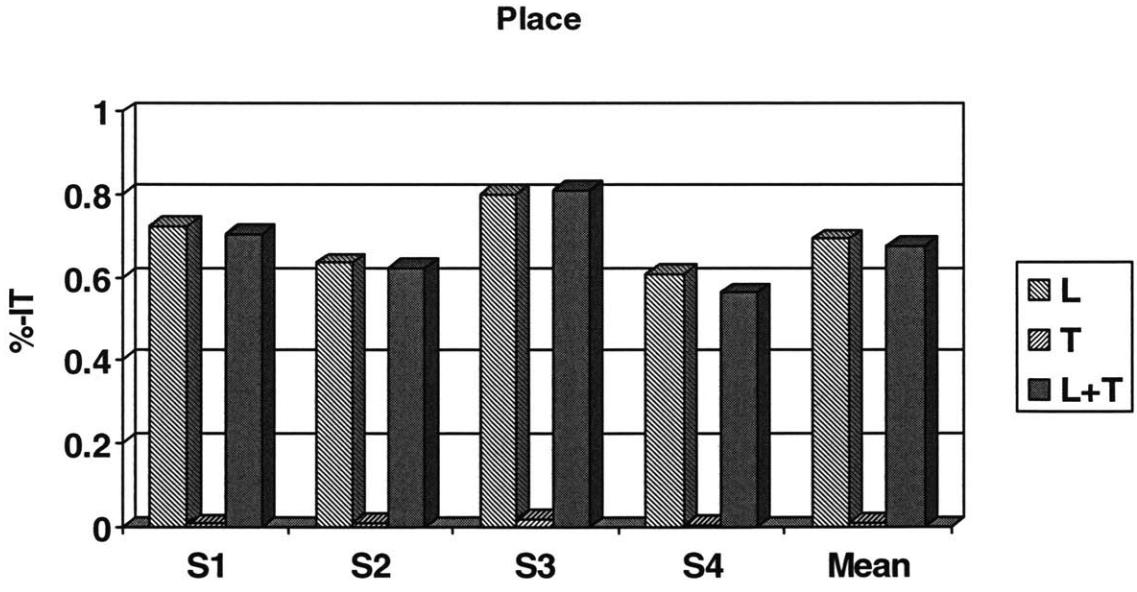

Fig. 7-29. Mean and individual performance in %-IT of place for the 16- consonant identification experiment under three modalities.

Measurements of %-IT averaged across subjects for the three features (voicing, manner, and place) are shown in Fig. 7-30. The features of manner and place are best transmitted through the modality L, and the feature voicing is best transmitted through the modality T. The features transmitted under the combined modality L+T appear to be a simple combination of the information transmitted through T and L alone.

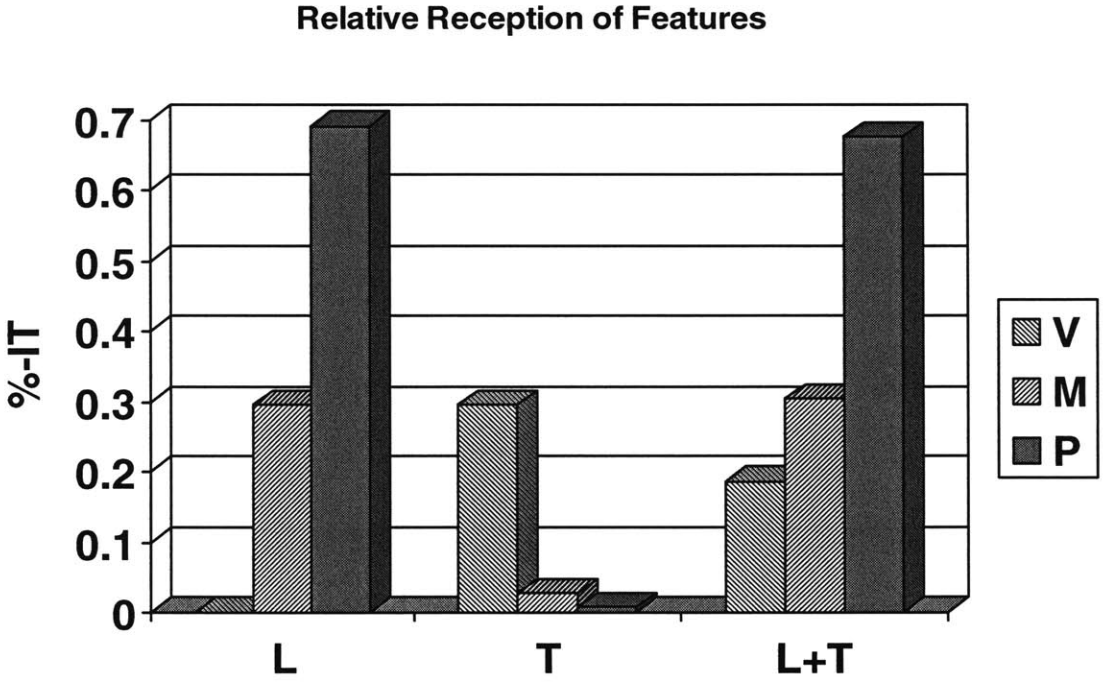

Fig. 7-30. Mean performance for %-IT of the three features for the 16- consonant identification experiment under three modalities.

7.5 CUNY Sentences

Results of the CUNY sentence test under the two modalities of lipreading alone (L) and lipreading supplemented by touch (L+T) are presented in Fig. 7-31 for each subject. Each data point represents the percent-correct score for each list presented under each modality. No learning over time was observed for any of the subjects. For each subject, performance appears to be similar for L and L+T. Averages and standard deviations of performance across the 54 lists presented under each modality for each subjects are shown in Table 7-13. Performance on the two modalities never differed by more than 3 percentage points. In addition, performance appeared to be fairly stable over time. Inter-list standard deviation ranged from 4-12 percentage points across subjects.

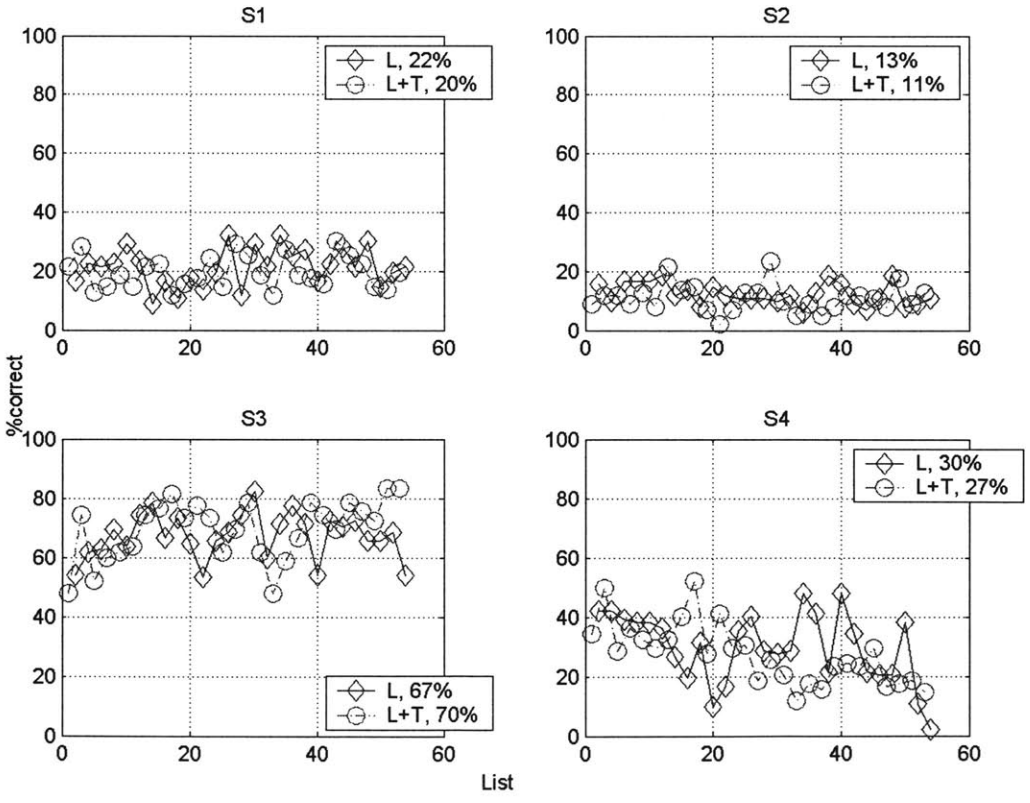

Fig. 7-31. Score in %-correct versus lists for individual subjects.

Table 7-13. Mean and standard deviation of the four subjects under modality L and L+T.

	S1	S2	S3	S4
Mean (L)	22%	13%	67%	30%
s.d. (L)	6.6	3.8	7.8	12.0%
Mean (L+T)	20%	11%	70%	27%
s.d. (L+T)	5.6%	4.8%	10.2%	10.3%

Chapter 8

Discussion

8.1 Absolute Threshold Measurements

An understanding of the response of the skin to ymechanical sinusoidal stimulation as a function of frequency is fundamental to the design of tactual communication systems. Numerous studies have examined the sensitivity of the skin to different frequencies of vibration. Measurements have been made at a number of different body sites, have employed different procedures, and have included a variety of stimulus parameters (such as contactor area, duration, rise time, etc.). Despite such differences, however, most studies have observed that the skin is differentially sensitive to frequency of vibration. The underlying mechanism for this differential sensitivity is attributed to the mechanoreceptor systems that mediate tactual sensations.

Bolanowski et al. (1988) proposed a four-channel model (P, NP I, NP II, and NP III) of mechanoreception for human somatosensory periphery and discussed possible corresponding physiological substrates (four afferent fiber types: PC, RA, SA II, and SA I) for each channel. The properties of each channel are summarized in Table 8-1. The psychophysically measured detection thresholds for the thenar eminence are shown in Fig. 8-1, as well as the responses of each channel that demonstrate some degree of overlap. The thresholds of each separate channel were determined using psychophysical procedures selected to isolate the particular channel being studied. These techniques exploited the temporal and spatial summation properties of each channel as well as *a*

priori knowledge of the frequency region of maximal sensitivity (see Gescheider et al., 2002). Systems that exhibit temporal summation (P and NP II) were studied using long-duration signals (e.g., 1 sec), while those that do not demonstrate temporal summation (NP I) were studied using short-duration signals (e.g., 50 msec). Systems that demonstrate spatial summation (P) were studied using large contactor areas (e.g., 1.5 cm^2) whereas those that do not exhibit spatial summation (NPI, NPIII) were studied using small contactor areas (e.g., 0.008 cm^2). Off-frequency masking was also used to isolate the shape of the response curve within each region under the assumption that masking occurs only within, and not across, channels (Gescheider et al., 1982, Hamer et al., 1983).

Table 8-1. Properties of each of the four channels.

	P	**NP I**	**NP II**	**NP III**
Frequency Range (Hz)	40 ~ 800	10 ~ 100	15 ~ 400	0.4 ~ 100
Temperature Dependency	Yes	No	Yes	Yes
Temporal Summation	Yes	No	Yes	Not known
Spatial Summation	Yes	No	Not known	No
Neurophysiological substrates	Pacinian corpuscle fiber (PC)	Rapidly adapting fiber (RA) with Meissner corpuscle receptor	Not known (possibly SA II fiber)	Slowly adapting fiber (SA I)
Perception	Vibration	Flutter	Vibration	Pressure

The absolute sensitivity at a particular frequency is determined by the channel having the lowest threshold at that frequency. The threshold curve, shown in Fig. 8-1, consists of five different regions. In Region 1 (NP III), the threshold is independent of

frequency at very low frequencies (below 2 Hz). In Region 2 (NP I), the threshold decreases slowly as frequency increases (2 to 30 Hz). In Region 3 (P), threshold decreases sharply as frequency increases until it reaches its maximum sensitivity in the range from 200 to 300 Hz. In Region 4 (P), the threshold increases as frequency increases up to about 600 Hz. In Region 5 (>1000 Hz), valid threshold measurements are difficult to obtain.

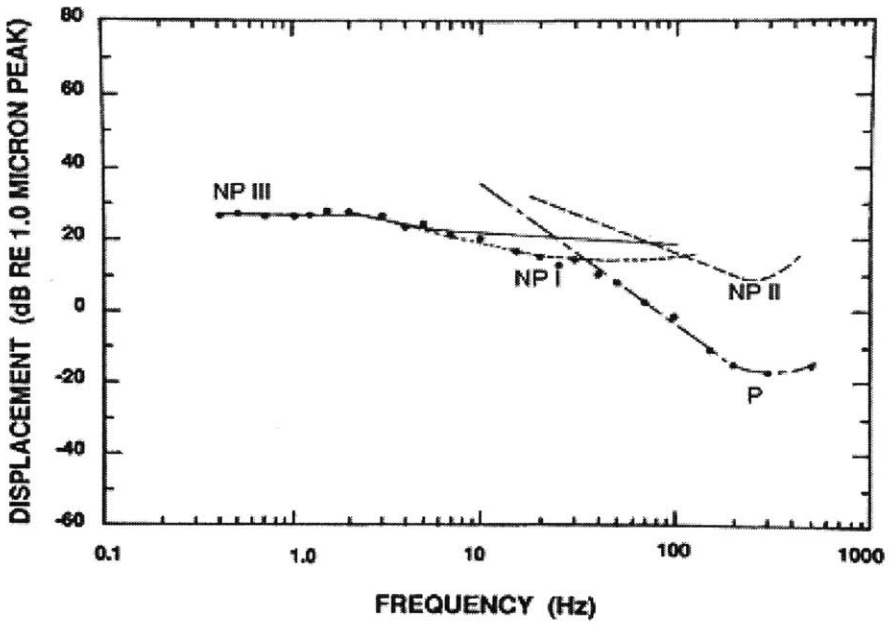

Fig. 8-1. The four-channel model of vibrotaction. The four curves represent the threshold-frequency characteristics of the P, NP I, NP II, and NP III channels, as determined by psychophysical measurements. The data points represent psychophysically measured detection thresholds. (From Bolanowski et al., 1988).

A comparison of the threshold measurements of the current study with those from previous studies is shown in Fig. 8-2. Those studies that were conducted on the fingertip

(Lamore et al., 1986 [LMK]; Rabinowitz et al., 1987 [RHDD]; Tan, 1996 [TAN]; Gescheider et al., 2002 [GBPV]) or thenar eminence (Bolanowski et al., 1988 [BGVC]) and used an adaptive forced-choice psychophysical procedure are used in the comparison. Parameters of each study are summarized in Table 8-2.

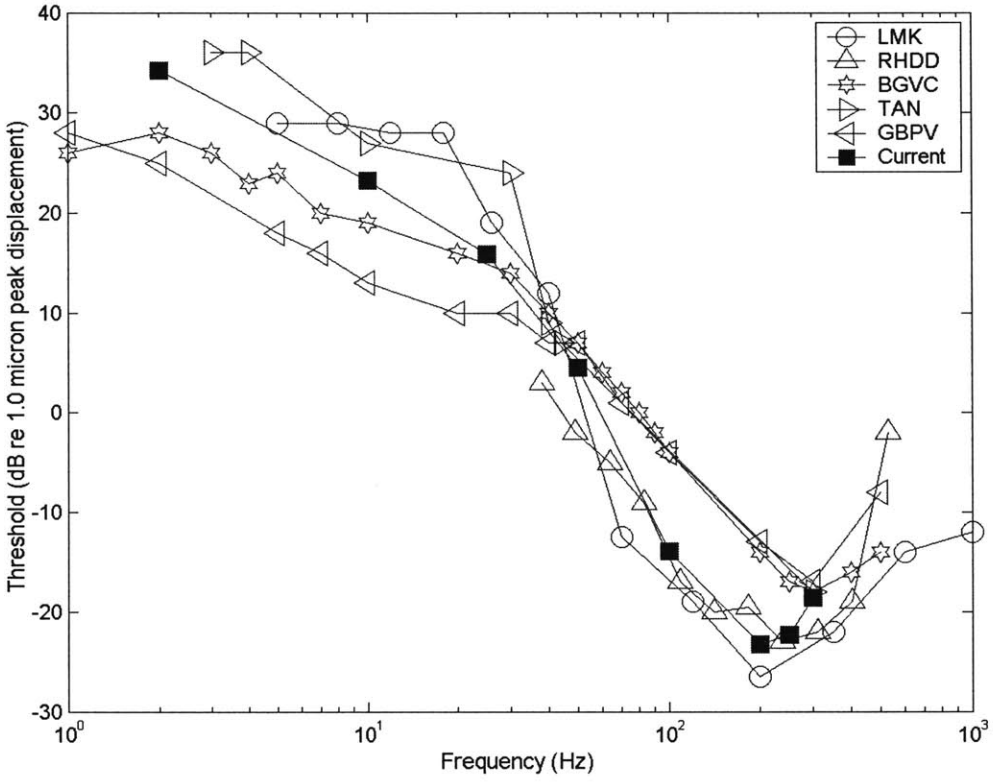

Fig. 8-2. Comparison of average threshold (index finger) of the present study with those from other studies that were performed on the fingertip or thenar eminence and used adaptive psychophysical procedures. Results from RHDD have been converted from "rms" to "zero to peak" by increasing the reported thresholds by 3 dB.

The level and frequency dependence of the thresholds measured in the present study are generally consistent with previous results. The studies BGVC and GBPV, using a static indentation of vibrator contactor of 0.5 mm into the skin, are more divergent from other studies. This may partly be due to the effects arising from the static indentation and

rigid surround employed in these two studies (and not in the remaining studies). The data can be approximated roughly by two straight lines with different slopes (see the left panel of Fig. 7-2). These slopes may reflect the characteristics of two different channels of Bolanowski's four-channel model. The shallower slope (from 2 to 25 Hz) reflects the sensitivity of NP I channel, and the steeper one (from 25 to 300 Hz) reflects the sensitivity of P channel. The expected flat portion of the NP III cannot be observed in the threshold curve for the current data, because frequencies below 2 Hz were not measured here.

Table 8-2. Summary of selected studies on tactual threshold measurements.

	Location	Frequency range (Hz)	Duration (msec)	ISI (msec)	Rise-fall (ms)	Contactor area (cm^2)	Procedure
Bolanowski et al. (1988)	Thenar eminence (static indentation of 0.5 mm)	0.4~500	700	1750	500	2.9	2I2AFC (3 down 1 up) => 75%
Rabinowitz et al. (1987)	Middle finger	40~500	500	500	50	0.38	2I2AFC
Lamore et al. (1985)	Left first phalanx of the middle finger	5~1000	100	NA	100	1.5	Adaptive forced-choice procedure
Tan (1996)	Left index	3~300	NA	NA	NA	≈0.5	1I2AFC (2 down 1 up) => 71%
Gescheider et al. (2002)	Right index fingertip (static indentation of 1 mm)	0.4~500	700	NA	500	0.72	2I2AFC (3 down 1 up) => 75%
Present Study	Left index fingertip	2~300	500	500	10	≈0.5	2I2AFC (2 down 1 up) => 71%

8.2 Temporal Onset-Order Discrimination

Temporal onset-order discrimination requires that subjects not only be able to discriminate two distinct stimuli, but that they also be able to differentiate the order in which they were presented. This ability is crucial to the recognition of acoustic signals such as speech and music. In speech, for example, temporal-order discrimination is necessary for the correct interpretation of the ordering of words within sentences, the ordering of individual speech components within words (mitts vs. mist), and the ordering of acoustic events that are important to the distinction of minimally contrasting speech sounds. In particular, the role of temporal-order ability has received special attention with regard to the discrimination of consonantal voicing (See Chapter 5, EOA or VOT measures). In addition, temporal-order discrimination threshold constrains the rate at which information can be transmitted through a sensory channel.

8.2.1 Comparison with Previous Research

Experimental studies of temporal-order discrimination ability are concerned with determining the stimulus-onset asynchrony (SOA) necessary for reliable discrimination of the temporal order of two distinct stimuli. Experiments have been conducted through different modalities including hearing, vision, and touch. A summary of previous studies of temporal-order discrimination relevant to the current study is provided in Table 8-3. The studies are described in terms of modality of stimulation, stimulus type, signal duration and rise-fall time, procedure, SOA at threshold, and other comments pertinent to the experimental outcome.

Table 8-3. Summary of selected studies of temporal-order discrimination.

Studies	Modality	Stimulus Type	Stimulus Parameters			Procedure	SOA at threshold	Comments
			Longer Duration (msec)	Shorter Duration (msec)	Rise-fall time (msec)			
Hirsh (1959)	Auditory (A)	5 pairs of sinusoids differing in frequency	500	500-SOA	20	Method of constant stimuli, repeated every 1.5 sec	20 ms for 75%- correct	Simultaneous offset of tones, to eliminate offset cue. However, energy and duration cue may contribute to the judgment of onset-order
		Tone vs. wideband noise	500	500-SOA	2, 7,15			
		Broadband noise vs. click	Longer noise	Brief click	7, 15			
		Click vs. tone	20, 50, 100	Brief	NA			
		Click vs. click	Brief	Brief	NA			Level of tones: 50, 90 dB
Hirsh and Sherrick (1961)	Auditory (A)	Pair of pulses differing in pitch, ear of stimulation, ear and pitch	0.1	0.1	NA	Method of constant stimuli, repeated every 1 sec	20 msec for 75%-correct	Offset information can contribute to the judgment of onset-order
	Visual (V)	Light flashes separated horizontally 18, 36, or 72 cm, or vertically 18cm	5	5	NA			
	Tactual (T)	Half-wave pulses (100Hz) to left and right finger tips	5	5	NA			
	Cross-modal:							
	A - V	Click vs. flash	5	0.1	NA			
	A - T	Click vs. tap	5	0.1	NA			
	V - T	Flash vs. tap	5	5	NA			

Pastore (1982)	Auditory (A)	1650 vs. 2350 Hz	CD+SOA	CD=10, 30, 100, and 300	0.5	Adaptive 2 down-1 up rule	4 ~ 12 (70.7%)	Simultaneous offset, duration and energy cues may contribute to temporal-order judgment
		1650 vs. 2350 Hz	CD+SOA	CD=10, 30, 100, and 300	0.5, 1, 2.5, 5.0, 10, 25, 50, 100		TOT rises with rising time for larger rising times	Thresholds increase as the common durations increased, as well as rising time.
		1650 vs. 2350 Hz + Trailing Tones (950, 1950, 2950 Hz)	CD+SOA	CD=30, 100, and 300			The trailers increased the threshold by roughly 4 to 5 msec for 30 msec signals	
Pastore (1983)	Auditory (A)	1800 vs. 2150 Hz 1600 vs. 2350 Hz 1600 vs. 2850 Hz 1100 vs. 2350 Hz 1100 vs. 2850 Hz	CD+SOA	CD=10, 30, 100, and 300	0.5	Adaptive 2 down –1 up rule	2 to 7 (71%)	Simultaneous onset, Measured stimulus-offset asynchrony threshold
Pastore and Farrington (1996)	Auditory (A)	5 msec click+10 msec silence+pair of tones (330 vs. 2200 Hz)	300 (2200 Hz)	300-SOA (330 Hz)		2I2AFC	35	
Jaskowski et al. (1990)	Auditory vs. visual	Auditory: click Visual: circular patch of light	100 or 30	5	NA	Adaptive	30	
Sherrick (1970)	Tactual	Pair of mechanical or electrotactile clicks	≈4 (mech) ≈0.4 (elec)	≈4 (mech) ≈0.4 (elec)	NA	Same as Hirsh & Sherrick (1961)	20 (bilateral) 25~35(ipsilateral)	Offset could be a potential cue
Marks et al. (1982)	Tactual	Pair of biphasic electrocutaneous pulses	0.5	0.5	NA	Adaptive 3 down – 1 up rule	>100 in most cases	
Craig and Baihua (1990)	Tactual	Horizontal vs. vertical patterns (230 Hz)	16	16	NA	2AFC	12 (same site) 68(ipsilateral) 62(bilateral)	Offsets may contribute to onset-order judgment
Eberhardt et al. (1994)	Tactual	Vibration vs. movement	733~900	236~788	NA	Method of constant stimuli	25~45	Details not available

Hirsh (1959) studied auditory temporal-onset-order discrimination for monaural presentation of pairs of stimuli differing in frequency, quality (tones vs. noise), duration, and both duration and quality. Temporal-onset order thresholds were in the range of 15 to 20 msec for 75%-correct performance across all conditions. Hirsh noted that the auditory temporal-order threshold is substantially larger than the fusion threshold (the minimum separation time of the two events that is required to differentiate them as successive instead of simultaneous) for two clicks (i.e., < 2 msec); therefore, he suggested that the process for judging temporal order is central to the peripheral auditory system. Hirsh noted that rise-time and duration might affect the threshold, but considered these to be secondary effects. Hirsh employed simultaneous signal offset for the two stimuli in a pair in order to eliminate offset cues; however, duration or energy cues may be available to the subject in this paradigm.

Hirsh and Sherrick (1961) carried out a series of experiments concerning temporal-order discrimination of pairs of clicks within and across the modalities of hearing, vision, and touch. Their results indicate that an |SOA| of approximately 20 msec was required for order judgments of 75% correct, a finding that was consistent across all three modalities as well as for cross-modal comparisons. They concluded that unlike judgments of simultaneity or successiveness (where temporal separations are dependent on the particular modality employed), temporal separation for judgment of order is longer and is independent of the sensory modality. These results have been interpreted as suggesting that simultaneity judgments are mediated at the peripheral sensory level (thus accounting for their modality dependence) and temporal-order judgments are mediated at

a more central cortical level (due to their modality-independence) (e.g., see Poppel, 1997; Wittmann, 1999).

Pastore (1982, 1983) investigated the effects of various stimulus parameters on auditory temporal-order judgments (both onset-order and offset-order) for pairs of tones. The parameters investigated included duration, separation in frequency of the pair of tones, rise time, relative intensity of the pair of tones, and presence of a trailing tone. Offset-order thresholds (Pastore, 1983) were substantially lower than onset-order thresholds (Pastore, 1982). Both types of thresholds increased as the common durations of the tones increased, but were insensitive to frequency separation. Onset-order thresholds were also dependent on the onset rise-time of the stimulus. No effects on offset-order threshold were observed as a function of the relative intensity of the tones in a pair; however, possible effects of relative intensity difference may have been canceled by averaging results over the two orders in which the pair of tones were presented.

Pastore and Farrington (1996) measured the ability to identify the order of onset of components within complex auditory stimuli that were created to approximate the temporal and spectral properties of consonant-vowel syllables. The temporal-order discrimination threshold was constant at roughly 30 msec for discrimination standards with short onset-time differences (< 25 msec) between F1 and F2. The size of the threshold increased linearly (with a slope of approximately 1.0) with the onset time of the discrimination standard for standards with longer onset-time differences (> 25 msec) between F1 and F2.

Jaskowski et al. (1990) measured temporal-order thresholds of roughly 30 msec for a mixed modality, auditory-visual stimulus condition (auditory click and visual flash)

using an adaptive procedure. The authors also reported an asymmetry in the perception of simultaneity, i.e., the auditory stimulus had to be delayed in order to be perceived as simultaneous with the visual stimulus.

In addition to the tactile research included in the experiments of Hirsh and Sherrick (1961), described above, there are several other relevant studies of temporal-order discrimination through the tactual sense.

Sherrick (1970) studied the ability to discriminate the temporal order of clicks presented at the thigh under various conditions including stimulation type (mechanical, electrotactile, and both), and place of stimulation (bilateral vs. ipsilateral). The thresholds measured ranged from 20 to 35 msec. Sherrick also investigated the effect of small intensity differences between two mechanical stimuli separated 10 or 20 cm along the length of the thigh. He found that thresholds increase with the intensity of the proximal stimulus, and this effect is more obvious for distal-stimulus first trials than for proximal-stimulus first trials.

Marks (1982) examined the dependence of temporal-order threshold for electrocutaneous stimulation on place of stimulation, distance between the stimuli, and stimulus orientation. Each of these three factors had an effect on temporal-order judgments. In addition, the obtained thresholds were substantially larger than those reported in other studies (greater than 100 msec in most conditions), which was ascribed both to possible procedural differences and to the poor temporal resolving capacity for electrocutaneous stimulation.

Craig and Baihua (1990) studied temporal-order discrimination for tactile patterns (a "vertical" stimulus versus a "horizontal" stimulus) presented through an Optacon

transducer array. Three conditions were tested: same site (left index fingerpad), ipsilateral (left index fingerpad and left middle fingerpad), and bilateral (left index fingerpad and right middle fingerpad). The temporal-order threshold was substantially lower for the same-site condition (12 msec) than for the ipsilateral (68 msec) and bilateral (62 msec) conditions. The effect of intensity differences between the two stimuli in a pair on the size of the temporal-order threshold was also examined. The results indicated that at short values of SOA in the same-site condition, the intensity pairing "Low in Interval 1-High in Interval 2" resulted in the best performance, and the pairing "High-Low" resulted in the worst performance. The effect of the intensity imbalance was much less in the ipsilateral and bilateral conditions compared to the same-site condition and appeared to occur in the opposite direction.

Results may differ across experimental studies due to various factors such as procedural differences, subject variability, the amount of training, and differences in stimuli. The values of temporal-order threshold fall into a relatively small range from several msec to tens of msec, with the exception of the results of Marks (1982) for electrocutaneous stimulation. Physical properties of the stimuli (duration, rise time, intensity, modality, etc.) also have an effect on the size of the temporal-onset-order threshold, although the relationship between these properties and their effects is still not entirely understood.

In the current study, stimuli roving in both amplitude and duration in each trial were employed with the purpose of depriving the subjects of cues other than temporal-onset-order (including, for example, possibly redundant cues of duration, energy, or stimulus offset), and also with the purpose of simulating the types of stimuli that might be

encountered in the speech experiment. Tactual temporal onset-order threshold ($d' = 1$) ranged from 18 msec to 43 msec across the four subjects (see Table 7-3), and was consistent with the range of values reported in the literature. Amplitude differences between stimuli (defined as the amplitude of the stimulus with an earlier onset minus the amplitude of the stimulus with a later onset) had a clear and consistent effect on the performance of all subjects at all values of |SOA| (see Fig. 7-6). Performance increased as the amplitude of the stimulus with an earlier onset increased relative to that of the stimulus with a later onset. This effect was larger at shorter values of |SOA| than at longer values. In addition, for some of the shorter values of |SOA|, the perceived order was the opposite of the order in which the stimuli were presented, particularly when the amplitude of the stimulus with a later onset was larger than that with an earlier onset. The size of the duration difference between the signals in interval 1 and interval 2 appeared to have little effect on performance (except for S3, see Fig. 7-10). In the case of S3, performance decreased as the difference in the duration of signals (D1 – D2) increased.

An interpretation of the effects of amplitude differences and duration differences on temporal-onset-order discrimination is discussed below.

8.2.2 Interpretation of Effects of Roving in Amplitude

8.2.2.1 Development of Hypotheses

The following Hypotheses were developed to explain the amplitude-difference effects observed in the current study.

1) The perceptual onset of the stimulus (t2) lags behind the physical onset of the stimulus (t1) by the value τ msec (see Fig. 8-3);

2) The time lag τ is dependent on the amplitude of the stimulus and on its masking by temporally adjacent stimuli in such a way that:

a) the larger the stimulus amplitude, the smaller the value of τ;

b) the greater the adjacent masking, the larger the value of τ.

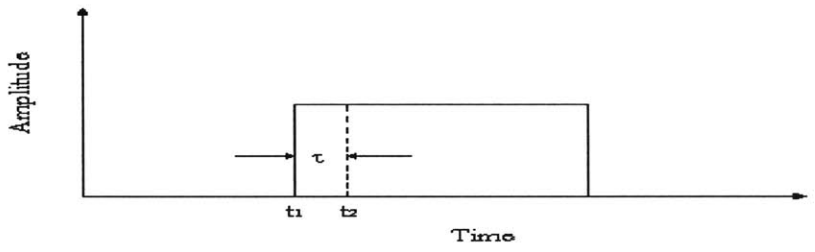

Fig. 8-3. Illustration of the lag between the physical onset and the perceptual onset of a given stimulus.

Although it is difficult to test Hypothesis (1) directly because it is difficult to measure the exact time of perceptual onset of a given stimulus, several arguments can be made in its support. The time lag (τ) can be argued to contain at least two components. One component is the transmission delay (μ) arising from the transmission time that is needed for the tactual signal to be transmitted from the finger to the cortex. Based on the conduction velocities of the mechanoreceptive afferent fibers (in the Aβ range from 35 to 70 m/sec, Johnson et al., 2000), 14 to 29 msec will be needed for the tactual signal to travel a distance of approximately 1 m. This transmission delay (μ) is independent of both the amplitude of the stimulus and the amount of masking from temporally adjacent stimuli assuming the conduction velocity is constant. A second component is the

perceptual delay (v) that is the time necessary for the subject to perceive the onset of the signal after its arrival at the cortex. This second component is dependent on the amplitude of the stimulus and the amount of masking produced by temporally adjacent stimuli. The existence of this second component (v) is implicitly predicted by models of temporal summation.

Temporal summation refers to the well-known phenomenon that the loudness of short sounds grows with their duration, or that threshold increases as duration decreases. Various mathematical models for the summation process have been proposed (see Zwislocki, 1960; Penner, 1978; Viemeister and Wakefield, 1991). These models employ different types of integration processes to account for the time-intensity trades that are observed both for detection and discrimination tasks. Despite the differences among individual models, all such models share the general notion that the subjective perception at some given time t is not the instantaneous stimulus intensity (or some transformation of the intensity at that time), but is, instead, based on a weighted average of stimulation that has occurred in the recent past. In addition, none of these models have taken transmission delay into consideration. Although these models were developed for results obtained through hearing, they may also be applied to tactual perception. Verrillo (1965), in experiments with vibrotactile stimulation, demonstrated that Zwislocki's (1960) theory of temporal summation can predict accurately the threshold shift as a function of pulse-repetition rate, number of pulses, and burst duration of sinusoidal signals for the Pacinian channel.

In the following discussion, the transmission delay will be ignored since its only effect is a shift in time. The first illustration discusses the existence of perceptual delay,

and the second illustration discusses the effect of amplitude on the perceptual delay. Both illustrations are discussed in relation to the temporal summation phenomena.

We can construct two stimuli S1 and S2 (see Fig. 8-4) satisfying the following constraints: a) Stimulus 1 (S1) is the same as stimulus 2 (S2) in every aspect except its duration is half of S2; b) S1 cannot be perceived (its amplitude is below the detection threshold); c) S2 can be perceived (its amplitude is above the detection threshold). S2 is composed of two consecutive repetitions of S1. Since S1 is below threshold and the first half of S2 is the same as S1, there is no obvious reason that S2 will be perceived during the period [T1, T2] if we assume that the human sensory system is a causal system. However, S2 can be perceived, so the perceptual onset must occur during the interval [T2, +∞]. Thus, the perceptual onset (somewhere in the interval [T2, +∞]) is later than the physical onset (T1) of S2.

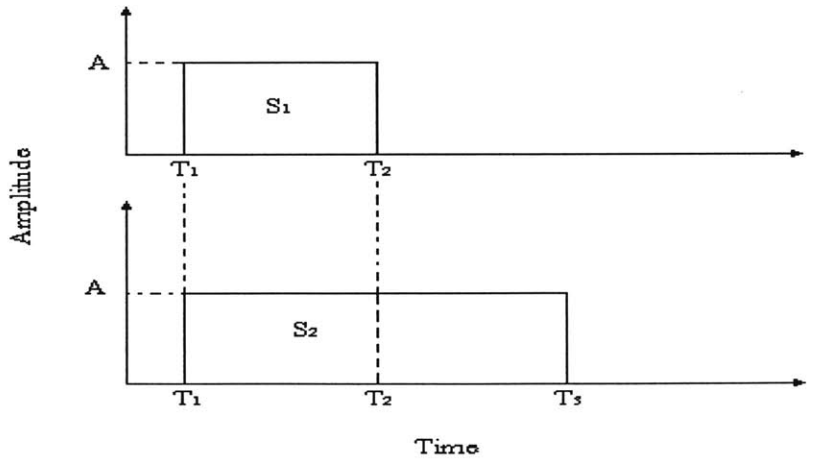

Fig. 8-4. Illustration of two stimuli with the property that S1 is below threshold, and S2 is above the threshold.

Hypothesis (2a) (the relation between stimulus amplitude and the perceptual lag) is shown in Fig. 8-5. The amplitude of S1 (A1) is larger than the amplitude of S2 (A2). The lag between the perceptual onset and physical onset of S1 ($\tau 1$) is smaller than that of S2 ($\tau 2$). This assumption is well-supported by models of temporal summation which predict the well-known relation between the signal duration and detection threshold: that is, that threshold decreases as duration increases over some range of duration. A larger-amplitude stimulus requires a shorter duration to be detected compared to a smaller-amplitude stimulus. Therefore, if S1 and S2 have physically simultaneous onsets, the perceptual onset of S1 ($T1+\tau 1$) is earlier than the perceptual onset of S2 ($T1+\tau 2$).

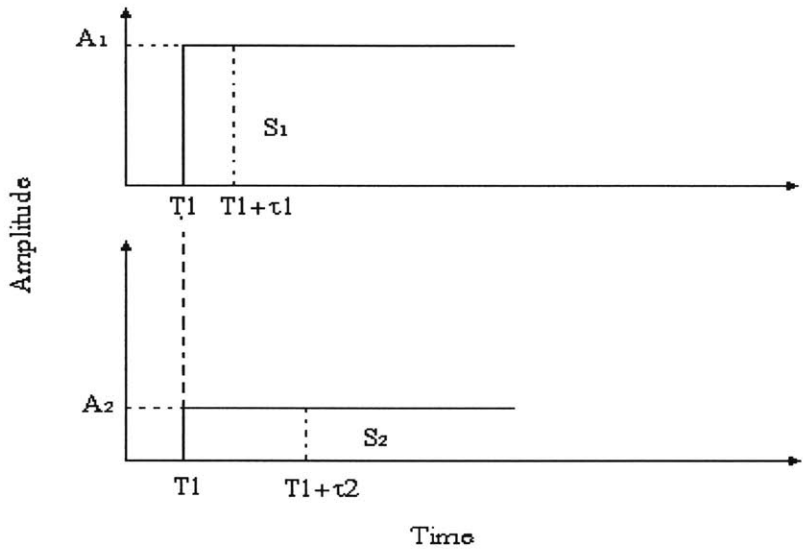

Fig. 8-5. Illustration of the effect of amplitude on perceptual lag.

Unlike Hypothesis 2(a) that describes the effect of the physical parameter of the stimulus (amplitude) on the perceptual onset, Hypothesis 2(b) takes into consideration the effect of the interaction between two signals (masking) on their perceptual onsets.

Masking refers to the increased difficulty in detection of a signal due to the existence of another signal (masker). Its underlying mechanism is still not completely understood. According to the relative temporal relationship between the signal and masker, there are three types of masking: 1) simultaneous masking in which the masker and the signal occur at the same time, 2) forward masking in which the masker occurs before the signal, and 3) backward masking in which the signal occurs before the masker. For all three types of masking, it is generally true that the larger the amplitude differences between the masker and the signal (amplitude of the masker minus the amplitude of the signal), the stronger the effect of masking. In addition, the more adjacent the masker and the signal, the stronger the masking will be. Furthermore, forward masking has a longer effective time range of several hundreds of msec, while backward masking decays mush faster with an effective range on the order of tens of msec (Sherrick, 1964). In the current experiment, all three types of masking came into play across trials depending on the combination of stimulus duration (which ranged from 50 to 800 msec) and SOA values (which ranged from 5 to 115 msec).

Hypothesis 2(b) is an inference based on both temporal-summation theory and the basic properties of masking (that the threshold of a signal is elevated in the presence of a masker and that the stronger the masking the more elevated the threshold). An illustration of the effect of masking on perceptual onset is shown in Fig. 8-6. The top panel represents the physical amplitude of the signal as a function of time. The bottom panel represents the perceived amplitude of the signal as a function of time using temporal-summation theory (although no specific integration function was assumed here). Threshold 1 is the detection threshold when there is no masking, and the signal begins to

be perceived at time $T_1 + \tau_1$. Threshold 2 is the detection threshold when there is a certain amount of masking that causes the signal to be perceived at time $T_1 + \tau_2$. Thus, the perceptual onset is delayed by $\tau_2 - \tau_1$ with the existence of the masker. Furthermore, this delay is assumed to increase with the strength of the masker. Note that a sufficiently strong masker is capable of elevating the detection threshold such that it is higher than the maximum of the perceived amplitude for a given stimulus; thus, the signal can be no longer perceived.

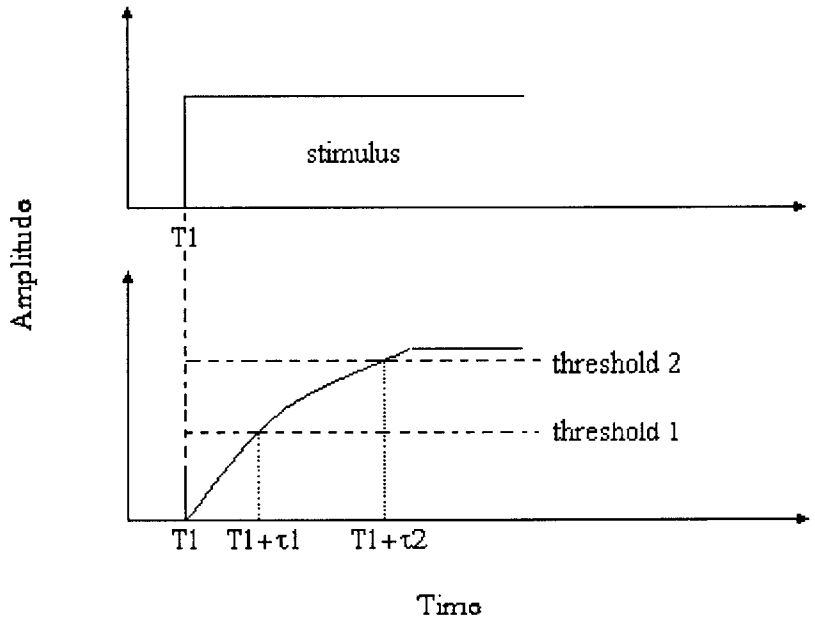

Fig. 8-6. Illustration of the effect of masking on perceptual onset.

8.2.2.2 Interpretation of Current Results According to the Hypotheses

In a temporal onset-order discrimination experiment, the subject's task is to judge the order in which two distinct stimuli are perceived to be presented. The physical onset

asynchrony (|SOA|) is the time difference between the physical onset of S1 (T_1) relative to the physical onset of S2 (T_2), i.e., $|SOA| = |T_2 - T_1|$. The two perceptual onsets (defined as $\tau 1$ and $\tau 2$ msec later than the two physical onsets, respectively) are two random variables due to effects of both internal and external noise sources. This noise will undoubtedly affect the discrimination performance; however, it is considered to be of secondary importance compared to the expected duration between the perceptual onsets (defined as $|POA| = |(T_2 + \tau_2) - (T_1 + \tau_1)| = |T_2 - T_1 + \tau_2 - \tau_1|$). In the following discussion, performance is assumed to be determined primarily by the expected mean duration between the two perceptual onsets (See Fig. 8-7).

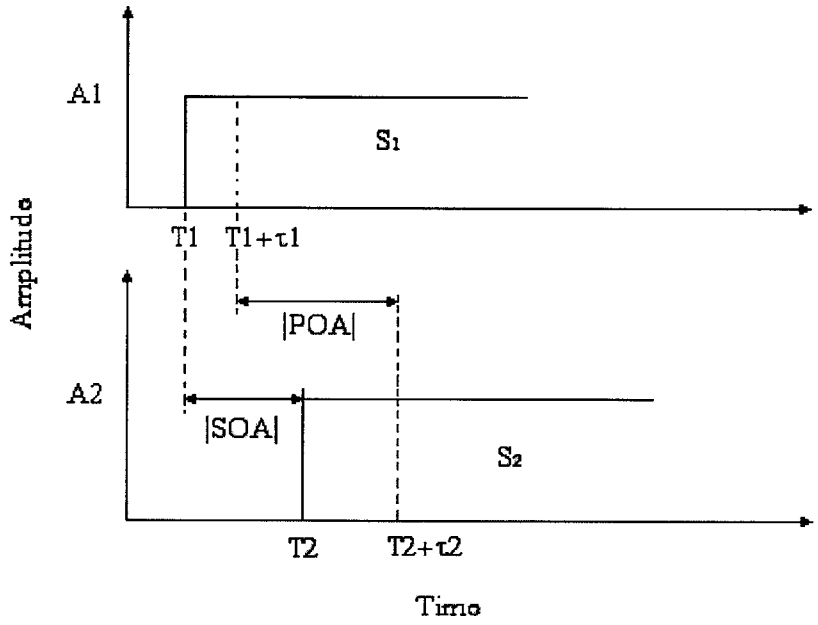

Fig. 8-7. Relation between perceptual-onset asynchrony (POA) and physical-onset asynchrony (SOA) for a pair of stimuli with equal amplitude.

The data collected from the current study are well explained qualitatively by the two hypotheses. Assume that in the case where the amplitude of the stimulus with an earlier onset (A1) equals that of the stimulus with a later onset (A2), i.e., $A_1 = A_2$, the duration between the perceived onsets of the two stimuli is $|SOA| - \tau 1 + \tau 2$. For convenience, the stimulus with an earlier onset will be called S1, and the stimulus with a later onset will be called S2 in later discussion. In the case of A1 > A2 (achieved by increasing A1 while keeping A2 the same), the lag between the perceptual onset and physical onset of S1 (L1) is shorter than $\tau 1$ (i.e., $\tau 1 - \Delta 1$), due to hypothesis 2(a). Thus, the perceived onset difference is $|SOA| - \tau 1 + \tau 2 + \Delta 1$ considering only hypothesis 2(a). Qualitatively, for a fixed value of $|SOA|$, $\Delta 1$ is a monotonically increasing function of A1 (or amplitude difference A1-A2, since A2 is fixed).

Furthermore, taking into consideration hypothesis 2(b), the lag between the perceptual onset and the physical onset of S2 (L2) is greater than $\tau 2$ (i.e., $\tau 2 + \Delta 2$), due to stronger forward masking exerted on S2 by S1. Qualitatively, for a fixed value of $|SOA|$, $\Delta 2$ is also a monotonically increasing function of A1 (or amplitude difference A1 - A2, since A2 is fixed). In addition, the more adjacent in time the masker and the signal are, the larger the value of $\Delta 2$ (in other words, for small values of SOA, the effect of masking is more significant).

Thus, the overall perceived onset difference is $|SOA| - \tau 1 + \tau 2 + \Delta 1 + \Delta 2$, longer than $|SOA| - \tau 1 + \tau 2$. Both $\Delta 1$ and $\Delta 2$ are monotonically increasing functions of A1; therefore, the larger the amplitude difference, the larger the value ($\Delta 1 + \Delta 2$), and the better the performance will be. Note that $\Delta 2$ can go to $+\infty$, which occurs when S2 is completely masked by S1. This explanation accounts for the observed results (shown in

Figs. 7-6 and 7-8) indicating that d′ increases as the amplitude difference between S1 and S2 increases, particularly for small values of |SOA|. At larger values of |SOA|, performance is already saturated. Similar argument can be applied to the case A1 > A2 by decreasing A2 while keeping A1 the same.

In the case of A1 < A2 (achieved by decreasing A1 while keeping A2 the same), the lag between the perceptual onset and physical onset of S1 (L1) is longer than $\tau 1$ due to two factors. The first factor is the decrease in amplitude (leading to longer delay, i.e., $\Delta 1$) and the second factor is elevation in threshold due to backward masking of S1 by the S2 with larger stimulus amplitude (leading further to a longer delay, i.e., $\Delta 2$). The overall lag between the perceptual onset and physical onset of S1 (L1) is $\tau 1 + \Delta 1 + \Delta 2$. Thus, the perceived onset difference ($|SOA| - \tau 1 + \tau 2 - \Delta 1 - \Delta 2$) is shorter than it is for the equal-amplitude case, leading to a reduction in performance. This can account for the observed results showing that performance decreases as the amplitude difference (A1-A2) decreases at a given value of |SOA| (see Fig. 7-6). The data indicate that in some cases (especially at smaller values of |SOA|), the perceived onset order is the opposite of the physical onset order, i.e., $d' < 0$. This result may be explained by assuming that in some cases $|SOA| - \tau 1 + \tau 2 - \Delta 1 - \Delta 2 < 0$, i.e., $\Delta 1 + \Delta 2 > |SOA| - \tau 1 + \tau 2$, thus leading to reversed judgment of the temporal order. This effect is less obvious when |SOA| is larger (it is less possible that $\Delta 1 + \Delta 2 > |SOA| - \tau 1 + \tau 2$). Similarly note that $\Delta 2$ can go to $+\infty$, occurring when S1 is completely masked by S2 (i.e., the subject perceives only one

stimulus, which is S2). In such a situation, the subject seems to respond with the stimulus perceived as having the earlier onset. [1]

8.2.3 Interpretation of Effects of Roving in Duration

The duration difference between the signals (S1 and S2) appears to have had a negligible effect on performance for three of the subjects (S1, S2, and S4). S3 showed a tendency for performance to decrease as the duration difference (D1-D2) increased (see Fig. 7-10). The effect is more obvious for positive larger duration difference (D1-D2 > 300 msec, i.e., categories 6 and 7). These results for S3 may be explained by judgments based on the probabilistic use of a "second" onset under the circumstances D1>>D2 (see Fig. 8-8). In Fig. 8-8, P11, P12 are the perceptual onsets of S1 with an earlier onset. P21 is the perceptual onset of S2 with a later onset. The occurrence of a second onset P12 of S1 is due to the long duration difference (D1-D2). S3 might mistakenly use P12 at some probability as the onset of the stimulus with a later onset, thus leading to decreased performance.

[1] It should be pointed out that a 3-dB/octave compensation was applied to the short-duration tones (50 and 100 msec) which were below the temporal integration constant. For any given amplitude difference, there are 49 duration pairs of the two stimuli in each trial. Among them, the amplitude differences of 25 out 49 pairs were not affected by the compensation. The amplitude differences of the remaining 24 pairs were either increased or decreased by 3 dB or 6 dB. Thus, the effect on temporal-onset order judgment might be cancelled by the equally likely increase and decrease. In sum, the effect of the 3-dB/octave compensation might be negligible in the analysis of the amplitude difference effect on temporal-onset order judgment. An analysis of the effect of the amplitude-difference including only the trials without 3-dB/octave compensation was conducted. The similar result further confirmed that the effect of the 3-dB/octave compensation was negligible on the analysis of the effect of the amplitude difference on temporal-onset order judgment.

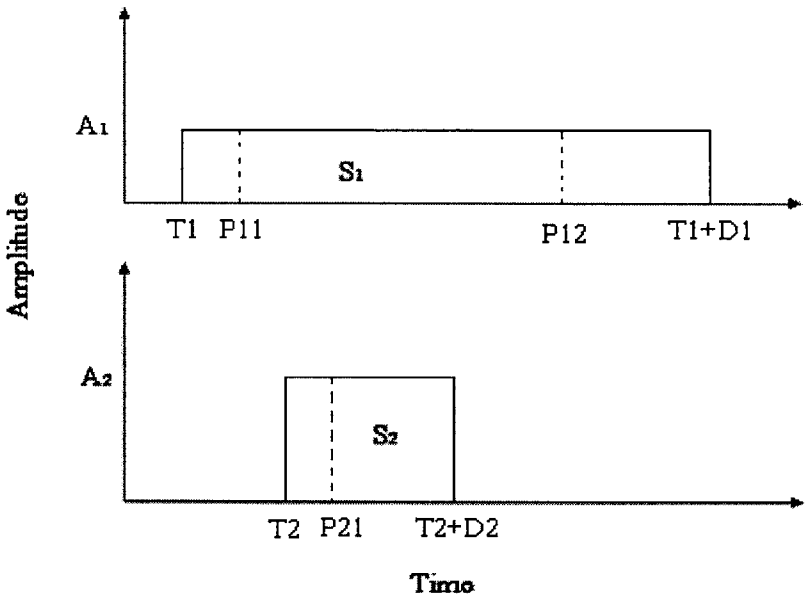

Fig. 8-8. Illustration of the second perceptual onset P12 of the stimulus with an earlier onset (S1) can be potentially used as the onset of the stimulus with a later onset by subject S3 when D1>>D2.

8.2.4 Other Issues

The stimuli employed in the current study of temporal-onset order were coded redundantly along two dimensions of finger and frequency: a 250-Hz tone was always presented at the index finger and a 50-Hz tone was always presented at the thumb. In other words, subjects could use either of the two dimensions in making judgments of order: i.e., frequency and finger carry the same information regarding temporal order. The effect of this redundant coding for temporal-onset order was not pursued in the current study. Taylor (1978), however, investigated dimensional redundancy in the

processing of vibrotactile temporal order. His results indicate that dimensional redundancy (dimensions are correlated) can facilitate temporal-order judgments.

Another issue relevant to the current study is the effect of the selection of the two particular frequencies (50 and 250 Hz) on temporal-order thresholds. Both the group delay for different frequencies and possible interactions between the two frequencies may have an effect on the psychophysical measurements. For the first question, previous studies have generally been concerned more with the relationship in amplitude rather than with the response time. However, it is very likely that the human tactual system behaves as most natural systems do, and thus will not have a flat group delay for different frequencies. It is important to known whether the magnitude of the difference between the group delays of the two frequencies is large enough to affect the temporal-onset order judgment. It should be noted that the difference in group delay between 250 Hz and 50 Hz for presentation through the Tactuator device is about 2 msec and was not compensated for in the current experiments. The second question, regarding the interaction between frequency separation and temporal-order threshold, is more complicated. Experiments exploring such effects are of interest to the development of tactual interfaces.

The results of the current study of temporal-onset order judgment clearly demonstrate that the amplitude difference between the two stimuli has a significant effect on the order judgment. This fact needs to be taken into serious consideration when designing a tactual display.

8.3 Pair-Wise Voicing Discrimination

8.3.1 Comparison with Previous Results

The results of the pair-wise voicing discrimination tests may be compared to results of similar tests conducted with other tactual displays. Delhorne, Rabinowitz, and Reed (1989) studied segmental discrimination for four different tactual displays. These displays included two wearable tactual aids (the Tactaid 2 and Tactaid 7 devices of Audiological Engineering Corporation) and two laboratory devices based on the Queen's University tactile vocoder (Brooks and Frost, 1983), a 16-channel version and a 9-channel version. The characteristics of each of these 4 devices are summarized in Table 8-4. Segmental discrimination was measured in three-normal-hearing subjects and included four pairs of voicing contrasts (/p-b/, /k-g/, /s-z/, and /ch-j/). The speech materials consisted of C-/a/ tokens recorded by two male speakers with three repetitions of each syllable. Testing employed a one-interval, two-alternative forced-choice procedure with correct-answer feedback.

Table 8-4. Summary of the characteristics of four tactual displays.

Device	No. of channels	Frequency characteristics	Body site	Carrier frequency
Tactaid 2	2	Low channel:100~1800 Hz High channel: 1500 to 8100 Hz	Wrist/hand	250/375/400 Hz vibrator
Tactaid 7	7	Three channels below 1200 Hz => F1 Three channels above 800 Hz => F2 Middle channel => (F1 or F2)	Forearm	250/375 Hz vibration
Queen's 16	16	Third-octave bands, with center frequencies from 160-8000 Hz	Forearm	100 Hz solenoids
Queen's 9	9	2/3-octave bands	Forearm	100 Hz solenoids

Reed et al. (1992) also collected segmental discrimination data through the Tactaid 7 device for four profoundly deaf tactile-aid users using the same stimuli and procedure as described for Delhorne et al. (1989). The body sites used by these subjects included the back of the neck, the abdomen, and the sternum.

A comparison of voicing-discrimination scores among the devices studied by Delhorne et al. (1989) and Reed et al. (1992) with the current tactual display (T) is summarized in Table 8-5. Percent-correct scores are provided for each of the four voicing pairs employed in the previous studies. The scores obtained in the current study with 2I2AFC procedure have been modified to be consistent with performance in a 1I2AFC procedure. Namely, d′ values were divided by $\sqrt{2}$, converted to percent-correct scores assuming no bias, and then averaged across subjects. Performance on the devices studied previously averaged 57% for the Queen's 16-channel device, 58% for the Tactaid 2, 62% for the Queen's 9-channel device, and 65-70% for the Tactaid 7. These scores are

substantially lower than the 81% average obtained for the current display. This superior result was obtained in spite of the fact that the current study employed a larger set of tokens to represent each consonant and introduced additional variability through the use of C_1VC_2 syllables.

Table 8-5. Voicing-discrimination scores in %-correct for four voicing contrasts obtained with different tactual displays.

Studies		# of channels	Procedure	# of tokens per C_1	Contrasting pairs				
					/p-b/	/k-g/	/s-z/	/ch-j/	Mean
Delhorne et al. (1989)	Tactaid 2	2	1I2AFC	6	60	53	64	56	58
	QS16	16		6	53	51	65	59	57
	QS9	9		6	60	56	69	64	62
	Tactaid 7	7		6	59	71	71	78	70
Reed et al. (1992)	Tactaid 7	7	1I2AFC	6	71	58	66	63	65
Current study	Tactual alone	2	2I2AFC	24	81	85	85	71	81

8.3.2 Cues Used in Pair-Wise Discrimination

The results of pair discrimination showed consistently good ability to discriminate the feature voicing under the two modalities T and L+T across all subjects and all pairs (see Fig. 7-16). The performance under L+T was similar to the performance under T alone. These observations indicate (1) that the tactual display of the onset-difference of the envelopes from the two frequency bands of speech is an effective perceptual cue for the feature voicing, and (2) that bimodal presentation did not enhance performance.

The perceptual strategies used by the subjects in the speech discrimination experiments can be explored by relating performance on this task to temporal-order thresholds and to EOA measurements. The measurements of EOA can be separated into

three categories based on the perceptual results of the temporal-onset order experiment. Averaged across the four subjects, threshold for determining the temporal order of a 250 Hz signal at the index finger and a 50-Hz signal at the thumb was roughly 34 msec. The acoustic measurements of EOA (Envelope Onset Asynchrony - see Tables 5-2 to 5-5) for the 16 consonants can be separated into the following three categories:

Category 1: The onset of the envelope obtained from the high-frequency band (ENV_{high}) precedes the onset of the envelope obtained from the low-frequency band (ENV_{low}) (>+34 msec).

Category 2: The onsets of the two envelopes are nearly simultaneous (in the range of -34 to +34 msec).

Category 3: The onset of ENV_{low} precedes the onset of ENV_{high} (<-34 msec).

The EOA measurements for the set of C_1VC_2 tokens representing each C1 were summarized according to the 3 categories defined above (see Table 8-6).

Table 8-6. EOA categories for the tokens with each initial consonant C1.

C1	**/p/**	**/t/**	**/k/**	**/ch/**	**/f/**	**/th/**	**/s/**	**/sh/**
Category	1	1	1	1	1, 2	1, 2	1	1
C1	**/b/**	**/d/**	**/g/**	**/j/**	**/v/**	**/tx/**	**/z/**	**/zh/**
Category	2	2	2	1	2, 3	2, 3	2, 3	2, 3

In the pair-wise discrimination experiment, the two envelopes modulated two different frequencies and were delivered to the two fingers (50T vs. 250I, as described in Chapter 6). The experiment employed a two-interval two-alternative forced-choice procedure in which the subject's task was to determine the order of the two tokens with C1 contrasting in voicing. The task might involve three steps for most pairs: 1)

Determine the onset order of the two channels for the signal in interval 1 and categorize the EOA according to Table 8-6 if possible, 2) Determine the onset order of the two channels for the signal in interval 2 and categorize the EOA according to Table 8-6 if possible, and 3) Make a final decision regarding the presentation order of voiced versus voiceless initial consonant. Because the EOA categories are based on the onset-order of the two envelopes, the decision regarding EOA category is therefore based primarily on the tactual temporal-onset order perception for most pairs (only order is relevant). It is worth noting that unlike in the temporal-onset order discrimination test where the amplitude of the carrier was constant, the amplitude of the envelopes derived from the speech was time varying (see Fig. 5-3). This may have introduced additional complexity to the task of order discrimination in the pair-wise discrimination test. Furthermore, since the envelopes were derived in real time, the phases of the two sinusoids were not controlled. In the temporal-onset order discrimination experiment, the phases of the two sinusoids always started from zero. Such phase randomization should have only a secondary effect on the order judgments.

For the stop consonants, the voiceless stops /p, t, k/ generally fall into Category 1 (exhibiting EOA measurements typically greater than 60 msec), and the voiced stops /b, d, g/ generally fall into Category 2 (exhibiting EOAs less than 30 msec). In discriminating the three pairs of voiceless-voiced stops (/p-b/, /t-d/, and /k-g/), the subject is presumed to determine the onset order of the two channels in each of the two intervals. If the onset of vibration on the index finger precedes that of the thumb, then the stimulus is assigned to Category 1 (voiceless), whereas if the two channels appear to have simultaneous onsets, the stimulus is assigned to Category 2 (voiced). Depending on the

particular speech tokens selected at random for each trial, the categorization may be more or less difficult. Occasionally, two stimuli may be assigned to the same category, requiring the subject to guess. The average d′ scores across subjects and stop pairs are 2.8 (T) and 2.7 (L+T).

For the fricatives, EOA values span two different categories for the sounds /f/, /v/, /th/, /tx/, /z/ and /zh/. In the case of the pairs /s-z/ and /sh-zh/, the EOA for the voiceless fricatives always lies within Category 1, whereas the EOA for the voiced fricative can assume values in Category 2 or Category 3. Thus, the onset order of the two channels of the display should be distinctive as described above for stop consonants. The average scores for these two pairs are 3.1 (T) and 2.9 (L+T). For the fricative pairs /th-tx/ and /f-v/, however, overlapping EOA categories may occur for the two members of a pair. For example, when discriminating tokens of /f/ versus /v/ or /th/ versus /tx/, it is possible that both tokens of a given trial may elicit a decision of Category 2 value of EOA. Thus, the subject would be forced to guess, thus leading to the relatively inferior discrimination performance observed on these two pairs. The average scores of these two pairs are 2.6 (T) and 2.5 (L+T).

The steps employed for discriminating the affricate pair /ch-j/ might be different from the steps mentioned above. For most of the tokens with initial consonants /ch/ and /j/, the onset of the envelope obtained from the high-frequency band (ENV_{high}) precedes that of the envelope obtained from the low-frequency band (ENV_{low}) (EOA for both consonants lies in category 1), although the EOA of /ch/ is generally longer than that of /j/. Order of onset of the two channels is thus not a reliable cue for this pair; instead, an ability to discriminate difference in the duration of the asynchrony may be required for

the decision. Such a task is similar to an onset-asynchrony discrimination task with roving values of asynchrony in each of the two intervals. The asynchronies of the two intervals fall into two regions of EOA: 1) EOA spans a range with larger mean 156 msec (/ch/), and 2) EOA spans a range with smaller mean 70 msec (/j/). The standard deviations for EOA for these sounds are 27 and 26 msec. The subject's task is to decide which interval has a longer EOA, and to label that interval as the voiceless stimulus.

The basic psychophysical experiment relevant to this task (i.e., the comparison between two EOA values, with both EOA values varying in a certain range from trial to trial) was not conducted in the current study. However, Pastore and Farrington (1996) did similar though less complicated experiments in auditory sense. The difference limen for discriminating the comparison from a standard delay were measured. They found that approximately 35 msec was required for discriminating the comparison from the 5-msec-delay standard, and for standard onset-delay above 25 msec, the size of the temporal-order threshold was approximately 12 msec above the standard onset-delay.

8.3.3 Comparison with the Performance of an Ideal Observer

The overall performance of the human subjects on the pair-wise voicing discrimination task is far inferior to the predicted performance of an ideal observer (see Chapter 5). This discrepancy may be due to differences in both the properties of the two types of observers and in their decision procedures. The ideal observer can discriminate any small difference in the onset difference, while human subjects have perceptual limitations on their ability to discriminate temporal order due to sensory noise, memory noise, lack of attention, fatigue, etc.

Though the observed performance in the pair-wise discrimination experiment was inferior to predictions of an ideal observer, the observed d′ values were fairly well correlated with the performance of the ideal observer derived from EOA measures. The performance is relatively high across the three pairs of stops (/p-b/, /t-d/, and /k-g/) and the three pairs of fricatives (/th-tx/, /s-z/, and /sh-zh/). Performance on the pairs /ch-j/ and /f-v/ is the least sensitive, and is consistent with the acoustic EOA measurements where the two pairs /ch-j/ and /f-v/ have the lowest values of d′. The voicing discrimination performance of individual subjects is also positively correlated with their individual sensitivity in the temporal onset-order discrimination task (see Fig. 7-18). The subject with the highest sensitivity in the onset-order discrimination task is the subject with the best performance in the discrimination experiment. In summary, the perceptual performance on pair discrimination seems to be dependent on both the EOA measurements and the subject's perceptual ability to discriminate the temporal onset orders.

8.3.4 Other Issues

a. Coding Redundancy

The coding of the EOA cue for speech signals is such that the two amplitude envelopes modulate two different sinusoids and the modulated sinusoids are delivered to two different fingers. The coding is redundant in the sense that both the onset order of the fingers and the onset order of the two frequencies carry the same information regarding the onset order of the two amplitude envelopes. Subjects could base their judgment of temporal order on either or both dimensions (finger and frequency). Though

the effect of redundancy was not investigated in our study, the results from previous work (Taylor, 1978) suggest non-redundant coding would lead to lower performance.

b. Bimodal Integration

Individual and mean d′ of the eight pairs under lipreading alone (L), touch alone (T), and lipreading supplemented by touch (L+T) are shown in Fig. 8-9. The predictions of d′ for the combined condition using the Prelabeling model of Braida (1991), assuming perfect integration, are also shown in the Figure. Across the pairs, the values of predicted d′ for the combined condition are similar to the values of observed d′ under the modality of T alone. This result is reasonable given that the performance under lipreading alone is near chance (d′ of roughly 0). The contribution of d′ under the modality L is negligible to the prediction using $d'_{prelabel} = \sqrt{(d'_L)^2 + (d'_T)^2}$. A trend is observed for slightly lower scores on L+T relative to T alone and relative to the predictions of the Prelabeling model. Although these differences are not significant, they suggest that the human observer may have been distracted by the non-informative visual information in the modality L+T.

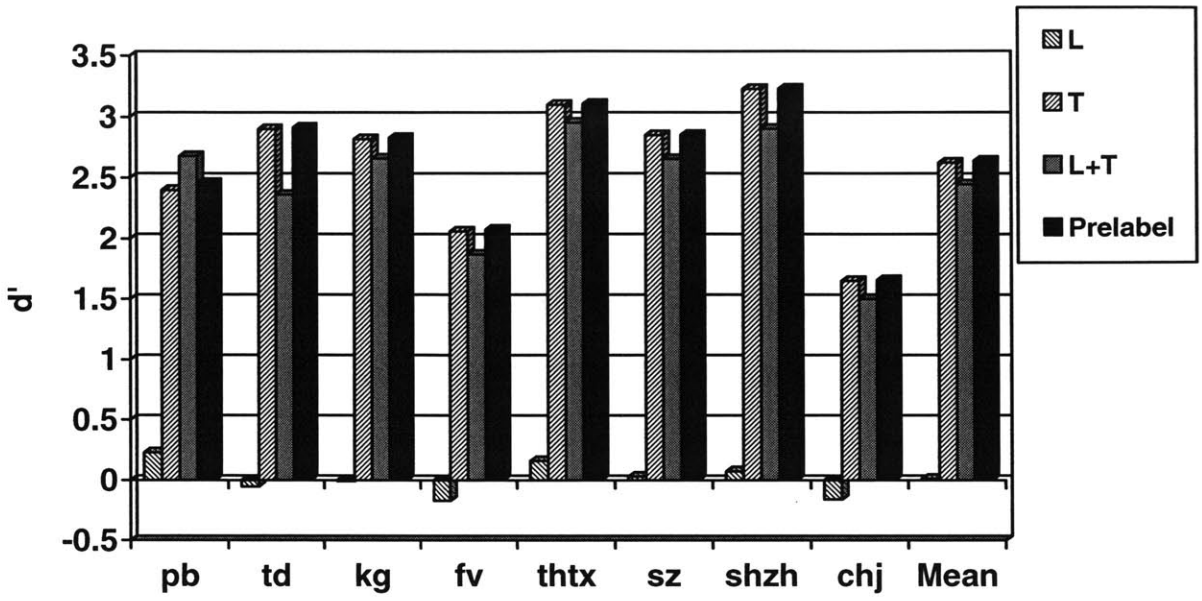

Fig. 8-9. Individual and mean d′ of the eight pairs under lipreading alone (L), touch alone (T), lipreading supplemented by touch (L+T), and prediction using the Prelabeling model of Braida (1991).

8.4 16-Consonant Identification

8.4.1 Comparison with Previous Results of Other Tactual Displays

The results of the 16-consonant identification tests may be compared to results of similar tests conducted with other tactual displays. Each of the four studies (Carney, 1988; Blamey, et al., 1988; Weisenberg, 1995; Bratakos, et al., 2001) evaluated consonant identification for an acoustic-based tactual display as a supplement to lipreading. These studies are summarized in Table 8-7.

Carney (1988) studied consonant identification with a single-channel tactile device (the Mini Fonator) and with a 24-channel tactual vocoder. The single-channel device directed the incoming unprocessed speech signal to a single, disk-shaped vibrator

that interfaced with the subject's finger. However, with the limited frequency and temporal resolution of human tactual system, the subjects essentially received only amplitude-envelope information of the wideband speech signal. The multichannel device is a 24-channel tactual vocoder that filtered the speech spectrum into 24 bands and coded the frequency into place of stimulation on the arm.

Blamey et al. (1988) studied phonemic information transmission through a multichannel electrotactile device (Tickle Talker) with normal-hearing subjects. The speech processor extracted three output parameters from speech signals: estimated second-formant frequency (F2), the scaled fundamental frequency (F0), and the speech amplitude envelope (A). These three parameters were coded by electrode position, pulse rate, and pulse width, respectively, through stimulators worn on the fingers.

Weisenberger et al. (1995) evaluated the transmission of phoneme-level information through a multichannel tactile aid (Tactaid 7) that is designed to convey the first two formant frequencies of speech.

Bratakos et al. (2001) investigated consonant-identification performance using a single-channel vibrator that delivered the envelope of a frequency band of speech centered at 500 Hz modulating a 200 Hz carrier.

A comparison of the results among these studies and the current study is difficult given the many differences in devices, procedures, materials, etc. Performance across studies generally followed the pattern of T < L < L+T. To normalize for different levels of ability on lipreading alone, a relative gain measure was used to assess the benefit provided for aided lipreading (see the final column of Table 8-7). Relative gain is calculated as [(L+T)-L]/(100-L), thus providing a measure of what proportion of the total

possible improvement to lipreading was actually carried by the tactual supplement. For the studies discussed here, relative gain ranged from approximately 7% to 31%, and was roughly 23% for the current study.

Table 8-7. Comparison between results of other studies and the average results across subject of 16-consonant identification of the current study.

Study	Device	Material	Subject	Training	Performance			
					L	T	L+T	$\frac{(L+T)-L}{100-L}$
Carney (1988)	Mini Fonator, 1–channel	20 consonant in CV syllable, live-voice by 1 speaker	6 normal hearing	≈16 hours per condition with feedback [2]	79%	33%	83%	0.19
	24-channel tactual vocoder		6 normal hearing		62%	26%	74%	0.31
Weisenberger (1995)	Tactaid 7, 7 channels	24 consonants in C/ae/t, C/u/t, C/i/t contexts	6 normal hearing	NA	57%	14%	64%	0.16
					46%	12%	52%	0.11
					40%	11%	44%	0.07
Blamey et al. (1988)	Tickle talker, Multi-channel	12 consonant /p, b, m, f, v, s,z, n, g, k, d, t/ in /a-C-a/, live-voice by 1 speaker	4 normal hearing	NA	44%	29%	56%	0.21
Bratakos, et al. (2001)	Minishaker vibrator, 1-channel	24 /C1-a-C2/, 8 tokens for each C1	3 normal hearing female subjects	3 single run blocks to 5 double run blocks, with each run containing 144 items	53%	11%	64%	0.23
Current study	2-channel	16 consonant in /C1VC2/, 2 speakers	4 normal hearing	8 runs with each run containing 80 items	34%	12%	49%	0.23

[2] If the subject was incorrect, the experimenter showed them the correct response and repeated the target several times. This repeated training made it possible that the subjects might remember the vibrotactile pattern for individual CV syllables and lead to the high lipreading scores reported here.

8.4.2 Estimates of Pair-Wise Performance Using Constant-Ratio Rule

The relation between the results obtained from the pair-wise discrimination task and from the 16-consonant identification task was examined using the constant-ratio rule (Clarke, 1957). According to the constant-ratio rule, the ratio between any two entries in a row of a submatrix is equal to the ratio between the corresponding two entries in the master matrix. Values of d′ of each of the eight pairs were computed from the corresponding 2×2 matrices extracted from the 16×16 confusion matrix obtained in the identification experiment for each subject. The values of d′ for each pair multiplied by a factor of $\sqrt{2}$ (to compensate for the use of a 1-interval procedure in identification) were averaged across the four subjects. These values are shown in Fig. 8-10. The difference in d′ between that obtained in the pair-wise discrimination experiment and that predicted from the 16-consonant identification experiment is plotted in Fig. 8-11. The differences in d′ range from -1.2 to 0.69 across pairs and conditions; however, the mean differences across pairs are small (-0.5 ~ 0.3) for each of the three conditions. Thus, the constant ratio rule is by and large satisfied. This result indicates that the voicing-discrimination ability observed in the pair-wise tests carries over to the 16-consonant identification task.

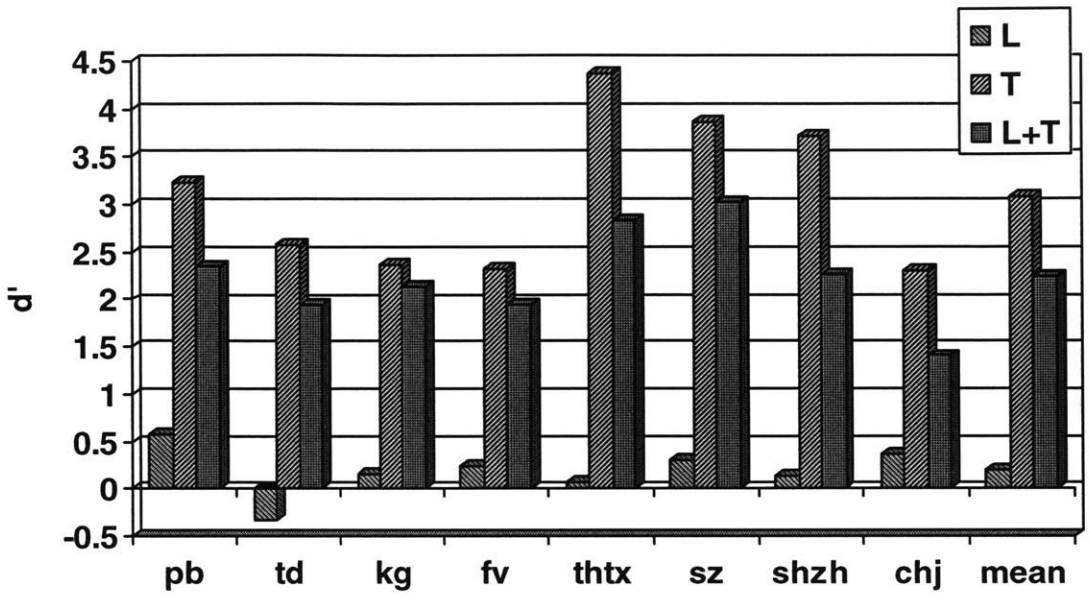

Fig. 8-10. d′ predicted from the confusion matrix, and multiplied by a constant $\sqrt{2}$.

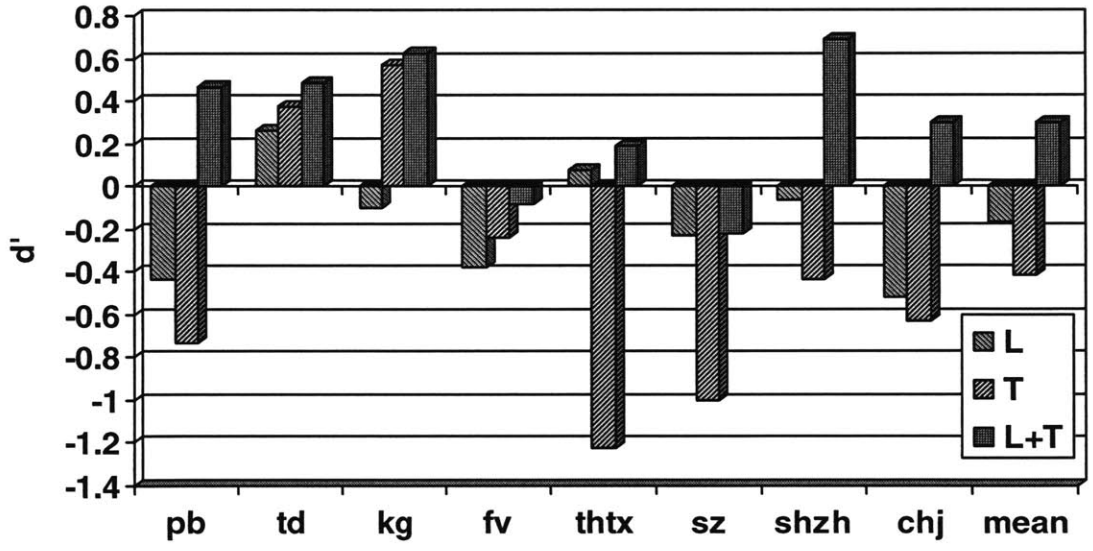

Fig. 8-11. Difference between d′ obtained in pair-wise discrimination and d′ predicted from results of 16-consonant identification ($d'_{pair\text{-}wise}$ - $d'_{16\text{-}c}$).

8.4.3 Predictions of Bimodal Performance

Predictions of bimodal performance under the combined modality of L+T were made using two different models described below.

The first model ("Perfect Voicing") assumes perfect reception of voicing through the tactual display combined with observed performance through lipreading. This model provides an upper bound on identification performance through L+T. The predicted structure of the confusion matrix under L+T (L+T $_{PV}$) is based on a modification of the confusion matrix observed under the modality L. The 16×16 confusion matrix under the modality L was modified in the following way: for each entry in the confusion matrix N(S, R),

$$N(S1, R1) = \begin{cases} N(S1, R1) & \text{if S1 and R1 have the same voicing feature.} \\ 0 & \text{if S1 and R1 have different voicing feature.} \end{cases},$$

The entry was left intact when the voicing feature of S1 was the same as that of R1. The entry was set to 0 when the value of the feature of voicing of S1 was not the same as that of the response (R1). This number was then added to the entry (S1,R2), where R2 has the same manner and place as R1, but assumes the same value of voicing as S1.

An example of the manipulation of two rows of the confusion matrix is shown in the following two tables. Table 8-8 shows two rows of the original confusion matrix, and Table 8- 9 shows the resulting two rows after manipulation.

Table 8-8. Two rows extracted from a confusion matrix under L condition.

	p	b	t	d	k	g	f	v	th	tx	s	z	sh	zh	ch	j
p	50	30	0	2	0	0	0	1	0	0	0	0	0	0	0	0
b	34	44	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 8-9. Two rows after manipulation.

	p	b	t	d	k	g	f	v	th	tx	s	z	sh	zh	ch	j
p	80	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0
b	0	78	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Comparisons among the perceptual results in percentage overall information transfer under the observed L and L+T [$(L+T)_{OBS}$] confusion matrixes and under the L+T matrix modified for the assumption of perfect voicing are shown in Fig. 8-12 for individual subjects and for means across subjects. Performance is ordered as $L < (L+T)_{OBS} < (L+T)_{PV}$ for each individual subject and for the mean across the subjects. Averaged across subjects, the predicted %-IT for the combined condition assuming perfect voicing was roughly 70% compared to 47% for L alone and 52% for observed $(L+T)_{OBS}$.

Percent-correct scores are presented for individual subjects and for means across subjects in Fig. 8-13. Averaged across subjects, the predicted score for $(L+T)_{PV}$ was roughly 63%-correct compared to 34% for L alone and 49% for $(L+T)_{OBS}$. Thus, the observed results indicate a 15 percentage-point improvement over L; if no voicing errors had been made, an additional 14 percentage-point improvement would have been possible. The pattern seen in Fig. 8-13 differs from that seen in Fig. 8-12 primarily in two aspects: 1) The ratio between the improvement measured in %-correct scores of $(L+T)_{OBS}$ over L and $(L+T)_{PV}$ over L: $\frac{(L+T)_{OBS} - L}{(L+T)_{PV} - L}$ is larger than that measured in %-IT scores

(0.52 vs. 0.22), and 2) The size of the improvement for $(L+T)_{PV}$ over L measured in %-correct scores is larger than that measured in %-IT scores (29 vs. 23 percentage points).

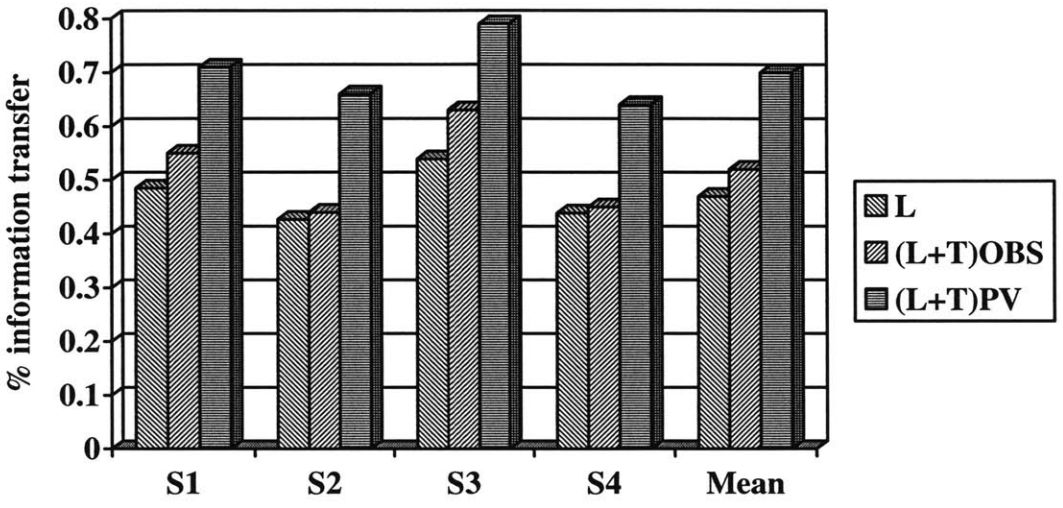

Fig. 8-12. %-IT for conditions of L, $(L+T)_{OBS}$, and $(L+T)_{PV}$.

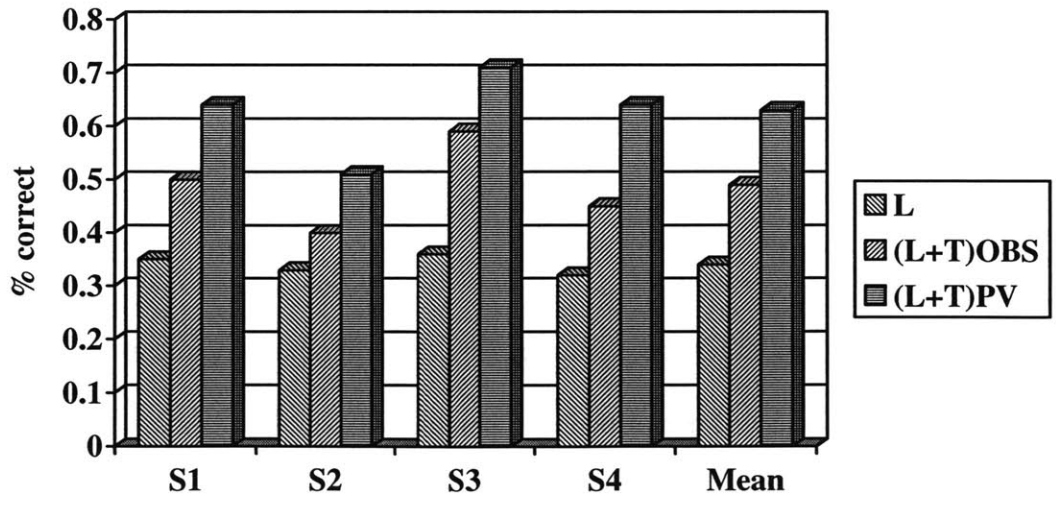

Fig. 8-13. Percent-correct scores for conditions of L, $(L+T)_{OBS}$, and $(L+T)_{PV}$.

In the second model, the predictions arise from the Pre-Labeling Integration model proposed by Braida (1991). This model is a multidimensional version of the theory of signal detection. The stimulus centers and response centers are estimated from the monomodal confusion matrix using multidimensional scaling (MDS). MDS is a method to geometrically represent similarity/confusability ratings such as those that arise from a confusion matrix. For bimodal stimuli, the cue space is the Cartesian product of each of the two monomodal cue spaces. Therefore, the continuously valued cues are combined across sensory systems to produce a vector of cues in a multidimensional space. The cue vector is displaced from the stimulus center S due to a shift by Gaussian noise. The potentially high-dimensional space of cue vectors is split into N compact "response regions" that are bounded by hyperplanes (where N is the number of possible stimuli). The subject is assumed to respond R_k if the distance between the cue vector and the response center $\overrightarrow{R_k}$ is the smallest compared to all the other response centers. This model assumes optimal integration of the two senses, and no perceptual interference occurs across modalities (i.e., masking or distraction). Therefore, the efficiency of the integration of information across two modalities (e.g., vision and touch) can be assessed by comparing predicted with observed performance.

The observed individual-subject confusion matrices for the modalities L and T were used to generate a predicted confusion matrix for the bimodal condition of L+T [$(L+T)_{PLM}$]. The bimodal response centers were chosen to be the same as the bimodal stimulus centers. Observed performance in percentage overall information transfer for L, T, and L+T, along with the bimodal prediction $(L+T)_{PLM}$, is shown in Fig. 8-14 for individual subjects and means across subjects. Performance is ordered as T < L < L+T <

$(L+T)_{PLM}$ both for individual and for mean results. The ratio of performance between $(L+T)_{OBS}$ and $(L+T)_{PLM}$: $[(L+T)_{OBS}/(L+T)_{PLM}]$ indicates the efficiency of the bimodal integration: the higher the ratio, the more efficient the bimodal integration. Inter-subject variance in integration efficiency is obvious: S3 is most efficient (ratio = 0.95) and S4 is least efficient (ratio = 0.80).

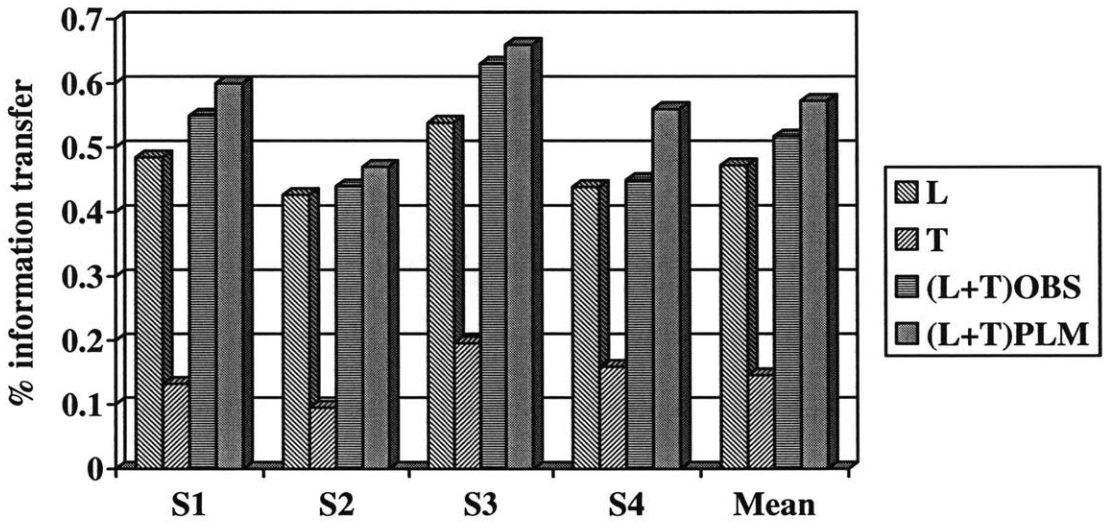

Fig. 8-14. %-IT for conditions of L, T, $(L+T)_{OBS}$, and $(L+T)_{PLM}$.

Percent-correct scores are presented for each subject and means across subjects in Fig. 8-15. The pattern seen in Fig. 8-15 differs from that seen in Fig. 8-14 primarily in that the improvement measured in %-correct scores for $(L+T)_{PLM}$ over $(L+T)_{OBS}$ is more than twice as large a that measured in %-IT scores.

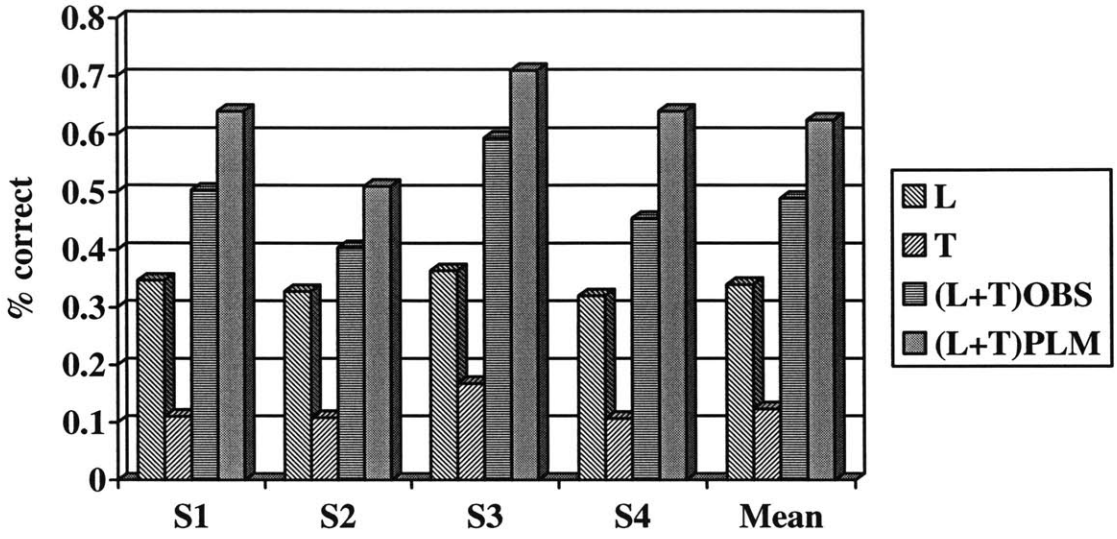

Fig. 8-15. Percent-correct scores for conditions of L, T, $(L+T)_{OBS}$, and $(L+T)_{PLM}$.

Performance on reception of the feature voicing is shown for L, T, $(L+T)_{OBS}$, and $(L+T)_{PLM}$ in Fig. 8-16 for %-feature IT, and in Fig. 8-17 for %-correct for individual and mean across subjects. Voicing reception is highest for the predictions $(L+T)_{PLM}$ which in turn is closest to performance through T alone under both measures. For each subject under each measure, performance for $(L+T)_{OBS}$ is lower than that obtained under T alone. The efficiency ratios for each subject on both measures (shown in Table 8-10) suggest that bimodal integration is not efficient for the feature voicing (<1). Across subjects, the integration efficiency ranged from 0.37 to 0.66 for %-IT measure scores and from 0.83 to 0.92 for %-correct measures.

Table 8-10. Integration efficiency ratios for individual and mean performance on voicing for %-IT measures and %-correct measures.

	S1	S2	S3	S4	Mean
%-IT	0.64	0.41	0.66	0.37	0.53
%-correct	0.92	0.90	0.92	0.83	0.89

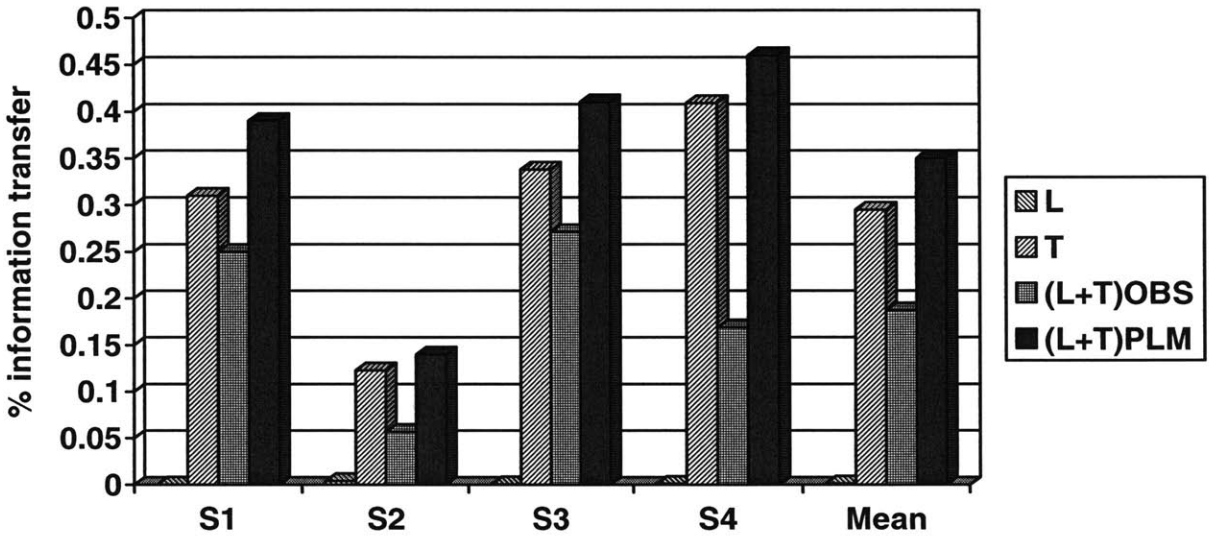

Fig. 8-16. Percentage Feature IT for voicing under L, T, $(L+T)_{OBS}$ and $(L+T)_{PLM}$.

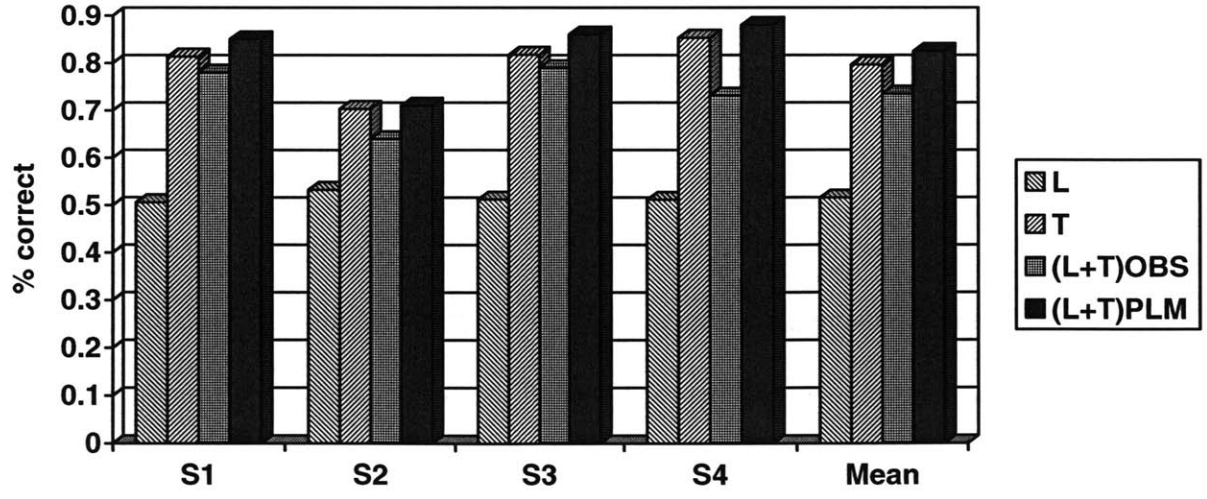

Fig. 8-17. Percent-correct voicing reception under L, T, $(L+T)_{OBS}$ and $(L+T)_{PLM}$.

8.5 CUNY Sentences

Miller et al. (1951) investigated the effects of speech material (digits, sentences, and nonsense syllables) on the intelligibility of speech under varied signal-to-noise ratio (S/N). Three functions of percent-correct scores versus S/N ratio were obtained for the three types of speech material, respectively. These results might be used to estimate the potential percent-correct score for word-recognition in sentences given the percent-correct score for nonsense syllables. The tactual cue is effective in improving consonant recognition at the segmental level from 34% correct for lipreading alone to 50% correct for the combined modality of L+T. According to Fig. 1 of Miller et al. (1951), such an improvement in performance at the segmental level translates into an improvement in word-recognition in sentences from 60% to 83%. A 16 percentage-point improvement at segmental level leads to a larger 23 percentage-point improvement in word-recognition in sentences. This amplification is due to the steeper slope of the "words in sentences" curve than that of the "nonsense syllables" curve in the region of S/N roughly from -10 to 10 dB.

The results of the CUNY sentence testing, however, indicated no benefit to lipreading with the tactual cue studied here. This result contrasts with those of previous studied which have demonstrated small improvements at the sentence level. For example, Bratakos et al. (2001) showed a 10 percentage-point improvement for aided lipreading over lipreading alone using a one-channel tactual display. This display was similar to that provided by the low-freqeuency envelope band of the current study. Thus, integration of information across the two envelopes may have been difficult for the subjects. In the auditory domain, however, Grant et al. (1994) were able to obtain improvements to

lipreading using a two-band auditory supplement similar to that used in the current tactual display. This result suggests that lack of familiarity with tactual input may also present a problem for the subjects.

The lack of benefit of the tactual cue for sentences may be due to a variety of factors. First, the amount of training subjects received on this task was limited to roughly 9 hours (5 hours for pair-wise discrimination, 4 hours for 16-consonant identification, and only roughly 10 minutes for continuous sentences). Training in the use of novel cues through the skin is critical in the evaluation of the tactual aids. In cases of good performance through the tactual sense (such as Tadoma), subjects received intensive training over a long period of time. Second, temporal masking of the tactual signals may play a greater role in continuous-sentence recognition than in segmental identification. Third, is the additional complexity of continuous speech compared to isolated syllables (such as consonant clusters, faster speaking rate, coarticulation, etc) will lead to increased variability in the acoustic cue provided by the tactual display. Fourth, the current display is designed specifically for initial consonant voicing. Its ability to resolve the ambiguity of consonant voicing in other places (final or middle) is unknown. For example, the two words pat and bad contrast in both the initial and final consonant voicing, in this case, even the initial voicing ambiguity is resolved with the aid of the tactual display, the ambiguity in final voicing can still lead to a wrong word. Fifth, the nature of the cue itself, which requires discriminating an onset-time difference between the vibrotactile signals presented at two fingers, may make it difficult to integrate with continuous speech.

Finally, it may be possible that the phonetic information of consonant voicing does not play a big role in continuous-sentence recognition due to the redundancy of language. In other words, even if the tactual cue of consonant voicing were to be perfectly perceived and perfectly integrated with lipreading, it may not provide much benefit to the recognition of continuous sentences, specifically the CUNY sentences. In continuous-sentence recognition, phonetic information is not the only information available. Other information such as lexical structure, syntax, and grammar all play a role in language processing. For example, lexical knowledge can help to solve the ambiguity for the word 'pool' when /p/ and /b/ are perceptually ambiguous because 'bool' is not a word. In this case, the voicing information is redundant with lexical knowledge. On the other hand, lexical knowledge is of no help in distinguishing between the words 'park' and 'bark' when /p/ and /b/ are perceptually ambiguous because both are legal words in the lexicon. In this case, voicing information can be used to solve this ambiguity. In a sentence, even more redundant information in the form of grammatical cues will be available. For example, consider the preceding sentence. The addition of the voicing information to distinguish between /s/ and /z/, /t/ and /d/ for the word 'sentence' doesn't really matter due to the lexicon, as well as for the words 'redundant', 'information' and 'available'. Though the lexical knowledge doesn't help to distinguish the word 'be' from 'pea', the knowledge of syntax will naturally exclude the word 'pea'. Once more, the information of consonant voicing is redundant with other information for the whole sentence we analyzed.

Although voicing information may not be necessarily required or may be redundant with lexical or other information, this does not imply that the voicing cue has

no contribution to word recognition. Voicing information enables the listener to limit the range of alternatives from which a response can be selected (reduce the subject's uncertainty), which, as Miller et al. (1951) suggested, can improve word recognition. Iverson, Bernstein, and Auer (1998) investigated the interaction of phonemic information and lexical structure with a computational approach. The phonemic information was obtained through experimental studies of the identification of nonsense syllables under conditions of lipreading, auditory alone, and audio-visual conditions. From the data, categories of perceptually equivalent phonemes were constructed under each condition. These phonemic equivalence classes were then used to retranscribe a lexicon. In general, the additional information under each condition divided the lexicon into classes of small size or even unique word (i.e., decreased the alternatives from which a response can be selected from), thus, can improve word recognition.

Chapter 9

Summary and Conclusions

The current research was concerned with the development and evaluation of a tactual display of speech designed to provide information about consonant voicing as a supplement to lipreading. Acoustic measurements of initial consonants in C_1VC_2 syllables were successful in identifying a physical cue to distinguish voiced from voiceless consonants. This cue was based on the envelope-onset asynchrony (EOA) of two bands of speech (a low-pass band and a high-pass band).

A tactual display was implemented for presentation of this cue with a two-finger tactual stimulating device and perceptual evaluations of the efficacy of this cue were conducted. These experiments indicated that:

(a) The tactual temporal-onset-order threshold is sufficiently small for subjects to perceive the onset order of the two envelopes derived from speech segments;

(b) Voicing is well-discriminated through a two-channel tactual display of the EOA cue for eight pairs of initial voicing contrasts. This cue provides significant improvement over lipreading alone.

(c) Consonant identification studies indicate that voicing information derived from the tactual display improved performance by 20-30 percentage points over lipreading alone.

(d) No significant improvement was observed over lipreading alone with the addition of the tactual cue for sentence reception.

The results of the current study indicate that the tactual cue for voicing was highly effective at the segmental level and led to levels of performance superior to those obtained with previous tactual displays. These results demonstrate that the approach taken here of selecting information to complement that available through lipreading was a judicious use of the capacity of the tactual channel. This strategy may likewise be applied to other features that are impoverished in the lipreading display. The tactual display, however, did not lead to improvements over lipreading at the continuous-speech level.

The major limitations of the current study include: restriction of the acoustic measurements to pre-stressed initial consonants in isolated syllables; the need for further exploration of psychophysical properties as background for establishing an optimal display; insufficient amount of training on the connected-speech reception task; and the need for further investigation of the properties of bimodal integration.

Chapter 10

Directions for Future Research

Future research is motivated primarily by the observed lack of improvement for the tactual cue in continuous-speech recognition compared to an observed benefit for segmental discrimination and identification. This result may arise from a variety of possible factors described below, including (1) the extent to which the voicing cue studied here (EOA) is effective in contexts other than obstruent consonants in pre-stress initial position in isolated syllables; (2) a lack of sufficient training in the use of the tactual display on continuous-speech reception; and (3) properties of the tactual display itself which may possibly be improved for a more optimal transmission of the EOA information.

Future work to help answer these questions is described below.

(1) Generalization of EOA cue to other contexts

The acoustical measure of envelope onset asynchrony (EOA) proposed in the current study is a robust acoustic cue of voicing but is limited to obstruent consonants in initial position and in isolated syllables. Its applicability to other contexts, such as syllable-final and medial position, consonant clusters, and the coarticulation encountered in continuous speech, remains unknown. An initial step towards exploring the generalizability of the cue involves investigation of voicing of the final consonant in the C_1VC_2 syllables used in the present study. This research will include (a) acoustic measurements of EOA for

final consonants and (b) perceptual evaluations of pair-wise discrimination of final consonants through the current tactual display. The results of such a study will determine the patterns of EOA for final compared to initial consonants and determine whether this cue is effective in perceptual discrimination of voicing for final consonants as well as for initial consonants.

(2) Effect of the Amount of Training

The current research demonstrated that bimodal integration occurred (although it was not perfectly efficient) in the identification of initial consonants in C_1VC_2 syllables; however, such integration appears not to have taken place under sentence reception. In addition, cases were observed where integration was negative in the pair-wise discrimination test, i.e., the performance under L+T was slightly worse than that under T alone. This inefficiency in bimodal integration may arise from unfamiliarity with the tactual cue and from the limited amount of training subjects received in the current study. Further study is necessary to determine whether extensive training will lead to improvement in the use of the tactual cue in conjunction with lipreading for the reception of connected speech.

Before initiating a long-term training study using continuous speech, it may be worthwhile first to conduct a training study using speech materials intermediate in difficulty between initial consonants and sentences. A task involving the complete identification of C_1VC_2 syllables will require subjects to use both lipreading and the tactual cue. Such a task is more difficult than the initial consonant-identification task, but more manageable than complete sentences. An analytic approach towards development

of a training protocol must be developed. For example, it may be useful to determine the errors that subjects are making, and then to focus training on the difficult distinctions with possible supplemental cues from another modality (audition or vision).

(3) Optimization of the Tactual Display

Before embarking on a time-consuming training study, it is important to investigate areas in which the tactual display may be improved for more optimal information transmission. Future work in this area may be focused on exploration of temporal-onset-order discrimination as a function of frequency separation in a two-channel display, as a function of the use of one or two fingers to encode EOA, and as a function of the redundancy of stimulation frequency and site of stimulation.

Tactual temporal masking also plays a role in the information-transmission capability of the tactual display. Studies of temporal masking (both within and across channels of the display) may shed light on possible changes to the display and will also be of use in the interpretation of roving-level effects observed in the temporal onset-order results.

Once improvements have been made to the tactual display, generalization of the EOA cue to other speech contexts has been established, and training has been demonstrated to improve performance in the identification of C_1VC_2 syllables, long-term training study for recognition of connected speech using the tactual display with lipreading may then be initiated. The results of such a study are necessary to determine whether the EOA cue for voicing can be used in connected-speech reception and thus should be considered in the design of practical aids for persons with profound hearing-impairments.

Bibliography

Ananthapadmanabha, T. V., and Yegnanarayana, B. (1975). Epoch extraction on voiced speech. *IEEE Transactions on Acoustics Speech and Signal Processing, 23(6),* 562-570.

Ananthapadmanabha, T. V., and Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE transactions on acoustics speech and signal processing, 27(4),* 309-319.

Baum, S. R., and Blumstein, S. E. (1987). Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *Journal of the Acoustical Society of America, 82,* 1073-1077.

Behrens, S. J., and Blumstein, S. E. (1988). Acoustic characteristics of English voiceless fricatives: a descriptive analysis. *Journal of Phonetics, 16,* 295-298.

Bernstein, L. E. (1992). The evaluation of tactile aids. In I. R. Summers (eds.), *Tactile aids for the hearing impaired.* London: Whurr Publishers, 167-186.

Besing, J. M., Reed, R. M., and Durlach, N. I. (1995). A comparison of auditory and tactual presentation of a single-band envelope cue as a supplement to speechreading. *Seminars in hearing, 16(4),* 316-327.

Blamey, P. J., Cowan, R. S. C., Alcantara, J. I., And Clark, G. M. (1988). Phonemic information transmitted by a multichannel electrotactile speech processor. *Journal of speech and hearing research, 31(4),* 620-629.

Bolanowski, S. J. Jr., and Verrillo, R. T. (1982). Temperature and criterion effects in a somatosensory subsystem: A neurophysiological and psychophysical study. *Journal of Neurophysiology, 48,* 836-855.

Bolanowski, S. J. Jr., Gescheider, G. A., Verrillo, R. T., and Checkosky, C. M. (1988). Four channels mediate the mechanical aspects of touch. *Journal of the Acoustical Society of America, 84,* 1680-1694.

Boothroyd, A., Hanin, L., and Hnath, T. (1985). *A sentence test of speech perception: Reliability, set equivalence, and short term learning.* (Speech and Hearing Science Report No. RC110). New York: City University of New York.

Boothroyd, A. (1988). Perception of speech pattern contrasts from auditory presentation of voice fundamental-frequency. *Ear Hearing, 9(6),* 313-321.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L. and Kishon-Rabin, L. (1988). Voice fundamental-frequency as an auditory supplement to the speechreading of sentences. *Ear and hearing, 9(6),* 306-312.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology-A, 43(3),* 647-677.

Bratakos, M. S., Reed, C. M., Delhorne, L. A., Denesvich, G. (2001). A single-band envelope cue as a supplement to speechreading of segmentals: A comparison of auditory versus tactual presentation. *Ear and Hearing, 22 (3),* 225-235.

Breeuwer, M., and Plomp, R. (1984). Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the Acoustical Society of America,* 76, 686-691.

Brooks, P. L., and Frost, B. J. (1983). Evaluation of a tactile vocoder for word recognition. *Journal of the acoustical society of America, 74(1),* 34-39.

Brughera, A. (2002). *Upgrade of the Tactuator computer, DSP, and control system.* Sensory Communication Group, RLE, MIT.

Carney, A. E. (1988). Vibrotactile perception of segmental features of speech – a comparison of single-channel and multichannel instruments. *Journal of speech and hearing research, 31(3),* 438-448.

Cheng, Y. M., and O'Shaughnessy, D. (1989). Automatic and reliable estimation of glottal closure instant and period. *IEEE transactions on acoustics speech and signal processing, 37(12),* 1805-1815.

Choi, J. Y. (1999). *Detection of consonant voicing: A module for a hierarchical speech recognition system.* Ph.D dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Clarke, F. R. (1957). Constant-ratio rule for confusion matrices in speech communication. *Journal of the Acoustical Society of America, 29(6),* 715-720.

Cole, R. A., and Cooper, W. E. (1975). Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America, 58 (6),* 1280-1287.

Cornett, R. O. (1967). Cued speech. *American Annals of the Deaf, 112,* 3-13.

Craig, J. C., and Baihua, X. (1990). Temporal order and tactile patterns. *Perception and Psychophysics, 47,* 22-34.

Crystal, T. H., and House, A. S. (1988). Segmental durations in connected-speech signals - current results. *Journal of the Acoustical Society of America, 83(4),* 1553-1573.

DeGroot, M. H., and Schervish, M. J. (2002). *Probability and statistics.* (3rd ed). Addison-Wesley, Boston, MA.

Delhorne, L. A., Rabinowitz, W. M., and Reed, C. M. (1989). Segmental discrimination performance through three tactile systems and through the symbion cochlear implant. (Unpublished).

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761-764.

Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., and Braida, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE transactions on biomedical engineering, 47(4),* 487-496.

Durlach, N. I. (1968). *A decision model for psychophysics.* Communication Biophysics Group, Research Laboratory of Electronics, MIT, MA.

Eberhardt, S. P., Bernstein, L. E., Barac-Cikoja, D., Coulter, D. C., and Jordan, J. (1994). Inducing dynamic haptic perception by the hand: System description and some results. *Proceedings of the ASME Dynamic Systems and Control Division, 1,* 345-351.

Ebrahimi, D., and Kunov, H. (1991). Peripheral-vision lipreading aid. *IEEE transactions on biomedical engineering, 38(10),* 944-952.

Erber, N. P. (1974). Visual perception of speech by deaf children - recent developments and continuing needs. *Journal of Speech and Hearing Disorders, 39(2),* 178-185.

Faulkner, A., Ball, V., Rosen, S., Moore, B. C. J., and Fourcin, A. J. (1992). Speech pattern hearing aids for the profoundly hearing-impaired: speech perception and auditory abilities. *Journal of the Acoustical Society of America, 91,* 2136-2155.

Fischer, R. M., and Ohde, R. N. (1990). Spectral and duration properties of front vowels as cues to final stop-consonant voicing. *Journal of the acoustical society of America, 88(3),* 1250-1259.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech Hearing Research, 11,* 796-804.

Formby, C., and Forrest, T. G. (1991). Detection of silent temporal gaps in sinusoidal markers. *Journal of the Acoustical Society of America, 89,* 830-837.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* New York: Wiley.

Gault, R. H. (1924). Progress in experiments on tactual interpretation of oral speech. *Journal of abnormal and social psychology, 14,* 155-159.

Gault, R. H. (1926). On the interpretation of speech sounds by means of their tactual correlates. *Annals of otology, rhinology, and laryngology, 35,* 1050-1063.

Gescheider, G. A., Verrillo, R. T., and Vandoren, C. L. (1982). Prediction of vibrotactile masking functions. *Journal of the Acoustical Society of America, 72(5),* 1421-1426.

Gescheider, G. A., Bolanowski, S. J., Pope, J. V., and Verrillo, R. T. (2002). A four-channel analysis of the tactile sensitivity of the fingertip: frequency selectivity, spatial summation, and temporal summation. *Somatosensory & Motor Research, 19 (2),* 114-124.

Goff, G. D. (1967). Differential discrimination of frequency of cutaneous mechanical vibration. *Journal of Experimental Psychology, 74,* 294-299.

Grant K. W., Ardell, L. H., Huhl, P. K., and Sparks, D. W. (1985). The contribution of fundamental-frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America, 77(2),* 671-677.

Grant, K. W., Braida, L. D., and Renn, R. J. (1991). Single band amplitude envelope cues as an aid to speechreading. *Quarterly Journal of Experimental Psychology – A, 43 (3),* 621-645.

Grant, K. W., Braida, L. D., and Renn, R. J. (1994). Auditory Supplements to speechreading-combining amplitude envelope cues from different spectral regions of speech. *Journal of the Acoustical Society of America, 95(2),* 1065-1073.

Green, D. M., and Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hamer, R. D., Verrillo, R. T., and Zwislocki, J. J. (1983). Vibrotactile masking of pacinian and non-pacinian channels. *Journal of the Acoustical Society of America, 73(4),* 1293-1303.

Hanin, L., Boothroyd, A., and Hnathchisolm, T. (1988). Tactile presentation of voice fundamental-frequency as an aid to the speechreading of sentences. *Ear and Hearing, 9(6),* 335-341.

Hansen, A. (1930). The first case in the world: Miss Petra Heiberg's report. *Volta Review, 32,* 223.

Heider, F., and Heider, G. M. (1940). An experimental investigation of lipreading. *Psychology Monographs, 52,* 124-133.

Hill, F. J., McRae, L. P., and McClellan, R. P. (1968). Speech recognition as a function of channel capacity in a discrete set of channels. *Journal of the Acoustical Society of America, 44,* 13-18.

Hillenbrand, J., Ingrisano, D. R., Smith, B. L., and Flege, J. E. (1984). Perception of the voiced-voiceless contrast in syllable-final stops. *Journal Acoustical Society America, 76(1),* 18-26.

Hirsh, I. J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America, 31,* 759-767.

Hirsh, I.J., and Sherrick, C.E. (1961). Perceived order in different sensory modalities. *Journal of Experimental Psychology, 62,* 423-432.

Hnathchisolm, T., and Boothroyd, A. (1988). Speechreading enhancement by voice fundamental-frequency - the effects of F0 contour distortions. *Journal of speech and hearing research, 35(5),* 1160-1168.

Hnathchisolm, T., and Kishonrabin, L. (1988). Tactile presentation of voice fundamental-frequency as an aid to the perception of speech pattern contrasts. *Ear and Hearing, 9(6),* 329-334.

Hnathchisolm, T., and Medwetsky, L. (1988). Perception of frequency contours via temporal and spatial tactile transforms. *Ear and Hearing, 9(6),* 322-328.

Horii, Y., House, A. S., and Hughes, G. W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the acoustical society of America, 49,* 1849-1856.

Houde, R. A. (1967). A *study of tongue body motion during selected speech sounds.* Doctoral dissertation, University of Michigan.

House, A. S., and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the acoustical society of America, 25 (1),* 105-114.

House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America, 33,* 1174-1178.

Iverson, P., Bernstein, L. E., and Auer Jr., E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26,* 45-63.

Jaskowski, P., Jaroszyk, F., and Hojan-Jezierska, D. (1990). Temporal-order judgments and reaction time for stimuli of different modalities. *Psychology Research, 52,* 35-38.

Jeffers, J., and Barley, M. (1971). *Speechreading.* Springfield, Ill.: Charles C Thomas.

Johnson, K., Yoshioka, T., and Vega-Bermudez, F. (2000). Tactile functions of mechanoreceptive afferents innervating the hand. *Journal of Clinical Neurophysiology, 17(6),* 539-558.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *Journal of the acoustical society of America, 85(4),* 1718-1725.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108(3),* 1252-1263.

Kadambe, S., and Boudreaux-Bartels, G. F. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory, 38(2),* 917-924.

Kiefte, M. (2003). Temporal information in gated stop consonants. *Speech communication, 40(3),* 315-333.

Kirman, J. H. (1973). Tactile communication of speech. *Psychological Bulletin, 80(1),* 54-74.

Kishonrabin, L., Boothroyd, A., and Hanin, L. (1996). Speechreading enhancement: a comparison of spatial-tactile display of voice-fundamental frequency (F0) with auditory F0. *Journal of the Acoustical Society of America, 10(1),* 593-602.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129-140.

Knudson, V. O. (1928). Hearing with the sense of touch. *Journal of general psychology, 1,* 320-352.

Lamore, P. J. J., Muijser, H., and Keemink, C. J. (1986). Envelope detection of amplitude modulated high-frequency sinusoidal signals by skin mechanoreceptors. *Journal of the Acoustical Society of America, 79,* 1082-1985.

Levitt, H. (1971). Transformed up-down procedures in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and speech, 1,* 153-167.

Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384-422.

Macmillan, N. A., and Creelman, C. D (1990). *Detection theory: a user's guide.* Cambridge university press.

Markel, J. D. (1972). The SIFT algorithm for fundamental frequency estimation. *IEEE Transaction on Audio and Electroacoustics, 20(5), 367-*377.

Marks, L.E., Girvin, J. P., O'Keefe, M. D., Ning, P., Quest, D. O., Antunes, J. L., and Dobelle, W.H. (1982). Electrocutaneous stimulations 3: the perception of temporal-order. *Perception and Psychophysics, 32(6),* 537-541.

Massaro, D. W. (1998). *Perceiving talking face: from speech perception to a behavioral principle.* MIT press, Cambridge, MA.

Miller, G. A., Heise, G. A., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329-335.

Noll, A. M. (1967). Cepstrum pitch determination. *Journal of the Acoustical Society of America, 41,* 293-309.

Ohde, R. N. (1984). Fundamental-frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75(1),* 224-230.

Owens, E., and Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28(3),* 381-393.

Pastore, R.E. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America, 71,* 430-436.

Pastore, R.E. (1983). Temporal order judgment of auditory stimulus offset. *Perception and Psychophysics, 33,* 54-62.

Pastore, R.E., and Farrington, S.M. (1996). Measuring the difference limen for identification of order of onset for complex auditory stimuli. *Perception and Psychophysics, 58,* 510-526.

Penner, M. J. (1978). Power law transformation resulting in a class of short-term integrations that produce time-intensity trades for noise bursts. *Journal of the Acoustical Society of America, 63(1),* 195-201.

Pirello, K., Blumstein, S. E. and Kurowski, K. (1997). The characteristics of voicing in syllable-initial fricatives in American English. *Journal of the Acoustical Society of America, 101(6),* 3754-3765.

Poppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences, 1,* 56-61.

Rabinowtitz, W. M., Houtsma, A. J. M., Durlach, N. I., and Delhorne, L. A. (1987). Multidimensional tactile displays-identification of vibratory intensity, frequency, and contactor area. *Journal of the Acoustical Society of America,* 82(4), 1243-1252.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America,* 51(4), 1296-1303.

Rauschecker, J. P., and Shannon, R. V. (2002). Sending sound to the brain. *Science, 295,* 1025.

Reed, C. M., Rubin, S. I., Braida, L. D., and Durlach, N. I. (1978). Analytic study of the Tadoma method – discrimination ability of untrained observers. *Journal of Speech and Hearing Research, 21(4),* 625-637.

Reed, C. M., Durlach, N. I., and Braida, L. D. (1982a). Research on tactile communication of speech: A review. *ASHA Monographs*, No. 20.

Reed, C. M., Durlach, N. I., and Braida, L. D. (1982b). Analytic study of the Tadoma method: identification of consonants and vowels by an experienced Tadoma user. *Journal of Speech and Hearing Research, 25, 108-116.*

Reed, C. M., Rabinowitz, W. M., Durlach, N. I., and Braida, L. D. (1985). Research on the Tadoma method of speech communication. *Journal of the Acoustical Society of America, 77,* 247-257.

Reed, C. M., Durlach, N. I., Delhorne, L. A., Rabinowitz, W. M., and Grant, K. W. (1989a). Research on tactual communication of speech: Ideas, issues, and findings. *Volta Review (Monograph),* 91, 65-78.

Reed, C. M., Durlach, N. I., Braida, L. D., and Schultz, M. C. (1989b). Analytic study of the Tadoma method – effects of hand position on segmental speech-perception. *Journal of Speech and Hearing Research, 32(4),* 921-929.

Reed, C. M., Delhorne, L. A., and Durlach, N. I. (1992a). Results obtained with Tactaid II and Tactaid VII. *Proceedings of the second international conference on tactile aids, hearing aids and cochlear implants*, Stockholm, Sweden.

Reed, C. M., Durlach, N. I., and Delhorne, A. D. (1992b). Historical overview of tactile aid research. *Proceedings of the second international conference on tactile aids, hearing aids and Cochlear Implants,* Stockholm, Sweden.

Ries, P. W. (1994). Prevalence and characteristics of persons with hearing trouble: United States, 1990-91. National Center for Health Statistics. *Vital Health Statistics, 10*, 188.

Ross, M. J., Shafer, H. L., Cohen, A., Frendberg, R., and Manley, H. J. (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustical Speech and Signal Processing, 22,* 353-362.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Shannon, R. V., Zeng, F. G., and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America,* 104(4), 2467-2476.

Sherrick, C. E. (1964). Effects of double simultaneous stimulation of the skin. *American Journal of Psychology, 77,* 42-53.

Sherrick, C.E. (1970). Temporal ordering of events in haptic space. *IEEE Transactions on Man-Machine Systems, MMS-11,* 25-28.

Sherrick, C. E. (1982). Cutaneous communication. In Neff, W. D. (eds.), *Contributions to sensory physiology, 6.* New York: Academic, 1-43.

Sondhi, M. M. (1968). New methods of pitch extraction. *IEEE Transactions on Audio and Electroacoustics, 16 (2),* 262-266.

Stevens, K. N. and Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America, 55,* 653-659.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowksi, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America, 91,* 2979-3000.

Stevens, K. N. (1998). *Acoustic Phonetics.* MIT press, Cambridge, MA.

Strube, H. W. (1974). Determination of instant of glottal closure from speech wave. *Journal of the Acoustical Society of America, 56(5),* 1625-1629.

Tan, H. Z. (1996). *Information transmission with a multi-finger tactual display.* Ph.D dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Tan, H. Z., and Rabinowitz, W. M. (1996). A new multi-finger tactual display. *Proceedings of the Dynamic Systems and Control Division, DSC-Vol. 58,* 515-522.

Taylor, B. (1978). *Dimensional Redundancy in the processing of vibrotactile temporal order.* Ph. D dissertation, Princeton University.

Turner, C. W., Souza, P. E., and Forget, L. N. (1995). Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners. *Journal of the Acoustical Society of America, 97(4),* 2568-2576.

Upton, H. W. (1968). Wearable eyeglass speechreading aid. *American Annals of the Deaf, 113,* 222-229.

Van Doren, C. L., Gescheider, G. A., and Verrillo, R. T. (1990). Vibrotactile temporal gap detection as a function of age. *Journal of the Acoustical Society of America, 87(5),* 2201-2206.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. (1987). Speech waveform envelope cues for consonant recognition. *Journal of the Acoustical Society of America,* 82, 1152-1181.

Van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (1992). Temporal cue for consonant recognition: Training, talker generalization, and use in the evaluation of cochlear implants. *Journal of the Acoustical Society of America,* 92, 1247-1257.

Verrillo, R. T. (1963). Effect of contactor area on the vibrotactile threshold. *Journal of the Acoustical Society of America,* 35, 1962-1966.

Verrillo, R. T. (1965). Temporal summation in vibrotactile sensitivity. *Journal of the Acoustical Society of America,* 37, 843-846.

Verrillo, R. T. (1966a). Vibrotactile sensitivity and the frequency response of the Pacinian corpuscle. *Psychonomic Science, 4,* 135-136.

Verrillo, R. T. (1966b). Vibrotactile thresholds for hairy skin. *Journal of Experimental Psychology, 72,* 47-50.

Verrillo, R. T. (1971). Vibrotactile thresholds measured at the finger. *Perception and Psychophysics, 9,* 329-330.

Verrillo, R. T., and Bolanowski, S. J. (1986). The effects of skin temperature on the psychophysical responses to vibration on glabrous and hairy skin. *Journal of the Acoustical Society of America, 80(2),* 528-532.

Verrillo, R. T., and Gescheider, G. A. (1992). Perception via the sense of touch. In: Summers, Ian R. (eds), *Tactile aids for the hearing impaired.* London: Whurr Publishers, pp. 13.

Viemeister, N. F., and Wakefield, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America,* 90(2), 858-865.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K. and Jones, C. J. (1977). Effects of training on visual recognition of consonants. *Journal of Speech and Hearing Research, 20(1),* 130-145.

Walden, B.E, Erdman, S. A., Montgomery, A. A., Schwartz, D. M., and Prosek, R. A. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research, 24(2),* 207-216.

Waldstein, R. S., and Boothroyd, A. (1995). Speechreading supplemented by single-channel and multichannel tactile displays of voice fundamental-frequency. *Journal of Speech and Hearing Research, 38(3),* 690-705.

Weisenberger, J. M. (1986). Sensitivity to amplitude-modulated vibrotactile signals. *Journal of the Acoustical Society of America, 80,* 1707-1715.

Weisenberg, J. M. (1995). The transmission of phoneme-level information by multichannel tactile speech-perception aids. *Ear and Hearing, 16(4),* 392-406.

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times, *Journal of the Acoustical Society of America, 93(4),* 2152-2159.

Wittmann, M. (1999). Time perception and temporal processing levels of the brain. *Chronobiology International, 16,* 17-32.

Wolf, C. (1978). Voicing cues in English final stops. *Journal of Phonetics, 6,* 299-309.

Yeung, E., Boothroyd, A., and Redmond, C. (1988). A wearable multichannel tactile display of voice fundamental-frequency. *Ear and Hearing, 9(6),* 342-347.

Zue, V. W. (1976). *Acoustic characteristics of stop consonants: a controlled study.* Ph.D. dissertation, Massachusetts Institute of Technology.

Zwislock, J. J. (1960). Theory of temporal auditory summation. *Journal of the Acoustical Society of America, 32,* 1046-1060.

Appendix A. An alternative method to calculate the performance of an ideal observer (d′) in pair-wise discrimination task

The procedures for an alternative calculation of d′ of the ideal observer using ROC curves are listed below:

1) For a given pair of consonants contrasting in voicing, sort the EOA measurements of the tokens in order from low to high for each of the two categories, respectively.
2) Select a criterion, compute the number of tokens with EOA measurement above the criterion for the two categories respectively, and convert the number of tokens into a percent-correct score by dividing by the total number of tokens of each category.
3) Transform these pairs of percent-correct score into z scores.
4) Plot ROC curves with these percent-correct score pairs in the z plane.
5) Fit the ROC curves by a straight line with unit slope.
6) Obtain the intercept of the fitting line on the y-axis, which is the estimated value of d′ for the ideal observer.

The method for the calculation of d′ described above has two major limitations. The first is that it makes use of only the overlapping region of the two EOA distributions. In the non-overlapping region, at least one z score will go to infinity; thus, the ROC curve contains only the information of the overlapping region of the two EOA distributions. In addition, d′ is not available for the case of two distributions without an overlapping region. The second is that fitting with a linear line of unit slope is very restrictive, thus not leading to good fits in some instances.

The values of d′ of the ideal observer in 1-interval experiment calculated by this method are shown in the following Tables A-1 and A-2 for the eight pairs under the two stimulus sets (3-vowel and 16-vowel). In general, these scores are lower than the values of d′ calculated by the method in Chapter 5.

Table A-1. d′ values of the eight pairs in 3-vowel stimulus set.

3-V	p-b	t-d	k-g	ch-j	f-v	th-tx	s-z	sh-zh
d′	3.43	3.04	NA	3.4	3.12	3.14	3.9	NA

Table A-2. d′ values of the eight pairs in 16-vowel stimulus set.

16-V	p-b	t-d	k-g	ch-j	f-v	th-tx	s-z	sh-zh
d′	3.07	2.5	0.73	2.93	2.09	2.52	NA	4.3

Appendix B. Summary of ANOVA for temporal onset-order experiment using |SOA| and categories of amplitude difference as the two main factors

B-1. Summary of ANOVA for S1.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	11.9424	5	2.38848	7.83	0.0003
I1-I2	18.881	4	4.72025	15.47	0
Error	6.1031	20	0.30515		
Total	36.9265	29			

Constrained (Type III) sums of squares.

B-2. Summary of ANOVA for S2.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	17.0196	5	3.40391	8.76	0.0002
I1-I2	20.8082	4	5.20205	13.39	0
Error	7.7711	20	0.38856		
Total	45.5989	29			

Constrained (Type III) sums of squares.

B-3. Summary of ANOVA for S3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	14.2154	5	2.84309	8.64	0.0002
I1-I2	32.1422	4	8.03556	24.43	0
Error	6.5785	20	0.32892		
Total	52.9361	29			

Constrained (Type III) sums of squares.

B-4. Summary of ANOVA for S4.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	46.686	5	9.3373	6.78	0.0008
I1-I2	66.79	4	16.6975	12.13	0
Error	27.538	20	1.3769		
Total	141.014	29			

Constrained (Type III) sums of squares.

Appendix C. Summary of ANOVA for temporal onset-order experiment using |SOA| and categories of duration difference as the two main factors

C-1. Summary of ANOVA for S1.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	22.9203	5	4.58407	24.16	0
D1-D2	1.1837	6	0.19728	1.04	0.4196
Error	5.6915	30	0.18972		
Total	29.7955	41			

Constrained (Type III) sums of squares.

C-2. Summary of ANOVA for S2.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	26.4973	5	5.29945	49.43	0
D1-D2	1.601	6	0.26684	2.49	0.0449
Error	3.2166	30	0.10722		
Total	31.3149	41			

Constrained (Type III) sums of squares.

C-3. Summary of ANOVA for S3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	28.9398	5	5.78797	41.61	1.31195e-012
D1-D2	26.6736	6	4.4456	31.96	9.65783e-012
Error	4.1734	30	0.13911		
Total	59.7868	41			

Constrained (Type III) sums of squares.

C-4. Summary of ANOVA for S4.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
|SOA|	75.0928	5	15.0186	121.14	0
D1-D2	1.6931	6	0.2822	2.28	0.0627
Error	3.7194	30	0.124		
Total	80.5053	41			

Constrained (Type III) sums of squares.

Appendix D. Summary of three-way ANOVA for pair-wise discrimination using modality, replication and pair as the three main factors

D-1. Summary of ANOVA for S1.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
modality	229.661	2	114.831	291.78	0
replication	4.646	4	1.162	2.95	0.0277
pair	15.862	7	2.266	5.76	0
modality*replication	6.025	8	0.753	1.91	0.0759
modality*pair	11.853	14	0.847	2.15	0.0221
replication*pair	13.957	28	0.498	1.27	0.2228
Error	22.039	56	0.394		
Total	304.043	119			

Constrained (Type III) sums of squares.

D-2. Summary of ANOVA for S2.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
modality	58.2993	2	29.1497	131.29	0
replication	2.6824	4	0.6706	3.02	0.0251
pair	6.2286	7	0.8898	4.01	0.0013
modality*replication	2.301	8	0.2876	1.3	0.2648
modality*pair	4.8215	14	0.3444	1.55	0.1233
replication*pair	5.4289	28	0.1939	0.87	0.6449
Error	12.4332	56	0.222		
Total	92.1949	119			

Constrained (Type III) sums of squares.

D-3. Summary of ANOVA for S3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
modality	135.833	2	67.9165	361.4	0
replication	3.295	4	0.8237	4.38	0.0037
pair	5.699	7	0.8142	4.33	0.0007
modality*replication	6.75	8	0.8437	4.49	0.0003
modality*pair	4.803	14	0.3431	1.83	0.0572
replication*pair	10.103	28	0.3608	1.92	0.0191
Error	10.524	56	0.1879		
Total	177.006	119			

Constrained (Type III) sums of squares.

D-4. Summary of ANOVA for S4.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
modality	174.627	2	87.3135	177.77	0
replication	10.238	4	2.5594	5.21	0.0012
pair	28.132	7	4.0188	8.18	0
modality*replication	18.981	8	2.3726	4.83	0.0001
modality*pair	20.785	14	1.4846	3.02	0.0016
replication*pair	18.453	28	0.659	1.34	0.173
Error	27.505	56	0.4912		
Total	298.719	119			

Constrained (Type III) sums of squares.