Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System

by

Jeung-Yoon Choi

B.S., Yonsei University (1992)
M.S., Yonsei University (1994)

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

June 1999

Author ....
Department of Electrical Engineering and Computer Science
May 25, 1999

Certified by ....
Kenneth N. Stevens
Clarence J. LeBel Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System

by

Jeung-Yoon Choi

Submitted to the Department of
Electrical Engineering and Computer Science
on May 25, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, a method for designing a hierarchical speech recognition system at the phonetic level is presented. The system employs various component modules to detect acoustic cues in the signal. These acoustic cues are used to infer values of features that describe segments. Features are considered to be arranged in a hierarchical structure, where those describing the manner of production are placed at a higher level than features describing articulators and their configurations. The structure of the recognition system follows this feature hierarchy. As an example of designing a component in this system, a module for detecting consonant voicing is described in detail. Consonant production and conditions for phonation are first examined, to determine acoustic properties that may be used to infer consonant voicing. The acoustic measurements are then examined in different environments to determine a set of reliable acoustic cues. These acoustic cues include fundamental frequency, difference in amplitudes of the first two harmonics, cutoff first formant frequency, and residual amplitude of the first harmonic around consonant landmarks. Hand measurements of these acoustic cues results in error rates around 10% for isolated speech, and 20% for continuous speech. Combining closure/release landmarks reduces error rates by about 5%. Comparison with perceived voicing yield similar results. When modifications are discounted, most errors occur adjacent to weak vowels. Automatic measurements increase error rates by about 3%. Training on isolated utterances produces error rates for continuous speech comparable to training on continuous speech. These results show that a small set of acoustic cues based on speech production may provide reliable criteria for determining the values of features. The contexts in which errors occur correspond to those for human speech perception, and expressing acoustic information using features provides a compact method of describing these environments.

Thesis Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering

Acknowledgments

I would first like to thank my advisor, Ken Stevens, for giving me an opportunity to learn and to grow. His encouragement and patience guided me in difficult times, and his insight and expertise provided a source of inspiration for me.

I would also like to thank my thesis committee, Prof. Louis Braida, Dr. Stefanie Shattuck-Hufnagel, and Dr. James Glass, for their continuous interest and support, and for their helpful suggestions.

The weekly lexical access meetings with Ken Stevens, Sharon Manuel, Stefanie Shattuck-Hufnagel, David Gow, Carol Espy-Wilson, Ariel Salomon, Wil Howitt and Aaron Maldonado were always enjoyable. The discussions formed the foundation on which much of my work has been based.

Many thanks go to the members of the Speech Communication group, for friendship and support. Kelly saw me through the stressful times, and Marilyn was always ready to lend a hand. Seth helped me with computer problems. Arlene took care of me better than I did myself. Corine gave her time to record speech data that was used in this thesis. Helen H., Krishna, Mark, Walter, Jeff and Lorin passed on good advice and help. Joe, Majid, Jane, Jennell, Melanie, Helen C., Janet, Harold, Gabriella and Jay always gave me encouragement.

My friends in the Korean Graduate Students' Association, especially Rena, Yoo-Kyung, Youngsook, Sunyoung, Seungsun, Yoonjung, Jinwoo, Heakyung, Jinyoung, Youngrae and Jaekyung made sure I had some fun. My fellow students in EECS also made studying at MIT an enjoyable one.

My husband, Seunghyun, stood by me in everything. I am grateful to him and to my parents-in-law for their support. My brother Junghwan deserves thanks for his encouragement. And finally, I thank my parents for making everything possible.

This work was supported by a grant from the National Institute of Health (DC02978), and by funds from the POSCO Scholarship Society.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech, as an effective means of communication between humans, has been used since prehistoric times, and studied extensively for the past several centuries. However, the mechanism of speech communication is still not fully understood. Much research is currently being conducted in such diverse fields as neurocognitive science, articulatory physiology, acoustics, auditory perception, and linguistics. In addition, speech communication has also become a focus of study in engineering. The widespread use of computers and communication networks has extended the concept of communication to include communication with and through machines. At present, the interface between humans and machines is largely through devices such as keyboards, but it is increasingly becoming desirable to communicate with machines in the natural mode of speech. Such a method for perceiving, representing and producing speech on machines may ultimately facilitate storing, searching, acquiring and communicating information between humans.

Recently, research in machine perception of speech, or speech recognition, has advanced rapidly, and current systems are able to recognize speech used in wide ranges of tasks with increasing accuracy. These systems employ methods of representing and recognizing speech which are largely based on statistical models [33]. These models represent distributions of observable quantities in localized regions of the speech signal and the transitions among these regions. Reliable procedures have been formulated for training the models from a corpora of representative examples, and for subsequent matching of input speech to those models. However, this framework is not amenable to changes in environment, or to expansion to include levels of abstraction higher than the level of the data on which it is trained. Also, the details of its performance are far from those observed in natural human communication.

Accordingly, the system proposed in this thesis tries to follow the representation and perception of speech in human communication, based on theories developed in fields such as articulatory phonetics and linguistics. These theories point to hierarchically arranged levels in the structure of speech, and suggest acoustic evidence for identification of the characteristics of speech units at each level.

1.1 Speech communication

Speech communication between humans involves two general processes: production and perception. In the production of speech, an idea is formulated and articulated through the speech production system, through movement of the respiratory system and of the various articulatory organs in the vocal tract. The resulting acoustic signal is propagated through the air to be received by the listener in the perception process, and the communicated idea is reformulated in the brain from auditory stimuli. The speech chain [9] thus involves a shared representation of speech between speakers and listeners, and shared mechanisms by which the representation of speech is enscribed onto and deciphered from the acoustic signal.

The representation of speech may be characterized by a hierarchical structure, with information at various levels. These levels include those at the prosodic, semantic, syntactic, lexical, morphological, syllabic and phonetic levels. The higher levels are more abstract or symbolic, while the lower levels have stronger acoustically measurable characteristics. In the perception process, information at all levels is used in extracting acoustic cues to identify units at all levels, so that the utterance is perceived as a coherent whole. The perception process starts with receiving the acoustic signal and extracting phonetic information, which is used to access lexical items. The information at the lexical level is then used at higher levels in arriving at the expressed idea. This thesis will examine a process for machine extraction of information from the acoustic signal at the phonetic level. The process is intended to be analogous to that in human perception. This process is termed labeling [4].

1.2 Segments and features

Speech at the phonetic level may be described in terms of segments and features [5]. A segment corresponds to a unit of speech such as a vowel or a consonant, and features are units that describe the characteristics of a segment, such as voicing and nasality. Words in the lexicon are stored in terms of segments and features. These features are represented by particular properties in the acoustic signal. The representation of a speech unit at the phonetic level is closely tied to the articulatory mechanisms responsible for its production. Classes of speech units correspond to those which have similar articulatory procedures in speech production, and hence have similar acoustic characteristics. Table 1.1 shows a list of segments in English, marked with the standard feature values.

The vowels and glides are shown in the upper portion of the table, and the lower portion shows the consonants. The features may be divided into three large classes. The first six are the *articulator-free* features, which describe the manner in which the segment is produced. Of these features, the upper three show the degree of constriction of the vocal tract. Vowels (/iy/ through /oi/) are the most open, followed by glides (/h/ through /r/), and consonants (/l/ through /ch/) are the most constricted. The lower three articulator-free features are only marked for segments that are consonantal. [Sonorant] is marked for consonants that do not produce a pressure buildup behind the constriction, such as the liquid /l/ and the nasal consonants (/m/, /n/ and /ng/). Fricative consonants (/v/ through /sh/) are marked [+continuant], since oral airflow is not completely blocked during production. On the other hand, stop consonants (/b/ through /k/) are marked [-continuant], as the airflow is blocked. Affricates (/dj/ and /ch/) have a stoppage of airflow that releases into frication, and are marked with both characteristics. Fricatives produced at and behind the alveodiscontinuity into a region of strong low frequency energy signals a sonorant closure. Likewise, a discontinuity out of that region is a sonorant release.

Figure 2.3 shows the landmarks for the consonants /sh/, /k/, /n/, /s/ and /ng/. The discontinuity between /n/ and /s/ (marked with an arrow in parentheses) is not considered a proper landmark, since it is produced when the velum is raised from the /n/ into the /s/, while the primary articulator (tongue blade) maintains the closure in the oral tract throughout both segments. The discontinuity between the /k/ release and the vowel /ae/ signals the start of phonation (voice onset). It is not considered a landmark since the change is produced by the state of the larynx and not by the three primary articulators. The start of frication of the /sh/ and end of low frequency energy of the /ng/ are also not landmarks. These points in the signal show a start or end of the phonation or frication source, but do not signal whether in fact the primary articulators made or released a constriction. These non-landmarks, which are produced by a secondary articulator such as the velum or larynx, are nevertheless important, since they may be used to infer the configuration of the vocal tract beneath the oral cavity.

2.3 Detection of acoustic cues

In order to determine the acoustic cues corresponding to the articulator and articulator-bound features, the signal is examined in more detail in the vicinity of the landmarks. These features, and hence acoustic cues, fall largely into four categories: place, nasality, tenseness, and voicing. Place is associated with the articulator features, and involves determining which primary articulator was used. The other three categories are associated with determining the configuration of the secondary articulators. Features in each of the four categories are determined differently for different types of sounds, i.e. vowels, glides, and consonants. The processing needed to determine features in the four categories will be described next.

Place in vowels is determined largely by examining the frequencies of the first two formants in relation to average values of those formants. The average values

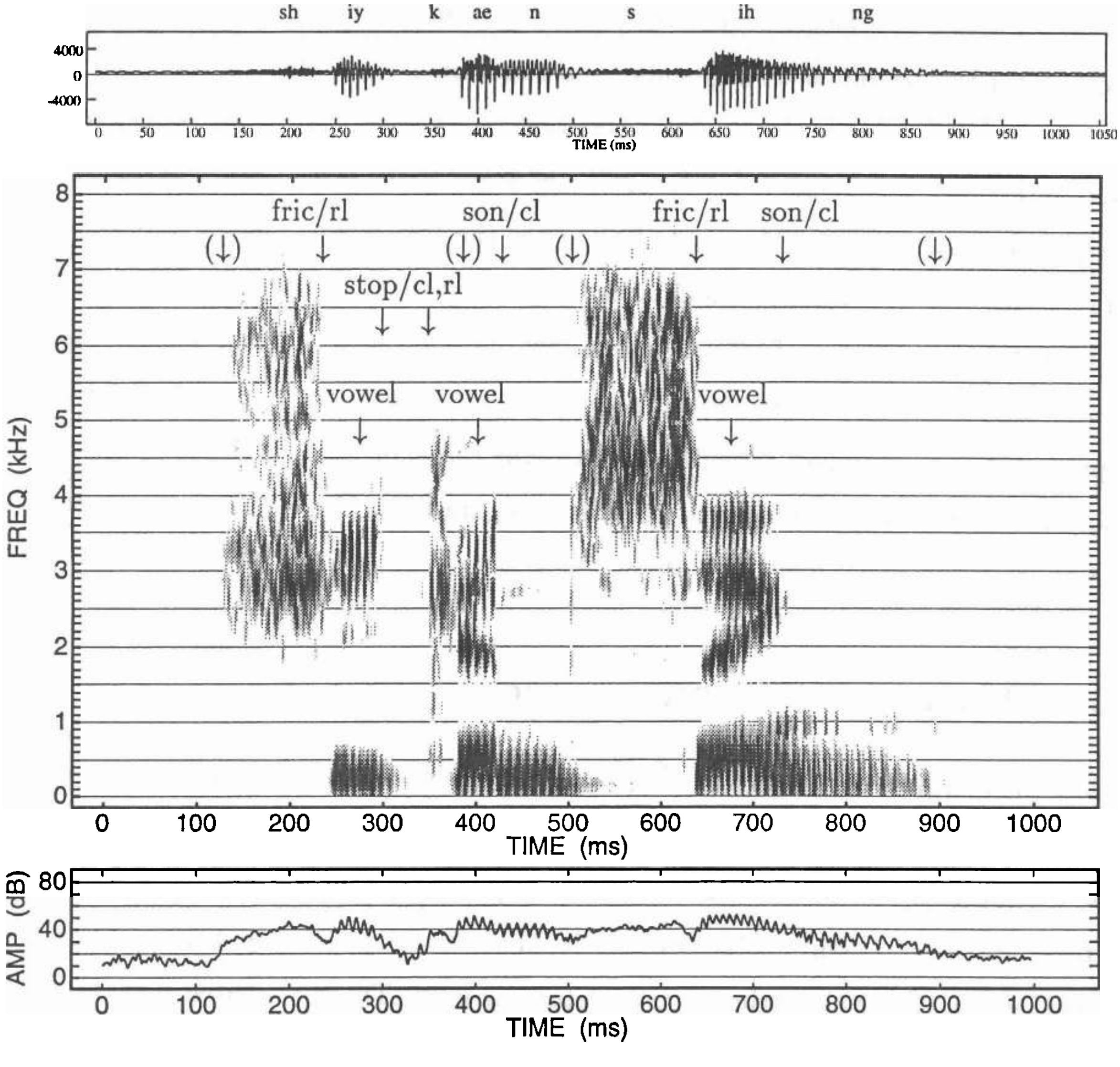


Figure 2.3: Vowel and consonant landmarks for the utterance "She can sing."

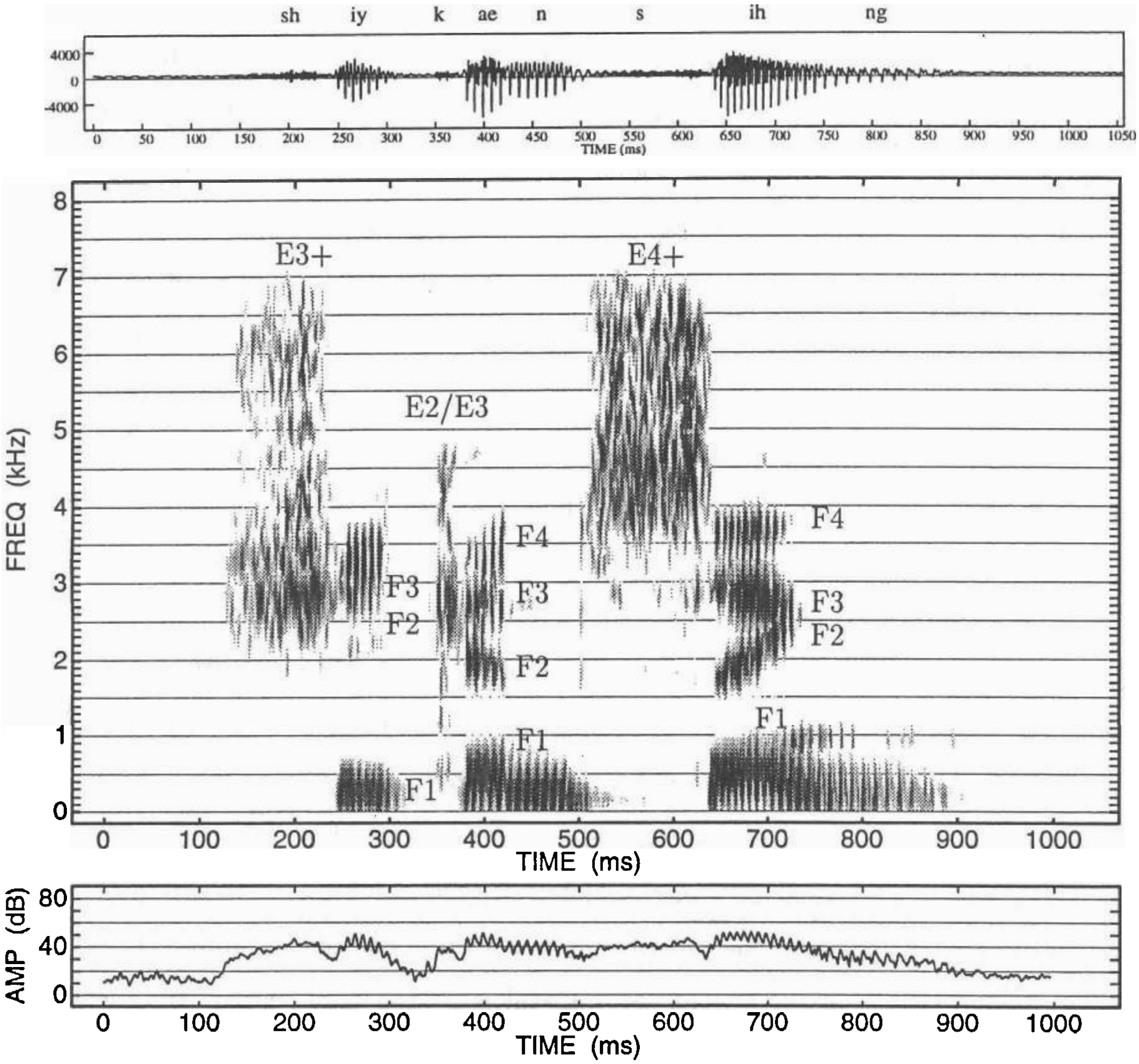


Figure 2.2: Finding primary spectral measurements in a spectrogram of the utterance "She can sing." Formants are shown for vowels and sonorant consonants. Concentration of noise energy is noted for burst and frication regions.

region is marked E3+, as for the consonant /sh/. Alternatively, a concentration of noise energy in a narrow spectral region may be specified, such as E2/E3, as shown in the burst region for the consonant /k/.

2.2 Detection of landmarks

Once the primary measurements have been made, the different regions in the signal are examined for landmarks. Phonation regions are examined for vowel and glide landmarks, and discontinuities in the signal are examined for consonant landmarks. During a phonation region, the overall amplitude of the signal and the frequency of the first formant are tracked to find maxima and minima. Maxima correspond to vowels, and minima correspond to glides. Glides are additionally constrained to appear adjacent to vowels [41]. (Diphthongs such as /ai/, /oi/ and /au/ exhibit different formants between the beginning and end of the phonation region, and should be marked with a vowel landmark at the beginning, and an off-glide landmark at the end. However, in a first pass, automatic detection will most likely yield one landmark for the diphthong. This is an instance where further examination of articulator-bound features results in an update, i.e. an addition, of articulator-free features.) As an illustration, vowel landmarks for the utterance "She can sing" are shown in Fig. 2.3. The utterance does not contain any glides or diphthongs.

Consonants are produced with a complete or very narrow closure by the primary articulators, which is later partially or completely released; this results in two or three discontinuities in the signal. These discontinuities are the consonant landmarks [26]. A discontinuity that leads into a silence interval is a stop closure; one that leads into a region of frication energy is a fricative closure. A discontinuity from a silence region into a burst, optionally followed by aspiration, is a stop release. Similarly, a discontinuity when frication noise ends is a fricative release. Stop and fricative consonants therefore have one closure and one release landmark. Affricates have one closure, similar to a stop closure, and two releases. The first release is similar to a stop release into a frication region, and the latter is similar to a fricative release. A

phonation regions	noise regions
F0 (fundamental frequency)	E1 (energy in F1 region)
H1, H2 (first and second harmonic frequencies and amplitudes)	E2 (energy in F2 region)
F1, A1 (first formant frequency and amplitude)	E3 (energy in F3 region)
F2, A2 (second formant frequency and amplitude)	E4 (energy in F4 region)
F3, A3 (third formant frequency and amplitude)	E5 (energy in F5 region)
F4, A4 (fourth formant frequency and amplitude)	E6 (energy in F6 region)
(F5, A5) (fifth formant frequency and amplitude)	
(F6, A6) (sixth formant frequency and amplitude)	

Table 2.1: List of primary speech measurements

ulation of the speech production system, such as sources of sound, the modulation of the sources, and the variation in time of these characteristics.

Quantities that describe a phonation source include fundamental frequency and harmonic structure. These quantities are useful in determining the configuration of the larynx in phonated regions of speech. These regions also show formant structure.

Aspiration and frication regions are characterized by a source with no distinct harmonic structure, and may be described as noise sources. In regions of aspiration, formant structure may be visible, but frication noise is usually concentrated in a characteristic region, depending on where the noise is generated.

The measurements that may be used to determine acoustic cues from the signal are listed below in Table 2.1. These quantities include the amplitude of the acoustic signal in particular frequency bands. The frequency of vibration of the vocal folds, or fundamental frequency F0, and the formant frequencies up to the fourth formant or higher are important spectral measurements. Additionally, the change in time of these concentrations of energy in frequency must be tracked. These measurements are necessary to determine the regions and characteristics of aspiration, frication and burst release noise in the signal.

Examples of marking these quantities are shown in Fig. 2.2. The phonation regions, i.e. vowels and the nasal consonants, have the formants marked. In these regions, the fundamental frequency and harmonics may also be measured; these quantities have not been marked in Fig. 2.2. The noise regions include intervals of burst and frication noise for obstruent consonants. The concentration of spectral energy is marked for these regions. For example, energy concentration above the third formant

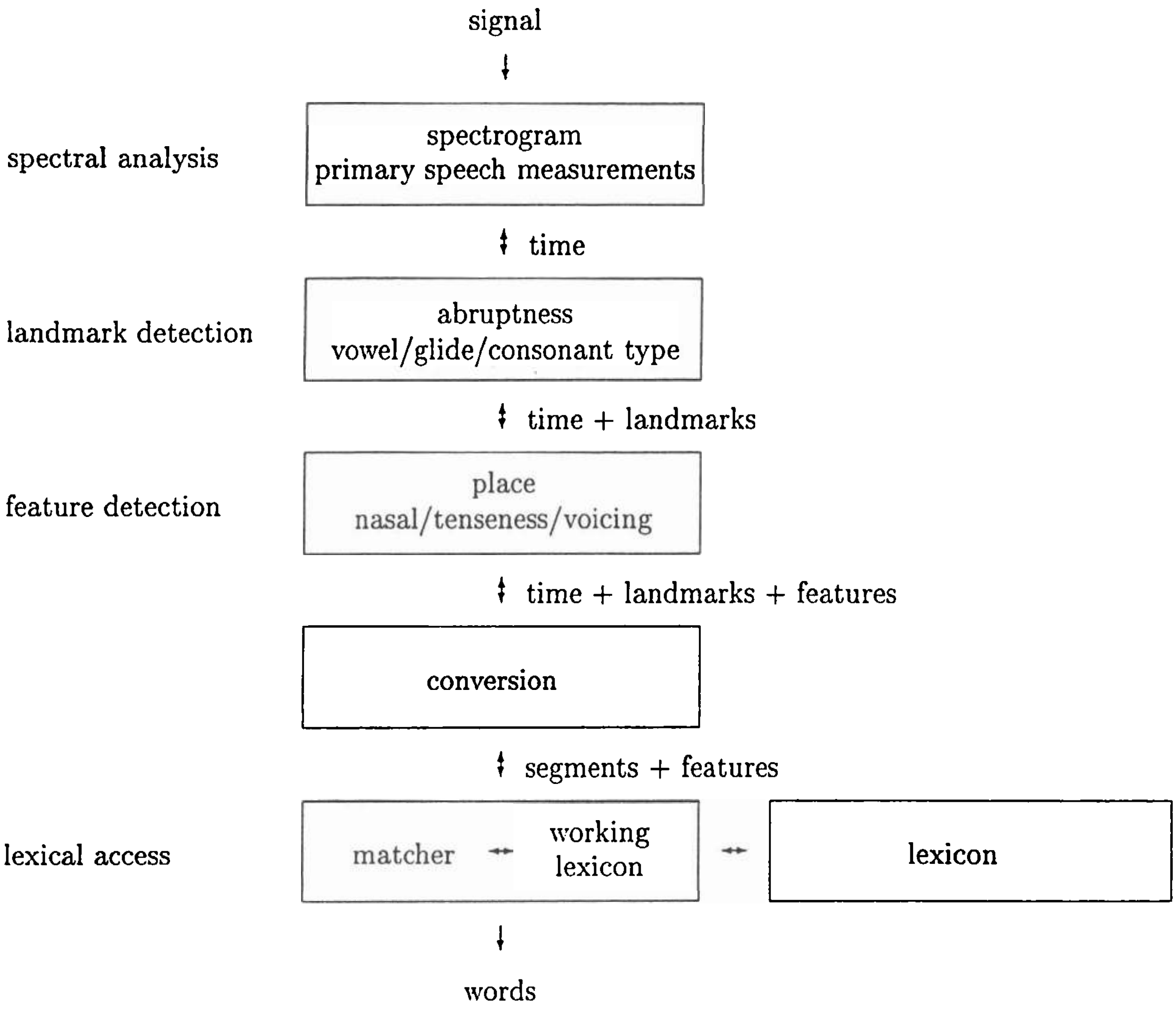


Figure 2.1: Flow diagram of processes for extracting words from the acoustic signal

and times are marked in the signal that may correspond to indicators for underlying segments. Inspecting spectral characteristics at these times leads to determination of landmarks and their types (which may be interpreted to determine the values of articulator-free features for the underlying segment). Further examination of acoustic cues in the signal around the landmarks yields values for corresponding (articulator and articulator-bound) features. The landmarks and features thus found are consolidated in the conversion process to produce a sequence of segments, with their associated features. The sequence of segments are then compared by the matcher with items in the working lexicon to find the best sequence of words. The working lexicon contains items from the canonical lexicon, and also entries that take into account possible augmentations and modifications, which are generated according phonological rules. The interactions between the modules have been schematically represented as bidirectional arrows. This is to indicate that lower-level units become synthesized into larger higher-level units, and that processing required to determine those lower-level units may be guided by the higher-level unit hypotheses. Each structure in the hierarchical system is described in more detail in the following sections.

2.1 Spectral analysis

In human speech perception, the incoming air pressure variations produced by the radiation of the articulated speech sound are processed by the auditory system into a time-frequency-amplitude representation. Similarly, a commonly used representation of the speech signal on machines involves a digitized time-frequency-amplitude function, or a spectrogram. A method for speech recognition by machine should be able to extract measurements such as formant frequencies and presence of frication noise from the digitized spectrogram. This detection process may be guided by higher level information, such as syllabic and segmental contextual information.

In order to determine the acoustic cues from the speech signal, there are various basic quantities that must be extracted from the signal. These primary speech measurements effectively describe the acoustic characteristics that result from manip-

Chapter 2

Overview of a hierarchical speech recognition system

In this chapter, a design for a speech recognition system that is based on the hierarchical feature representation of speech is outlined. Figure 2.1 shows a flow diagram of the processes involved in extracting words from the acoustic signal. First, the signal is transformed into a spectral representation and measurements relevant to speech are obtained. From these measurements, landmarks and acoustic cues are found, which are than consolidated into segments and features. These segments are then used in accessing the lexicon. The processes involved in these steps are carried out by a set of modules. These modules and their interactions will be described in the following sections. As an example of designing a component in this system, a module for detecting consonant voicing will be examined in the detail in this thesis. This chapter provides an overview of the overall structure in which the consonant voicing module will operate. In this system, contextual information is marked in a hierarchically arranged structure as features and higher level symbols (e.g. position of a segment within a syllable). Signal processing modules are used to extract measurements from the signal to infer the values of lower level features, guided by the values of the higher level features and units.

Figure 2.1 shows a schematic diagram of the structures and information flow in the hierarchical system. Spectral analysis is carried out on the input speech signal which to base decisions about which measurements to use in determining consonant voicing. These measurements are used in Chapter 4 to examine isolated utterances, in order to verify the predicted acoustics, and to refine the sets of measurements used. Chapter 5 examines continuous speech using the measurements developed above. The measurements that are described up to this point are made by hand. Chapter 6 describes a scheme for automatic detection of consonant voicing, which closely follows the procedures used in examining the hand measurements, and the results of testing both isolated and continuous speech are given. Finally, discussions of the issues involved in designing a component within the hierarchical speech recognition system and directions for further work are discussed in Chapter 7.

glide, and is characterized by a very low F3 near the local minimum amplitude of the signal.

Consonants are produced with an extreme narrowing or complete closure of the oral tract. Stop consonants show a period of silence, followed by an abrupt burst, which releases into the next segment. Fricatives are characterized by a period of high frequency frication noise, which also releases into the next segment. Sonorant consonants, on the other hand, show an interval of concentration of spectral energy at low frequencies.

The transition patterns of neighboring vowel formants into the closure and out of the release landmarks for consonants, along with the concentration of energy during the closure interval, are good indicators of which articulator (lips, tongue blade, or tongue body) is involved. At the same time, cues such as low frequency energy near the fundamental frequency (F0) may be used as indicators for determining the voicing features, +/- stiff/slack vocal folds and +/- spread/constricted glottis.

These observations provide a background on which to base the design and implementation of a speech recognition system that attempts to follow the perception of speech in humans. Careful examination of the representation, production and acoustics of speech sounds will be used in determining the various levels of representation, the component modules, and processing necessary in each module. The overall flow of information, in which physical quantities in the signal are extracted and interpreted into symbolic speech units, will also attempt to follow that in human perception.

1.4 Outline of the thesis

Chapter 2 describes the overall structure of a hierarchical speech recognition system, based on the processing requirements and relationships among the various units in a representation of speech at the phonetic level. In order to identify the processes involved in designing a component in this system, a module for detecting consonant voicing is implemented in this thesis. Chapter 3 examines the relevant speech production and acoustics related to consonant voicing, as a theoretical background on values of features for that segment are most easily found at these landmarks. These places include regions of maximum and minimum constrictions in phonated intervals for vowels and glides, and discontinuities corresponding to closures and releases for consonants. The characteristics of the signal at both sides of a discontinuity identify the manner of production of the consonant, as well as other features, such as those for place. Finding landmarks corresponds to finding the articulator-free features at the three nodes in Fig. 1.1. At the landmarks, further examination of the signal is carried out, and the type of analysis is selected to be appropriate for that type of segment. Harmonic structure may be examined for phonated regions, while overall energy concentration may be useful in frication regions. Acoustic cues that indicate different features may not be equally reliable, and may be influenced by neighboring context. Those that are directly related to the implementation of a feature during production are usually more robust, and may be regarded as primary acoustic cues. Other cues that result from assisting movements of the articulators are usually more subtle, and may be regarded as secondary acoustic cues. Some acoustic cues that are observed for speech sounds are described below.

Vowels are produced with a maximally open vocal tract within a syllable, and resonant frequencies are observed as concentrations of spectral energy called formants. Various vowels are produced by moving the pharynx, the tongue body, and the lips to change the target resonances and their trajectories. Vowels are classified with features that include high, low, back, advanced tongue root and constricted tongue root. Corresponding acoustic cues that may be used are low first formant frequency (F1) for [+high] vowels, and high first formant frequency for [+low] vowels [32]. Low second formant frequency (F2) may be used to infer [+back] vowels, and extreme F1 and F2 frequencies are found for [+advanced tongue root] or [+constricted tongue root] vowels, with [+constricted tongue root] being allowable for [+low] and [+back] vowels only.

Some acoustic cues for glides, which have a narrowing of the vocal tract, include minimum amplitude of the signal, along with a maximum or minimum of F1 and F2, depending on the glide. The sound /r/ in American English is also classified as a

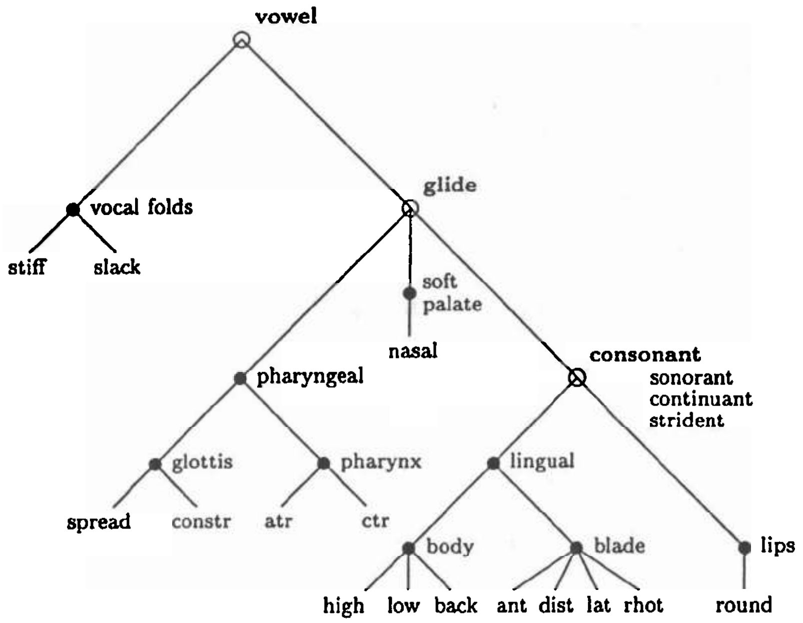


Figure 1.1: Hierarchical feature tree. Open circles indicate articulator-free feature nodes.

tract at some point at or above the larynx. The consonant (or supranasal) node is dominant for consonants, which have a narrow constriction, or closure, in the oral tract. Designation of one of these nodes as dominant, together with specification of the three manner features for consonants provides a specification of the articulator-free features, and indicate the presence of a segment. At each node, there are one or more articulators that may be involved in the production of the speech sound. At each articulator, there are features describing the configuration of the articulator, and these features further define the characteristics of the speech sound. These features correspond to the articulator and the articulator-bound features, respectively. This overall structure for a segment may be implemented as a data structure to be used in automatic labeling.

1.3 Landmarks and acoustic cues

To infer the presence of segments, the acoustic speech signal must be examined to find the corresponding landmarks. Acoustic cues that may be used to determine the lar ridge are marked [+strident]. For these consonants, the airflow directed onto an obstacle results in a sound that is stronger than for fricatives produced at the teeth or with the lips. The articulator-free features are said to describe the *manner* of the speech sound.

The three features named [body], [blade] and [lips] are *articulator* features, and denote which of these three primary articulators is used to produce a constriction in the vocal tract. The articulator features are only specified for consonant segments. Specifying the articulator features for consonants is also referred to as specifying the *place* of the consonant.

The remaining features are the *articulator-bound* features. The features [stiff], [slack], [spread] and [constricted] describe the configuration of the larynx. [Advanced tongue root] and [constricted tongue root] describe the pharyngeal configuration. [Nasal] shows whether the velum is lowered, so that airflow occurs through the nasal tract. The larynx, pharynx and velum are considered to be secondary articulators for consonants. The features [high], [low] and [back] describe the position of the tongue body, and are specified for the vowels and most glides, as well as for [+body] consonants. Constrictions produced by the tongue blade in front of the alveolar ridge are marked [+anterior]. Consonants produced with a wide area of the tongue blade forming the constriction are [+distributed]. [+Lateral] consonants have side paths around the constriction produced by the tongue blade, through which airflow may occur. An example is the consonant /l/. [+Rhotic] segments are produced with the tongue blade bunched up, such as for the glide /r/.

The features described above may be arranged in a hierarchical manner [19], and the geometrical form of this hierarchical arrangement is not dissimilar to the structural relationships between articulators in the speech production system. As shown in Figure 1.1, there are three nodes (marked with open circles), which correspond to the three broad classes of sounds.

The vowel (or root) node indicates that the vocal tract has no major constriction, and this node is designated as dominant for vowel segments. The glide (or supralaryngeal) node is the dominant node for glides, which have a narrowing of the vocal

symbol	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	uh	rr	ex	au	ai	oi	h	w	y	r
vowel	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				
glide														+	+	+	+	+	+	+
cons																				
son																				
cont																				
strid																				
stiff																				
slack																				
spread																	+			
const																				
atr	+	-	+	-	-	-	-	+	-	+	-	-	-	- -	- -	- -		+	+	
ctr	-	-	-	-	-	+	+	-	-	-	-	-	-	+ -	+ -	+ -		-	-	
nasal																				
body																				
blade																				
lips																				
high	+	+	-	-	-	-	-	-	-	+	+	-	-	- +	- +	- +		+	+	-
low	-	-	-	-	+	+	+	-	-	-	-	-	-	+ -	+ -	+ -		-	-	-
back	-	-	-	-	-	+	+	+	+	+	+	+		+ +	+ -	+ -		+	-	+
ant																			-	-
dist															+	+			+	-
lat																				-
rhot												+								+
round							+	+		+	+			+		+		+		

symbol	l	m	n	ng	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	dj	ch
vowel																				
glide																				
cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
son	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
cont	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-	-	- +	- +
strid						-	+	+		-	+	+							+	+
stiff									+	+	+	+				+	+	+		+
slack					+	+	+	+					+	+	+				+	
spread																				
const																				
atr																				
ctr																				
nasal	-	+	+	+																
body				+											+			+		
blade	+		+			+	+	+		+	+	+		+			+		+	+
lips		+			+				+				+			+				
high	-			+											+			+		
low	-			-											-			-		
back	+			+											+			+		
ant	+		+			+	+	-		+	+	-		+			+			
dist	-		-			+	-	+		+	-	+							+	+
lat	+																			
rhot																				
round																				

Table 1.1: Feature chart for standard segments in English

would also be obtained by configuring the vocal tract to approximate a uniform tube, which corresponds to a neutral vowel such as /ex/. The formant frequencies in this configuration depend primarily on the length of the vocal tract, and may be thought of as characteristic of the speaker. A low F1 corresponds to [+high], and a high F1 corresponds to [+low]. An intermediate value is [-high, -low]. A vowel is marked [+back] if the F2 frequency is low, and [-back] if F2 is high. The same criteria are used for the glides /w/, /y/ and /r/. The glide /h/ is produced by aspiration generated at the larynx, so that [+spread] is marked to denote both the place of production and the configuration of the laryngeal source.

Place in consonants denotes which primary articulator was used to produce the constriction. At a consonant landmark with an adjacent phonation region, the movements of the formants may be used to help to determine the place of articulation. Consonants produced with the lips show formants falling in frequency as they approach the consonant landmark. Those produced with the tongue blade show a falling first formant, and a second formant that approaches a target frequency of around 1.8 kHz. Consonants produced with the tongue body tend to show a meeting of the second and third formants.

In addition to these acoustic cues in an adjacent phonation region, spectral characteristics during the frication region, and at the burst may be used, for fricatives and stops, respectively. For fricatives, a concentration of energy in E3+ or higher signals a segment produced with the tongue blade. Those corresponding to E3+ are palatals ([+blade, -ant, +dist]), and E4+ or E5+ correspond to alveolars ([+blade, +ant, -dist]). E6+ (or higher, which would be observed as a dispersion of energy over all frequencies) signals dental fricatives ([+blade, +ant, +dist]). In a similar vein, labial fricatives ([+lips, -round]) show energy concentration dispersed over all frequencies. For stops, the primary articulator may also be the tongue body. The spectral profile at the burst release of a velar stop shows a concentration of energy at E2/E3, and the segment is marked [+body, +high, -low] since the tongue body is raised. The feature [back] is variable according to context – a spectrally lower concentration of energy is [+back], and energy concentration nearer E3 is marked [-back]. The feature

[back] is highly influenced by the adjacent vowel, and may also be inferred from the value of [back] for that vowel. Alveolar stops show most energy in the burst region at E4/E5, and are marked [+blade, +ant, -dist]. Labial stops have a dispersion of energy, similar to labial fricatives, and are marked [+lips, -round]. For sonorant consonants, formant transitions in an adjacent vowel or glide region are usually sufficient to determine the place of articulation. The liquid /l/ is produced with the tongue blade, but the tongue body is used a secondary articulator, and the features [high, low, back] are additionally marked.

Detection of nasality is carried out at a sonorant closure or release. In an adjacent phonation region, an extra peak around 1.0 kHz may be observable, due to an extra pole-zero pair that is produced by coupling the oral and nasal tracts. In the low frequency region, higher formant amplitudes are suppressed.

Tenseness is expressed through the features [atr] and [ctr], for the vowels and glides. A very high or low value for the first and second formant frequencies can be seen for tense vowels and for the glides. For [+/-back,-low] segments, the tongue root (pharyngeal region) is advanced ([+atr]), as for the vowels /iy/, /ey/, /ow/ and /uw/. The vowel /aa/ and /ao/ are both [+back, +low], so that the pharynx is constricted. These segments are marked [+ctr].

The features [stiff, slack, spread, constr] show the configuration of the larynx, and are used to mark the distinction between voiced and unvoiced consonants in English. Voiced consonants are [+slack] vocal folds, and unvoiced consonants are [+stiff] vocal folds. The features [spread] glottis and [constricted] glottis function as "helping" features, in determining voicing for English consonants. A [+spread] or [+constr] segment is perceived as an unvoiced consonant. Voiced consonants show low frequency energy near the fundamental frequency at the closure and release landmarks, as phonation is extended through the closure interval. Also, characteristics of the phonation region adjacent to consonant landmarks may be analyzed to determine the laryngeal configuration leading into and out of the consonant, and hence determine the features for voicing.

As an example of examining the signal to find acoustic cues, the features for two segments, /sh/ and /ih/, have been marked in Fig. 2.4. The fricative closure landmark may be interpreted into the articulator-free features [+cons,-son,+cont] for the segment /sh/. The concentration of energy in E3+ signals a palatal place of articulation, giving rise to the articulator-free feature [+strident], and the articulator feature [+blade]. In addition, the configuration of the primary articulator is specified by the articulator-bound features [-ant, +dist]. The landmark does not show low frequency energy near the fundamental frequency (at the large circle in the figure), so that it is an unvoiced consonant, and marked [+stiff]. For the segment /ih/, the vowel landmark is directly converted into the [+vowel] feature. The first and second formant frequencies are examined, and yield moderately low and moderately high values, respectively, when compared with the average values. Thus, the features [+high, -low, -back] and [+atr, -ctr] are marked.

The processing required to extract landmarks and acoustic cues from the signal, as described in this section, may be implemented through modules dedicated to each task. The total number of modules required is less than ten – three for detecting vowel, glide and consonant landmarks, and one each for place, nasality, tenseness, and voicing. The information that needs to be passed between these modules is conveniently represented by a data structure that is fundamentally identical to the hierarchical structure shown in Fig. 1.1. The modules that determine landmarks are employed first, and the articulator-free features are marked. These landmarks are then further examined for acoustic cues corresponding to the articulator and articulator-bound features by the remaining four modules, to fill in the remaining relevant features. These landmark/feature units are then consolidated into a stream of abstract segments. This process is described in the next section.

2.4 Conversion into segments and features

A vowel of a glide landmark may be directly mapped into a single segment, with the features of that segment taken from the landmark. However, consonant segments have multiple landmarks, each with its set of feature values inferred from the sig-

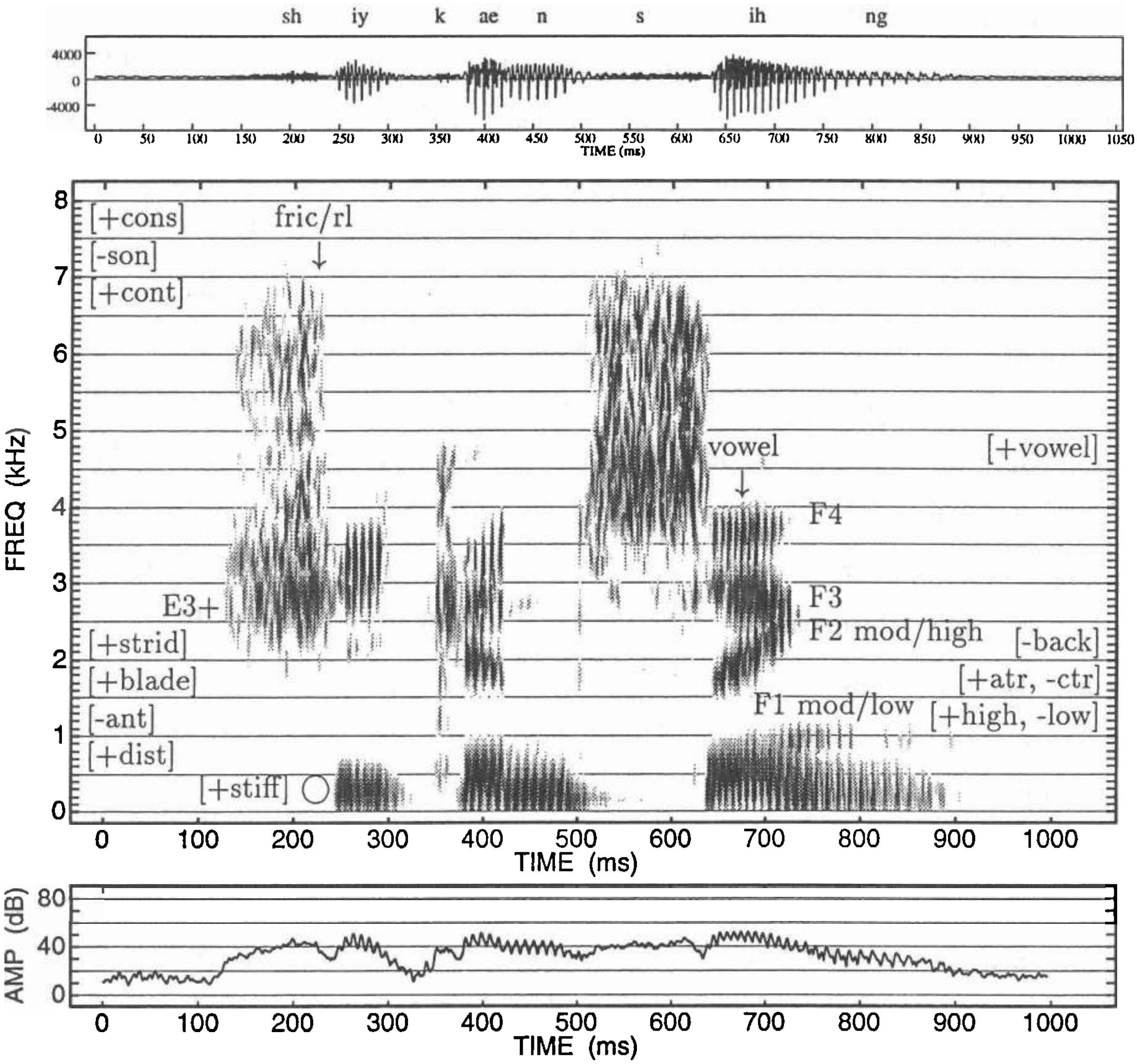


Figure 2.4: Acoustic cues and corresponding feature values for the segments /sh/ and /ih/ in the utterance “She can sing.”

nal. Therefore, a scheme for consolidating the landmarks must be determined for consonants. A simple method is to combine adjacent closures and releases if all the respective features are the same. In practice, it is more probable that features at the closures and releases for the same consonant may differ, due to contextual effects. As such, it may be more reasonable to compare only those features that are distinctive for that type of segment. For example, the [spread] features need not be the same for both the closure and release, although the [stiff] and [slack] features must be in order for the two landmarks to be grouped into a single segment. Various schemes may be possible for the treatment of non-distinctive landmarks that disagree. The features may be dropped, or may both be retained. Alternatively, the possible combinations may be explicitly mapped out to yield the resulting feature values.

The sequences of segments thus obtained will not necessarily match those of words in the canonical lexicon. Differences from the canonical lexicon, however, fall into categories that depend on the type of segment, and its context. These modification rules may be formalized and used to infer the sequence of segments that matches a string of words from the lexicon. A brief description of such a procedure is described next.

2.5 Accessing the lexicon

Items from the lexicon may initially be compared with those given by the string of segments extracted from the signal. It is most probable that a match will not be made directly. In such a case, items from the lexicon may considered in conjunction with the context that is offered by the extracted segments, to identify if the context may give rise to a possible modification of features. If so, lexical items that fit the sequence of segments after modifications occur are retained as possible matches [46]. An example of a modification rule is changing of a dental fricative /dh/ into a dental nasal (a nonstandard segment) if the segment preceding the fricative is a nasal, as in the sequence "in the." Modification rules can be formalized conveniently through the use of the hierarchical arrangement of features. As an example, an alveolar stop /t/

may be produced as a velar stop /k/ in the sequence "late cruise." In this case, all the features remain the same, except that features under the consonant (or supranasal) node of the /t/ assume those of the following /k/. Following the scheme described above results in a final list of possible matches. Further processing using higher level knowledge, such as syntax and semantics, may ultimately reduce the number of possible matches to the sequence that was intended by the speaker.

2.6 Summary

In this chapter, a brief outline of the components and the interrelations among various components in a hierarchical speech recognition system has been presented. The signal is first transformed into a time-frequency-amplitude representation, and regions in the signal that are characteristic of speech are identified, such as phonation and noise intervals. Phonation regions are examined for vowel and/or glide landmarks. Discontinuities between these regions are examined for consonant landmarks. At each landmark, the signal is examined in more detail in order to determine the features that correspond to the acoustic cues that are present. The process for detecting landmarks and acoustic cues may be performed by a small number of modules that are dedicated to each task. The landmark/feature units are then consolidated into segments. These segments are then matched with items in the lexicon, considering the possibility of modifications of features. These modifications are expressed in terms of sets of features affected at specific contexts. This process results in sequences of possible lexical matches, which may further be reduced by higher level knowledge.

A representative example of a module – voicing detection in obstruent consonants – will be examined next. Designing and implementing such a module should take into consideration the underlying physical production mechanisms that are involved, how these actions manifest themselves as acoustic cues, and decisions on which spectral measures are most indicative of such cues. These issues are discussed in the next chapter.

Chapter 3

Examining production and acoustic cues for a consonant voicing module

3.1 Production of consonant voicing

The term "voicing" refers to the distinction made between two classes of segments, where voiced segments exhibit vocal fold vibration as a primary characteristic during production. All vowels, glides and sonorant consonants are voiced, with the exception of the glide /h/, which is produced with an aspirant noise source. 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 but be distinguishable in voicing. In the feature representation of speech, the primary features that describe voicing are stiff/slack vocal folds. In English, the features spread/constricted glottis are additionally used to describe the voicing of consonants in certain phonetic environments. These features are related to the state or configuration of the vocal folds that may encourage or discourage vocal fold vibration during production. The acoustic signal resulting from production of an obstruent consonant shows different characteristics, for example, in low frequency energy corresponding to the promotion or inhibition of vocal fold vibration, and these acoustic cues may be used to infer whether the consonant is underlyingly voiced or voiceless.

As part of the system described in this thesis, a module for detecting the features classifying voicing in consonants will be implemented. In order to implement a module for detecting the features for voicing, production models of voicing will first be examined, and acoustic cues corresponding to articulatory movements that produce voicing will be proposed. These acoustic cues in the signal will be used to infer the features corresponding to voicing in obstruent consonants, and will form the basis of measurements to be used in the following chapters.

3.2 Production models for voicing in obstruent consonants

To produce vibration of the vocal folds for an obstruent consonant, there must be a pressure drop across the glottis sufficient to create a flow of air through the vocal folds, and the vocal folds themselves must be placed together and remain slack [42]. In other words, there are at least three conditions that must be met in order to produce voicing. Meanwhile, to produce an obstruent consonant, there must be a closure or narrow constriction at some point in the oral tract. This leads to buildup of air pressure below the point of closure, so that there is a decrease in the pressure drop across the glottis. If this pressure drop is decreased enough, there will not be sufficient airflow through the glottis and voicing will cease [38].

Therefore, to produce an unvoiced consonant, the air pressure above the glottis is allowed to build up, causing cessation in the airflow through the glottis, and cessation of vocal fold vibration. This situation is also assisted by either spreading the vocal folds apart, or forcing them together into a constriction, or stiffening them. On the other hand, to produce a voiced consonant, it becomes necessary to try to keep the pressure buildup from becoming too great. This may be accomplished by actively expanding the pharyngeal region [1]. At the same time, the vocal folds must be kept together (adducted) and slack.

3.3 Acoustic cues for consonant voicing

The acoustic cues for voicing in consonants may be identified by considering the production mechanism involved. The context in which a consonant occurs must be taken into account in specifying the production mechanism. The contexts examined in this thesis will assume that a non-nasal phonated segment either precedes or follows the consonant or both. These segments in this set include vowels, glides (with the exception of the aspirant /h/), and the sonorant consonant /l/. The nasal consonants have been excluded, as the phonation source in these cases becomes modified by the nasal tract, unlike the other phonated sounds. Other cases which have been excluded are obstruent consonant clusters in which any consonant is not immediately adjacent to a segment in the set described above, such as in “*s*pot” and “pig*s*ty.” It is to be noted that within a syllable in English, if there is a sequence of obstruents, they are all voiced or all voiceless. However, this rule does not apply across word boundaries, as in “hi*s* *f*arm” and “ba*ck* *d*oor” where voicing assimilation may or may not occur – these instances will be of particular interest in analysis of the data in this thesis.

Of the three simple contexts, the first is the case where a consonant is released into a phonated segment at the beginning of a syllable (*syllable-initial*). The second case is where the consonant at the end of a syllable is preceded by a phonated segment (*syllable-final*), and the third case is where the consonant appears between two phonated segments (*intersonorant*).

For the special cases of syllable-initial consonants which are not preceded by a sonorant segment, there is no voicing from the preceding segment, so that evidence for voicing features must be found from the region immediately preceding and following the burst or frication region of the consonant until the onset of voicing of the following segment. For stop consonants, aspiration noise due to a spread glottis is usually present for unvoiced cases, while voiced stops have little or no aspiration noise. On the other hand, vocal fold vibration at low frequencies is observable along with high frequency frication noise during the closure interval for voiced fricatives, and less so for unvoiced fricatives. Also, presence of vocal fold vibration preceding the burst or frication may be observed for voiced obstruents (prevoicing).

For syllable-final consonants which are not followed by a sonorant segment, the acoustic cues are examined in the region leading from the preceding voiced segment into the closure for the consonant. Here, the falloff of vocal fold vibration from the preceding segment is a good indicator of the voicing characteristic of the consonant. Also, secondary acoustic cues such as glottalization due to stiffened vocal folds and adducted glottis may be observed.

Finally, for intersonorant consonants, both the closure and the release out of and into the adjacent sonorant segments may be examined, with their respective acoustic cues.

In addition, it is possible to infer some information about the voicing features of the consonant at regions further away from the consonant landmarks, by observing the attributes of the neighboring segments, such as vowel duration and first formant structure [40, 10]. Shifts in fundamental frequency in a sonorant segment following the consonant release may also be examined for syllable-initial consonants. The degree of stress in neighboring vowel nuclei may also affect the consonant character.

3.4 Measurements for finding acoustic cues for voicing

A module for detecting the acoustic cues for voicing involves extracting several types of information from the signal. One of these measures is the falloff of low frequency amplitude after the closure of the consonant and preceding the onset of vocal fold vibration of the following vowel. Another identifies and measures the intensity of aspiration following the release of the obstruent. The presence of glottalization at the end of voicing of the preceding vowel is another attribute that is also an indicator in determining unvoiced obstruents.

Chapter 4

Acoustic analysis and classification of consonant voicing in isolated utterances

4.1 Description of the database

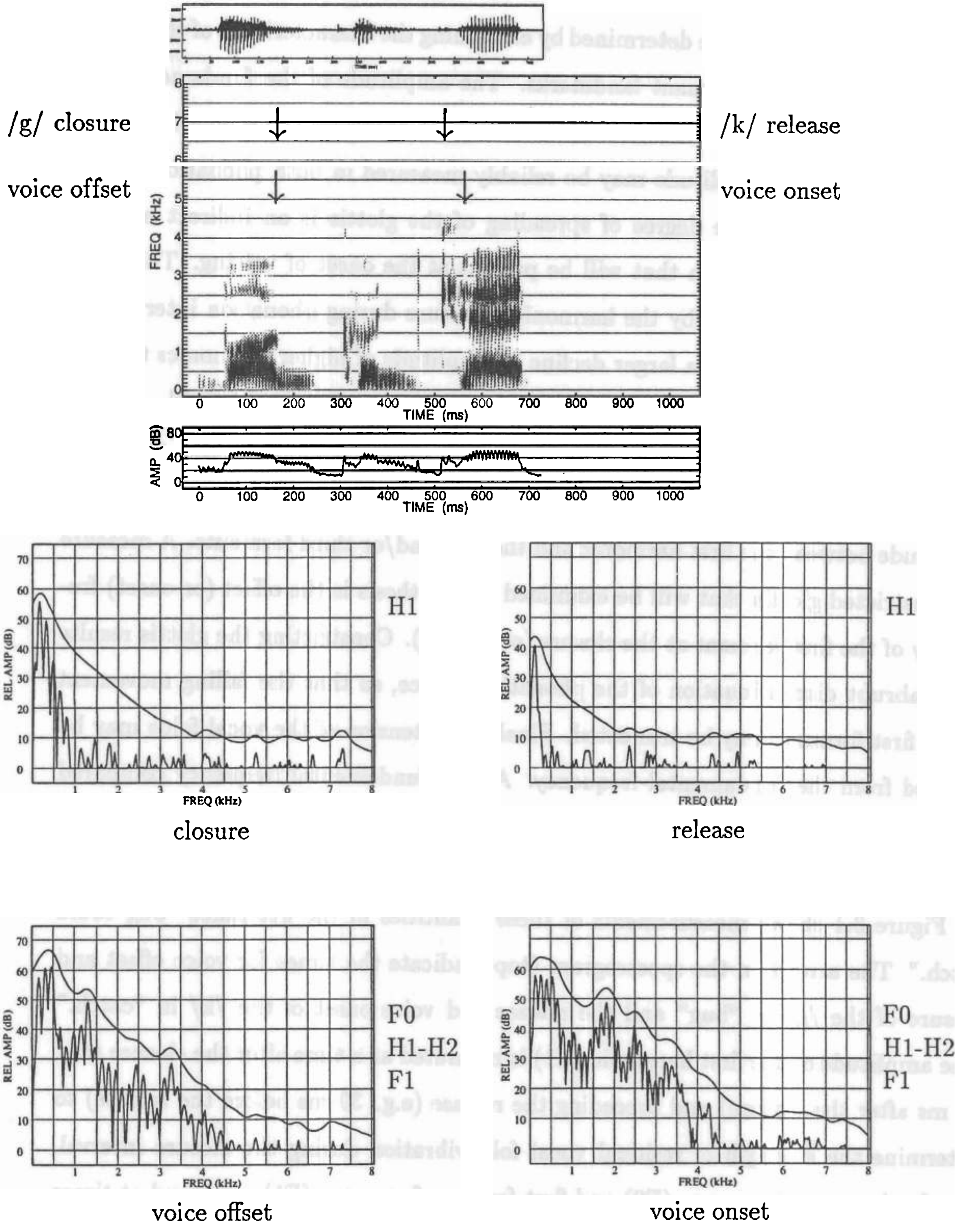
The isolated utterances examined in this chapter have been extracted from a corpus of VCV and CVC utterances. The consonants C are from the set of 16 obstruent consonants, i.e. C = { b, d, g, dj, v, dh, z, zh, p, t, k, ch, f, th, s, sh }. The vowels V are either /aa/ or /eh/. Examples of VCV utterances are /aadaa/ and /ehdjeh/; examples of CVC utterances are /kaak/ and /shehsh/. These utterances were spoken once by two speakers, one male (ks) and one female (cb).

4.2 Measurements

Each utterance was marked with times where the closure and release of the consonant occurred. In addition, the primary spectral measurements described in Chapter 2 were extracted at 100ms intervals centered at each time, at 10ms intervals. In the case of unvoiced stops, measurements were further carried out to include times up to 50ms after the onset of voicing after the release. These measurements include tion, such as syllabic structure, to determine the acoustic cues. For example, estimates of the voicing features for syllable-initial (or word-initial), prestressed consonants will require acoustic cues, and hence, measurements, at the release of the consonant. In this case, measurements at the release will take precedence over those at the preceding closure. It is also possible that in some instances, measurements at the release are all that are available, since the preceding closure may be absent. In this thesis, syllable structure to be used in examining these measurements is determined manually, from orthographic notation and by listening.

3.5 Summary

In this chapter, the mechanisms involved in vocal fold vibration and the production of consonants have been described. The resulting acoustic cues, for different types of consonants, in various contexts have also been discussed. Measurements that show the configuration of the vocal folds and glottis in phonation and during the closure interval have been proposed. These measurements will be further examined in the following chapters, under various contexts, to assess the extent to which they may be used reliably in determining consonant voicing.



closure

release

voice offset

voice onset

Figure 3.1: Measurements for determining consonant voicing in the utterance "bug could catch." Spectra are obtained for times indicated by arrows in the spectrogram, at the voice offset and closure of the /g/ in "bug" and the release and voice onset of the /k/ in "catch."

These quantities may be determined by examining the characteristics of the phonation source near the consonant landmarks. The amplitude of the fundamental frequency (or first harmonic) is a suitable measure for assessing the strength of vocal fold vibration. This amplitude may be reliably measured in both phonation and silence/noise regions. The degree of spreading of the glottis is an indirect measure of the amount of aspiration that will be present at the onset of voicing. This measure may be characterized by the harmonic structure during phonation intervals. A spread glottis gives rise to a larger decline in amplitude of higher harmonics than an adducted glottis. A broader first formant bandwidth may also be observed. Accordingly, the differences in amplitudes between the first two harmonics may provide a good measure for detecting the presence of a spread glottis, as well as the difference in amplitude between the first harmonic and the first and/or third formants. A measure for constricted glottis that will be examined in this thesis is the offset (or onset) frequency of the first formant at the closure (or release). Constricting the glottis results in an abrupt discontinuation of the phonation source, so that the falling movement of the first formant may be truncated. Finally, the tension of the vocal folds may be inferred from the fundamental frequency. A high fundamental frequency compared to the average value for a speaker is the result of stiffened vocal folds, while a low fundamental frequency may signal slackened vocal folds.

Figure 3.1 shows measurements of these quantities in the utterance "bug could catch." The arrows in the spectrogram (top) indicate the times for voice offset and closure of the /g/ in "bug" and the release and voice onset of the /k/ in "catch." The amplitude of the first harmonic (H1) is measured at a time after the closure (e.g. 30 ms after the closure) and preceding the release (e.g. 30 ms before the release) to determine the strength of residual vocal fold vibration during the closure interval. The fundamental frequency (F0) and first formant frequency (F1) are found at times just preceding the voice offset (e.g. 10 ms before the voice offset) and just following the voice onset (e.g. 10 ms after the voice onset). In addition, the relative amplitudes of the first two harmonics (H1-H2) are also found at the voice offset/onset.

These measurements will be examined in conjunction with higher level informathe fundamental frequency, amplitudes of the first two harmonics and the formant frequencies and amplitudes up to the third formant. All measurements described in this chapter were made by hand.

4.3 Extracting acoustic cues

4.3.1 Initial set of measurements

The results of plotting the measurements for determining consonant voicing for a set of VCV utterances are shown in Fig. 4.1. The consonants were spoken in the context of the vowel /aa/ by speaker ks. The first column shows measurements for voiced consonants and the second column, for unvoiced consonants. Each plot shows two regions: the interval around the closure into the consonant is centered at -100ms, and the release region is centered at +100ms. Circles represent fricatives and dots represent stops and affricates. The first two rows show acoustic measures related to the acoustic cues for determining the features [stiff] and [slack]. The next three rows show measures related to the feature [spread], and the last row shows the movements of the first formant, which may be used to infer the feature [constr].

From these plots, it can be seen that in general, there are asymmetries in the rise or fall of the measures at the closure and at the release. For example, in the second row, the average fundamental frequencies at the closure are similar for the voiced and unvoiced consonants, whereas there is a difference of about 20 Hz at the release.

In the first row, it can be noted that a large difference exists in the falloff of H1 after closure and the rise before the release, between voiced and unvoiced consonants. This difference is readily observable at points 30ms after the closure and 30ms before the release. These times are marked with solid vertical lines in the plots.

Outside the interval between the closure and release, i.e. before the closure and after the release, the laryngeal configuration may be inferred by examining measurements that characterize the phonation source. These measurements are shown in rows 2 through 6. Here, although there are differences in the means between the voiced

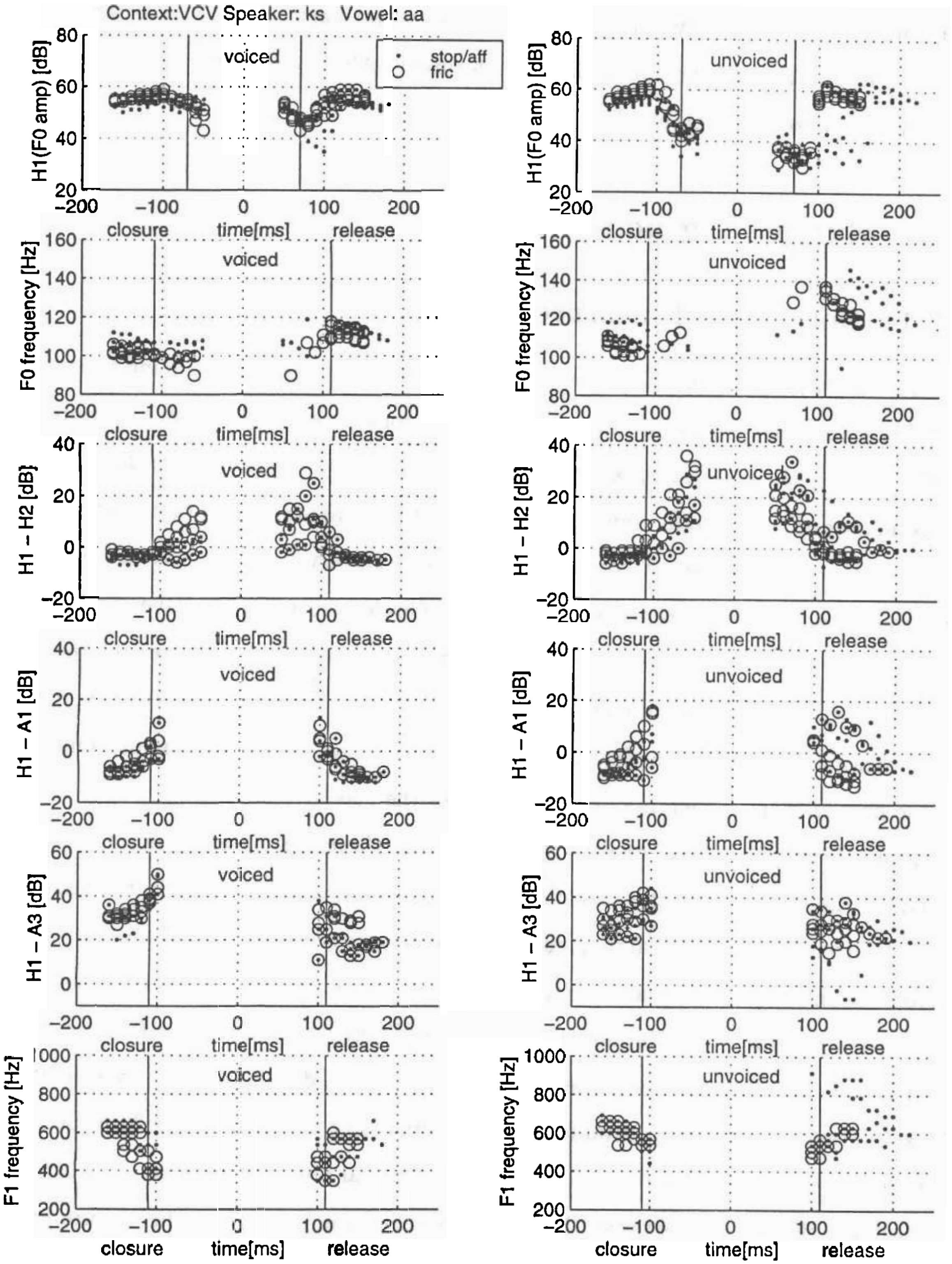


Figure 4.1: Measures for determining consonant voicing at closure and release for voiced (left) and unvoiced (right) consonants in VCV utterances with vowel /aa/ for speaker ks. Dots represent measures for stops, and circles for fricatives. The solid vertical line in the top two plots are at 30ms after the closure and 30ms before the release. Solid vertical lines in the remaining plots are at 10ms before voice offset and 10ms after voice onset.

and unvoiced stops, there are discrepancies between the fricatives and the stops and affricates, especially after the release, as can be seen in the releases of the unvoiced stops. In particular, the measurements of the stops may be interpreted as having been shifted according to the amount equivalent to the time between the release and the onset of voicing. This is seen most clearly for the aspirated (unvoiced) stops, which have a relatively long voice onset time. Accordingly, it becomes important to take measurements that characterize the phonation source after phonation has started, i.e. after the onset of voicing. A time 10ms before offset of voicing and 10ms after onset of voicing have been chosen as suitable times for extracting measurements. These times have been marked with solid vertical lines in rows 2 through 6. (Again, it is to be noted that these times do not align with equivalent times in the case of stops in the plots.)

Next, the measurements in the context of CVC utterances were examined, as shown in Fig. 4.2. As before, the first column shows measurements for voiced consonants and the second column shows those for unvoiced consonants. In this case, however, the release at the end of the first consonant precedes the closure, which leads into the second consonant.

Again, asymmetries exist between the measurements in the release and in the closure. In addition, the releases and closures of the consonants in the CVC utterances do not show the same characteristics as those in VCV utterances. For example, there is a much sharper dropoff of H1 at the closure of voiced consonants in CVC utterances, as seen in the first plot in the figure. The plot for the unvoiced consonants also show a larger range in the difference between the amplitude of H1 before the release and after the closure, when compared to the amplitude during the vowel, which is the same for both contexts. The fundamental frequency at the closures are also somewhat higher than in VCV utterances, although the fact that in both instances, this measurement does not provide a good means of discriminating between voiced and unvoiced consonants, remains the same. The measures for [spread] remain relatively similar, although there seems to be somewhat less variation in the CVC utterances.

Figure 4.3 shows the same measures for VCV utterances where the vowel is /eh/.

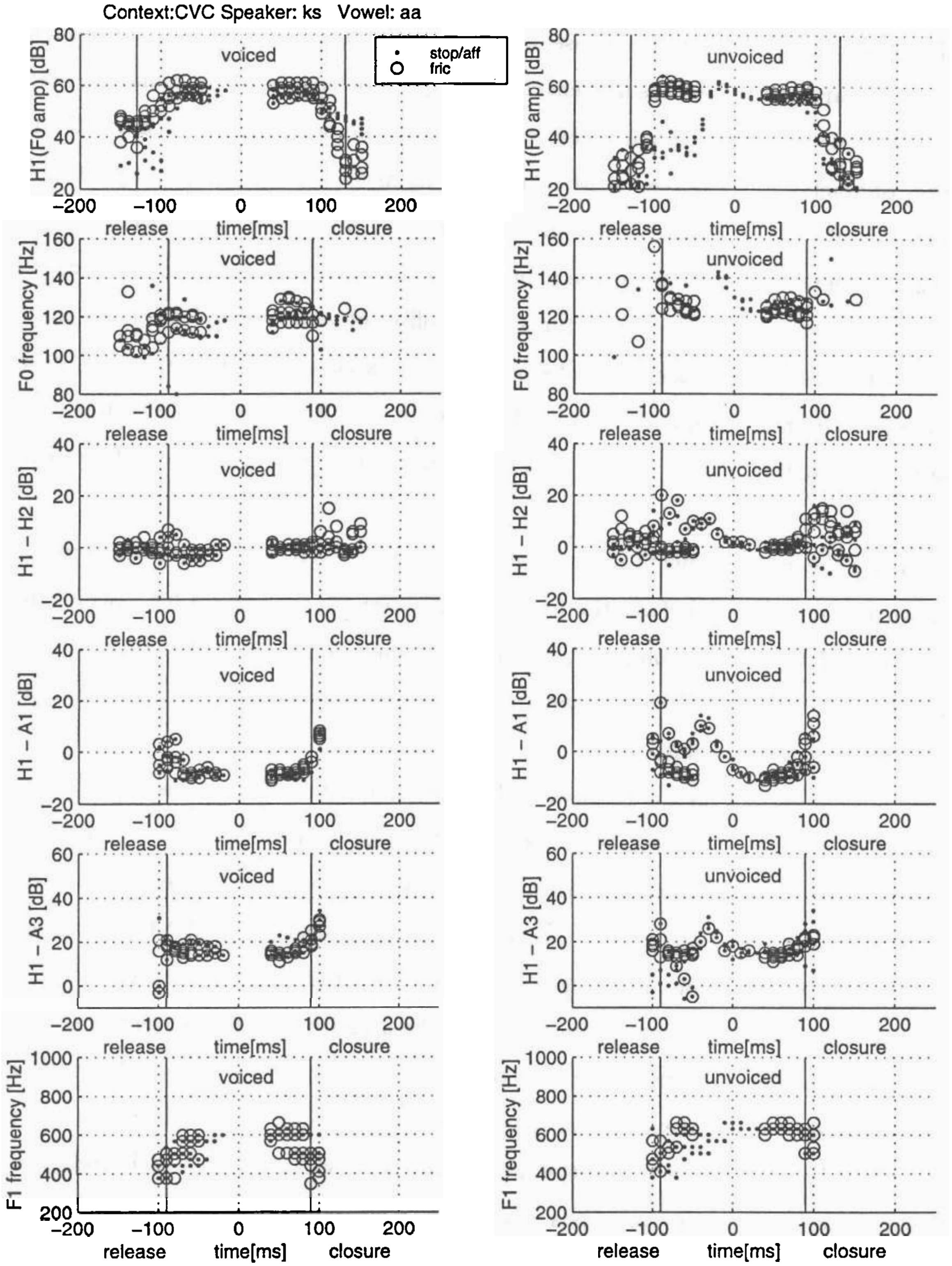


Figure 4.2: Measures for determining consonant voicing at closure and release for voiced (left) and unvoiced (right) consonants in CVC utterances with vowel /aa/ for speaker ks.

Overall, the plots are similar to those in Fig. 4.1, except in the plots for H1 - A3. The values in the case where the vowel is /eh/ are displaced downwards from those where the vowel is /aa/. This may be explained by the fact that the frequency of the second formant is higher for /eh/, so that the amplitude of the third formant is boosted, leading to a smaller difference between H1 and A3. Therefore, it should be noted that this measure may be less reliable than other measures when different vowels occur adjacent to the consonants being examined.

The measurements for CVC utterances with the vowel /eh/ are similar to those of CVC utterances with the vowel /aa/, except for the measure H1 - A3 as described above, but to a lesser degree than in the VCV cases. These results suggest that it may be possible to pool measurements from utterances where vowels adjacent to the consonants may be variable.

The same measurements described above were made for the utterances spoken by the female speaker cb. The plots for VCV utterances with the vowel /aa/ are shown in Fig. 4.4. Comparison with Fig. 4.1 shows that there are differences in the ranges of the fundamental frequency and the first formant, but the overall characteristics are similar. Also, it should be noted that the measures for the feature [spread] are slightly higher than for speaker ks, indicating a more breathy voice in speaker cb. The measurements for the features [stiff] and [slack] are similar, but it is interesting to note that the fundamental frequency at the releases of voiced and unvoiced stops differ less than for speaker ks. On the other hand, the cutoff frequency of the first formant at the closures for voiced and unvoiced consonants are more distinct.

The measurements for CVC utterances for speaker cb are again similar. One of the differences is that there is less distinction in the amplitude of H1 at the closures between voiced and unvoiced consonants. These overall similarities also remain valid for the female speaker in both the VCV and CVC utterances where the vowel is /eh/ instead of /aa/.

In conclusion, these results suggest that measurements used in the preceding figures show similar trends between speakers and vowel environments, as well as in closures and releases. It may also be noted that although a general trend may be

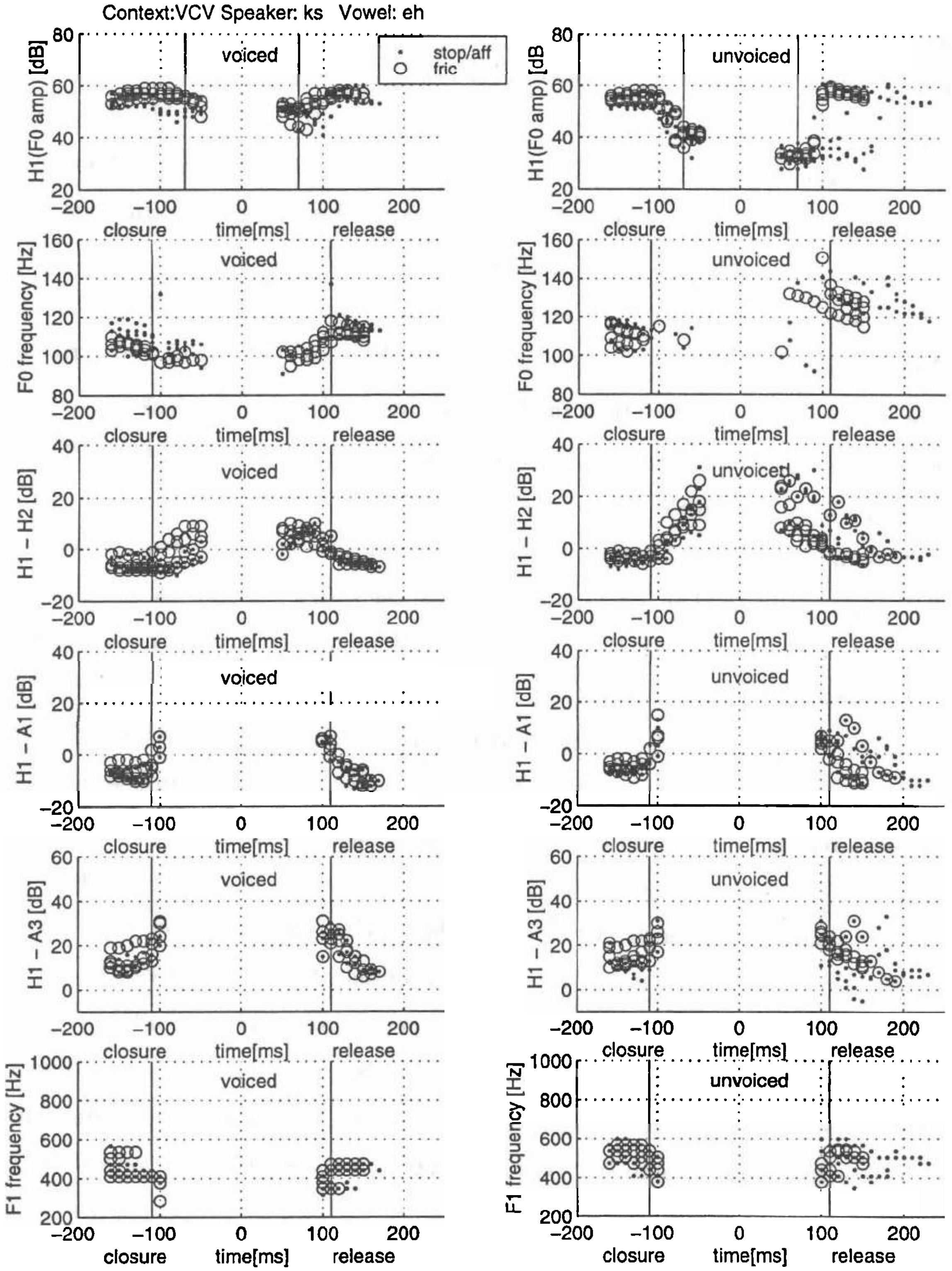


Figure 4.3: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in VCV utterances with vowel /eh/ for speaker ks

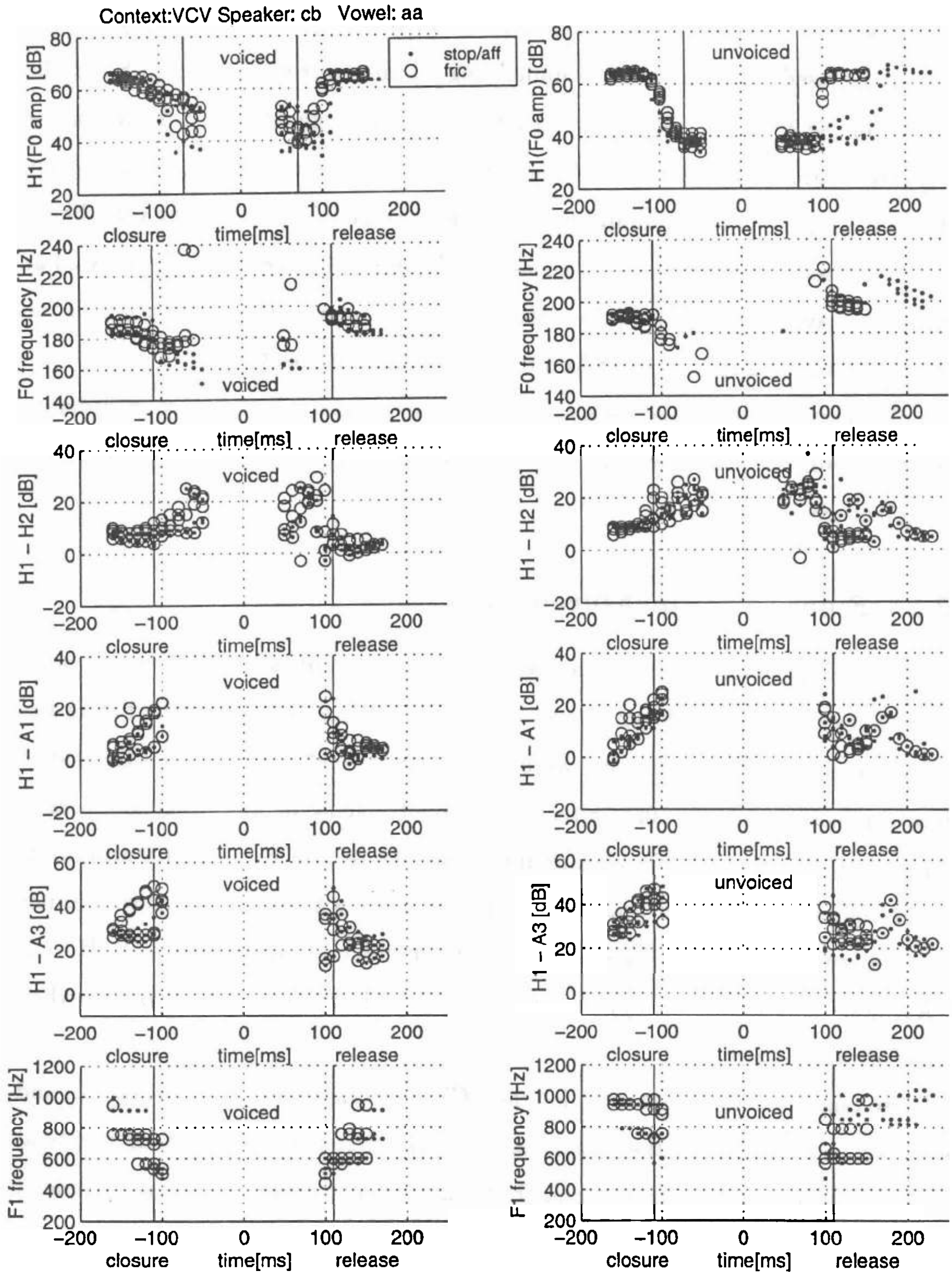


Figure 4.4: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in VCV utterances with vowel /aa/ for a speaker cb.

present, there are individual differences, so that no one measure is completely reliable, but several measurements in conjunction may provide evidence for acoustic cues that may be used to determine the underlying features fairly accurately. Since the character of the adjacent vowel does not affect the measurements critically, it may be possible to examine these utterances together. On the other hand, measures that vary with speaker, such as fundamental frequency and range of the first formant, should be examined separately for different speakers. If these utterances are to be treated as a group, it may be necessary to use measures that are normalized in relation to the average fundamental frequency or formant values for that speaker.

In the next section, these measurements are examined more closely in order to assess the conditions under which they are most useful in determining the underlying voicing features.

4.3.2 Final set of measurements

The measurements described above were taken at the times marked by the solid vertical lines in the plots shown above. These times correspond to -10ms before the offset of voicing from the previous vowel at a closure, +30ms after a closure, -30ms before a release, and at +10ms after the onset of voicing of the vowel following a release. At the times after the closure and before the release, only the H1 amplitude is measured, since the other quantities that characterize a phonation source cannot be reliably measured. Those measurements (fundamental frequency, H1-H2, H1-A1, H1-A3, first formant frequency) were taken at the offset and onset of voicing when phonation was present.

These quantities were measured for the VCV utterances spoken by speaker ks and are shown in Fig. 4.5. Of the upper six plots, the first five show measurements at -10ms before the offset of voicing, and the remaining plot shows the H1 amplitude at +30ms after the closure. Of the six lower plots, the first plot represents the amplitude of H1 at -30ms before the release, and the next five plots are measurements at +10ms after the onset of voicing. The means of the measurements with an adjacent vowel of /aa/ and /eh/ are denoted with a triangle and a square, respectively. The means of utterances with either vowel are denoted by a circle. Each plot has 4 groups of measurements, corresponding to voiced stops, unvoiced stops, voiced fricatives and unvoiced fricatives. The standard deviation of each group is delineated by the short lines, and the stars mark the ranges of the measurements. In grouping consonants into these four groups, the offset of voicing, closure and (burst) release of affricates were grouped with measurements for stops, while the onset of voicing from a frication region into a phonation region was grouped with the measurements for fricatives.

As can be seen from the plots, the H1 amplitudes at +30ms after closure and -30ms before the release show a good separation between the voiced and unvoiced consonants, while being relatively compact about each mean. In general, the measures H1-A1 and H1-A3 show a large variation about the mean, and do not provide good discrimination between the voiced and unvoiced consonants. Of the remaining measurements, the fundamental frequency is a good indicator at the voice onset, but not at the offset of voicing. H1-H2 is a reliable discriminatory measure at the voice onset for stops, but not for fricatives. Finally, the cutoff frequency of the first formant provides some means for distinguishing voicing in fricatives at the offset of voicing, but is strongly dependent on the adjacent vowel.

The plots in Fig. 4.6 for CVC utterances show similar trends as for the VCV utterances. The H1 amplitudes at +30ms after closure and -30ms before the release remain good discriminatory measures, except in the case of fricatives at the closure. This is offset by the reliability of the H1-H2 measure for fricatives at the offset of voicing, as well as the F1 cutoff frequency. The measures for the release are similar, with the H1-H2 measure again showing slightly more separability between the two classes of consonants.

The measurements for the VCV and CVC utterances for the female speaker cb are shown in Fig. 4.7 and 4.8. Overall, the measurements show similar patterns of reliability in determining voicing. However, the relative discriminatory ability of each measure differs from speaker ks. The amplitude of H1 is not as strong a cue; it is difficult to use this measure to judge voicing at the release of both stop and fricative consonants. However, the cutoff frequency of the first formant a the onset of voicing

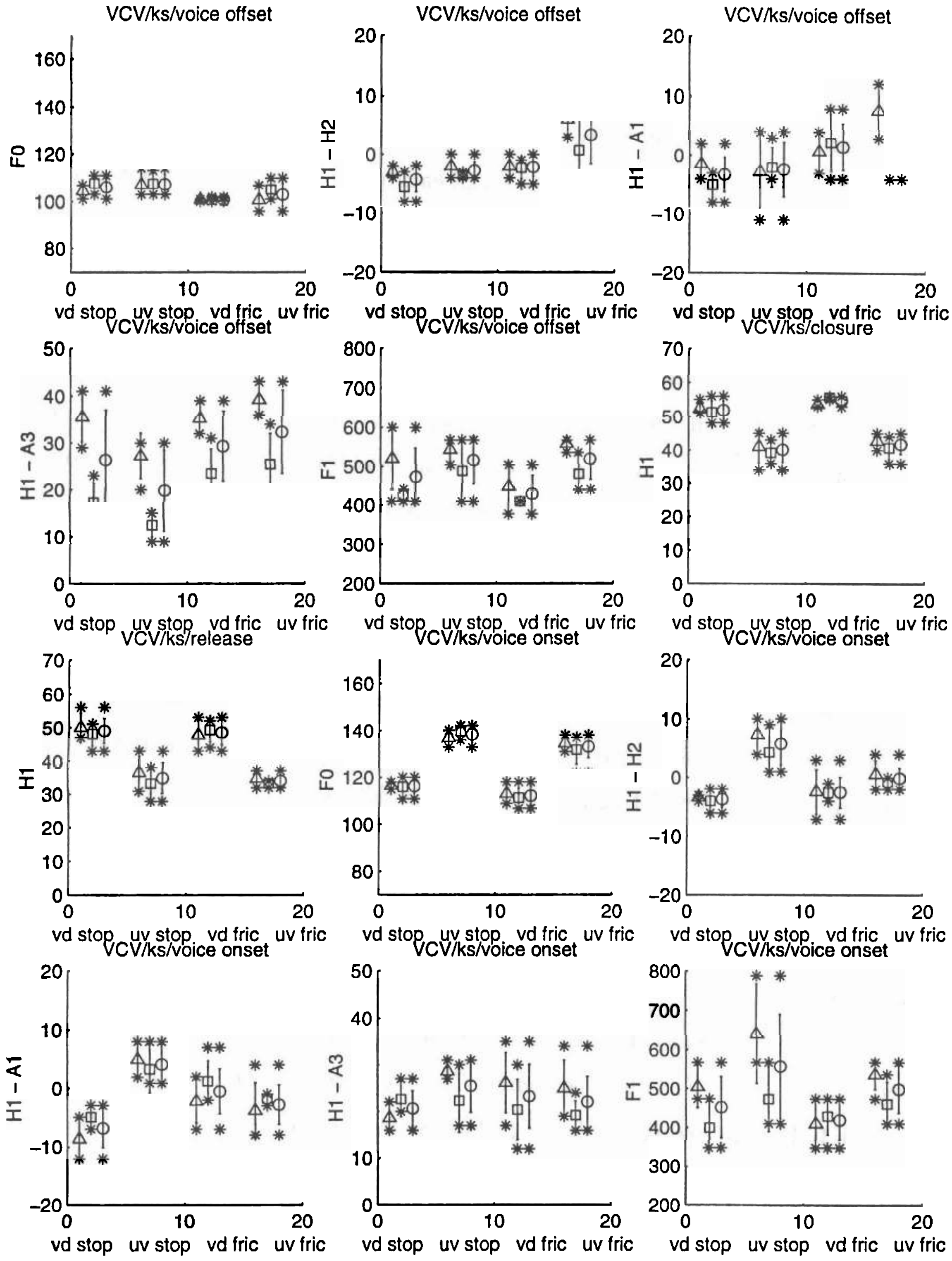


Figure 4.5: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in VCV utterances for speaker ks. Measures for the voice offset and closure are shown in the top six plots. The remaining plots show measures for the release followed by plots for the voice onset. The triangles represent means for the /aa/ utterances, and the squares for /eh/ utterances. The circles are the overall means. Standard deviation is shown by the short lines, and ranges are denoted by the stars. Each of the four groups in a plot shows measures for voiced stops, unvoiced stops, voiced fricatives and unvoiced fricatives, respectively.

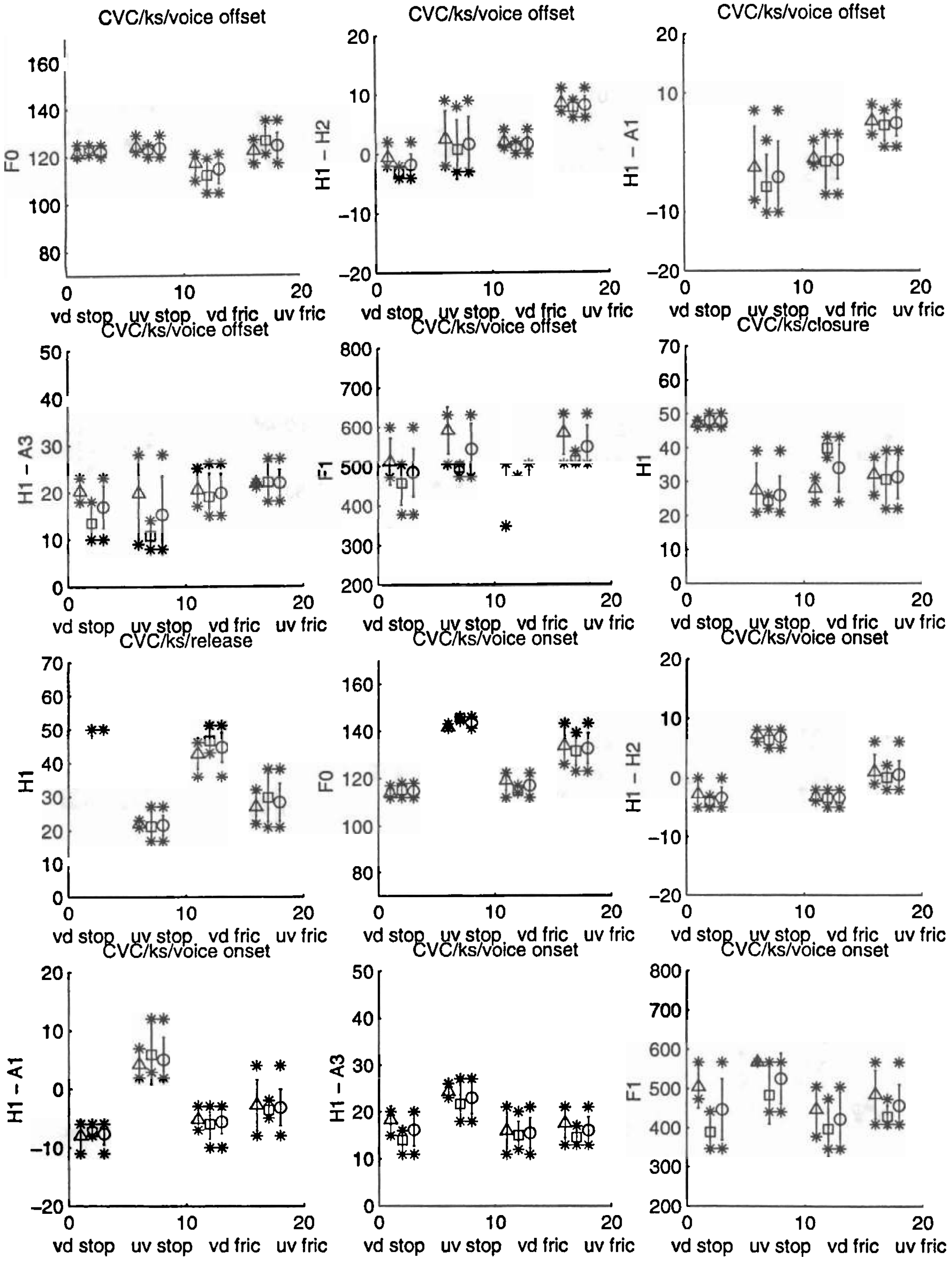


Figure 4.6: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in CVC utterances for speaker ks

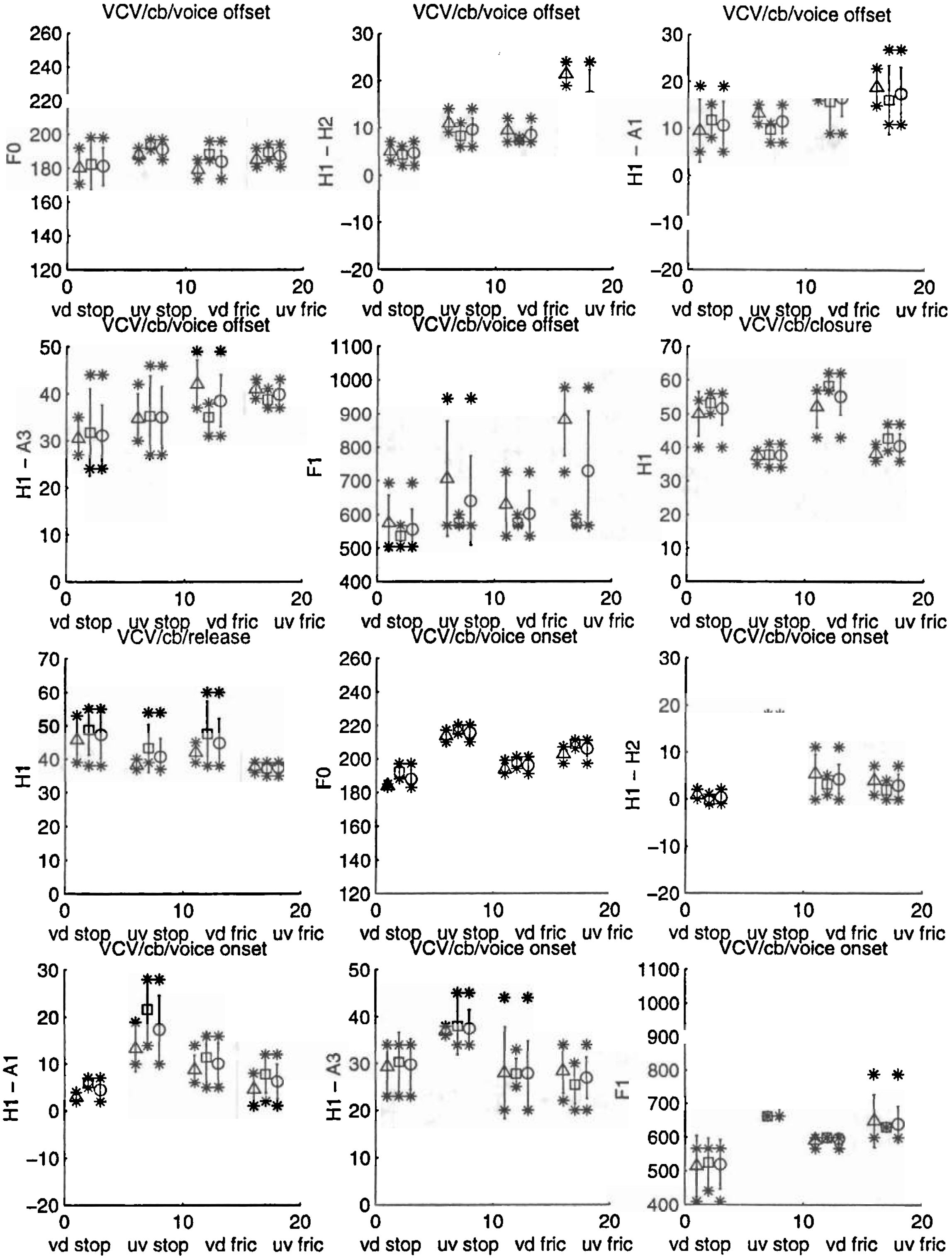


Figure 4.7: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in VCV utterances for speaker cb

training set	LAFF/1-5 (48)	LAFF/6-10 (54)	all (102)
LAFF/1-5	8 (16.7)	14 (25.9)	22 (21.6)
LAFF/6-10	12 (25.0)	13 (24.1)	25 (24.5)
LAFF/1-10	10 (20.8)	10 (20.8)	23 (22.5)
totals	30	37	70

Table 5.1: Classification results for LAFF sentences for speaker ks

training set	LAFF/1-5 (48)	LAFF/6-10 (54)	all (102)
VCV/aa	13 (27.1)	14 (25.9)	27 (26.5)
VCV/eh	13 (27.1)	13 (24.1)	26 (26.5)
VCV/all	12 (25.0)	13 (24.1)	25 (24.5)
CVC/aa	19 (39.6)	22 (40.7)	41 (40.2)
CVC/eh	17 (35.4)	26 (48.1)	43 (42.2)
CVC/all	18 (37.5)	26 (48.1)	44 (43.1)
VCV+CVC/aa	13 (27.1)	20 (37.0)	33 (32.4)
VCV+CVC/eh	14 (29.2)	19 (35.2)	33 (32.4)
VCV+CVC/all	13 (27.1)	20 (37.0)	33 (32.4)
totals	132	173	305

Table 5.2: Classification results for LAFF utterances for speaker ks using isolated utterances for training

disjoint sets.

Next, utterances from isolated data for speaker ks were used in training and classification was carried out for the LAFF sentences. The results are shown in Table 5.2. The error rates are about 26% when VCV utterances are used, but rise to around 43% when CVC utterances are used. Using both sets gives intermediate error rates of 32%. These results show that using a set of selected isolated utterances (such as VCV utterances) in training may yield performance that is comparable to that of training on a subset of continuous speech. The second set of sentences shows a higher error rate than the first five sentences.

The cues for voicing at the closures and releases were next consolidated into voicing decisions for each consonant. The manner and place of closures followed by releases were compared, and if identical, the two landmarks were considered to belong to the same segment, and the measures were averaged to produce the final voicing decision. Using this criterion resulted in closure and release landmarks for geminates and also sequential consonants differing in only voicing to be combined. Examples include overlap between the two classes. Previously reliable measures, such as H1-H2 at voice onset for stops (corresponding to degree of aspiration) are much less separable. The F1 frequency at the offset and onset of voicing appear to be better measures for fricatives at voice offsets and for stops at voice onsets.

The distributions for speaker ss in Fig. 5.2 also show a large spread. H1 amplitudes at closure and release seem good measures for fricatives, but less so for stops, and F0 amplitude seems to show relatively good separation as well. Again, measures for [spread] (H1-H2) and [constricted] (F1 frequency) are much less reliable. In both speaker ks and ss, the H1-A1 and H1-A3 measurements seem less reliable, as for isolated utterances. The measurements shown in this section were included in various sets in the classification experiments discussed next.

5.3 Classification procedure and experiments

The same procedure as discussed in the previous chapter for isolated utterances was used to classify consonant voicing in the continuous speech database. To recapitulate, H1 amplitudes were measured after the closures and before the releases, and F0, H1-H2 and F1 frequency were measured at voice offsets and onsets. The voicing classification from these individual measurements were selected according to landmark type (closure or release) and consonant manner (stop or fricative) and the averaged value was taken as the voicing decision for that landmark.

The first ten sentences were divided into two groups (sentences 1 through 5, and sentences 6 through 10) to examine effects of disjoint training and test sets. The resulting voicing decisions for each landmark were compared with the underlying lexical voicing for the cues associated with the landmarks. The results are shown in Table 5.1. The numbers of landmarks tested are shown in parentheses across the tops of the columns. The number of errors from each trial are given, along with the error rate in percent in the parentheses. Overall, the error rates ranged from 25.9% (train: LAFF/1-5, test:LAFF/6-10) to 16.7%(train:LAFF/1-5, test:LAFF/1-5). As expected, training and testing on the same utterances give better results than using

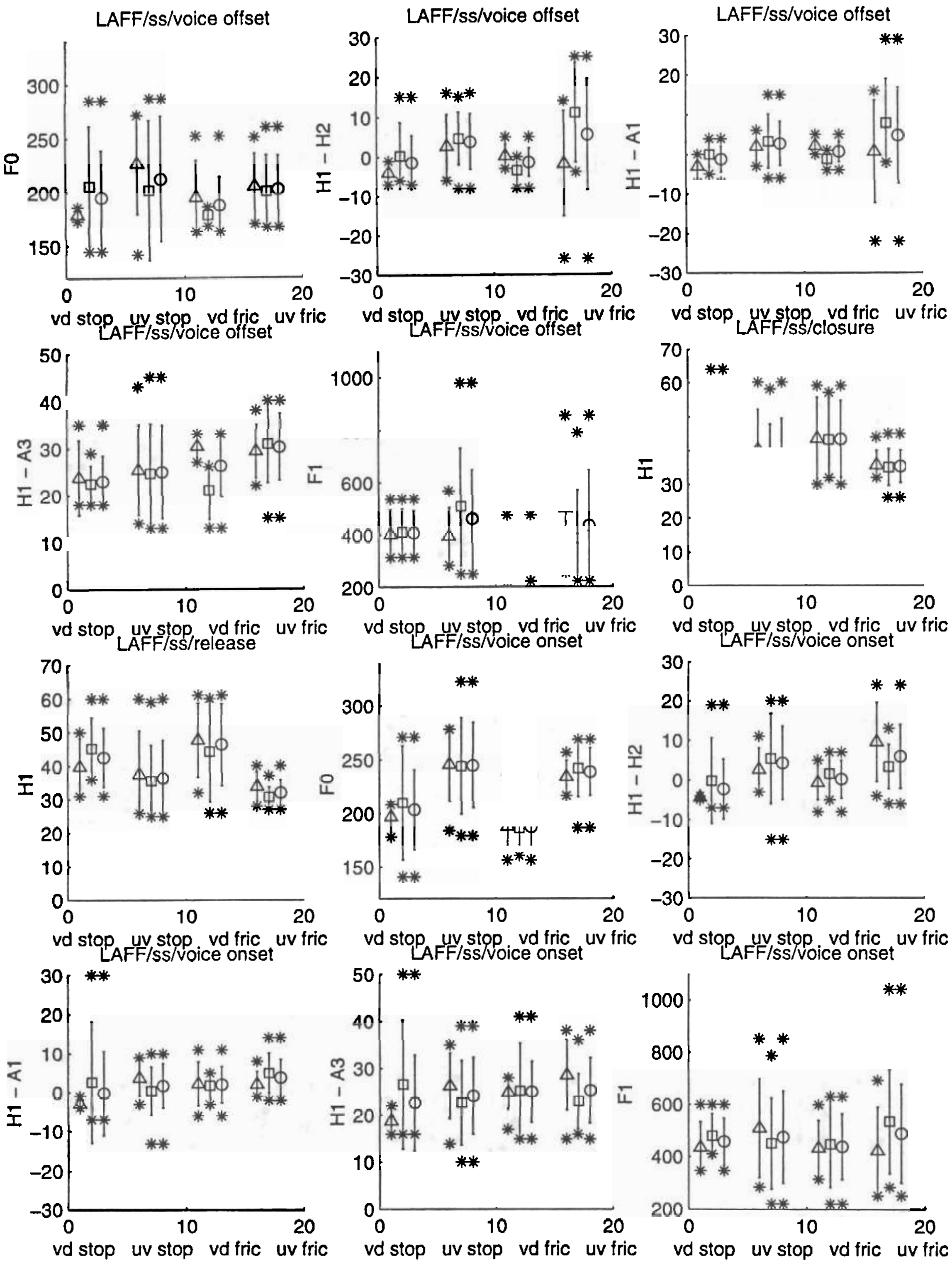


Figure 5.2: Distribution of consonant voicing measures for LAFF sentences for speaker

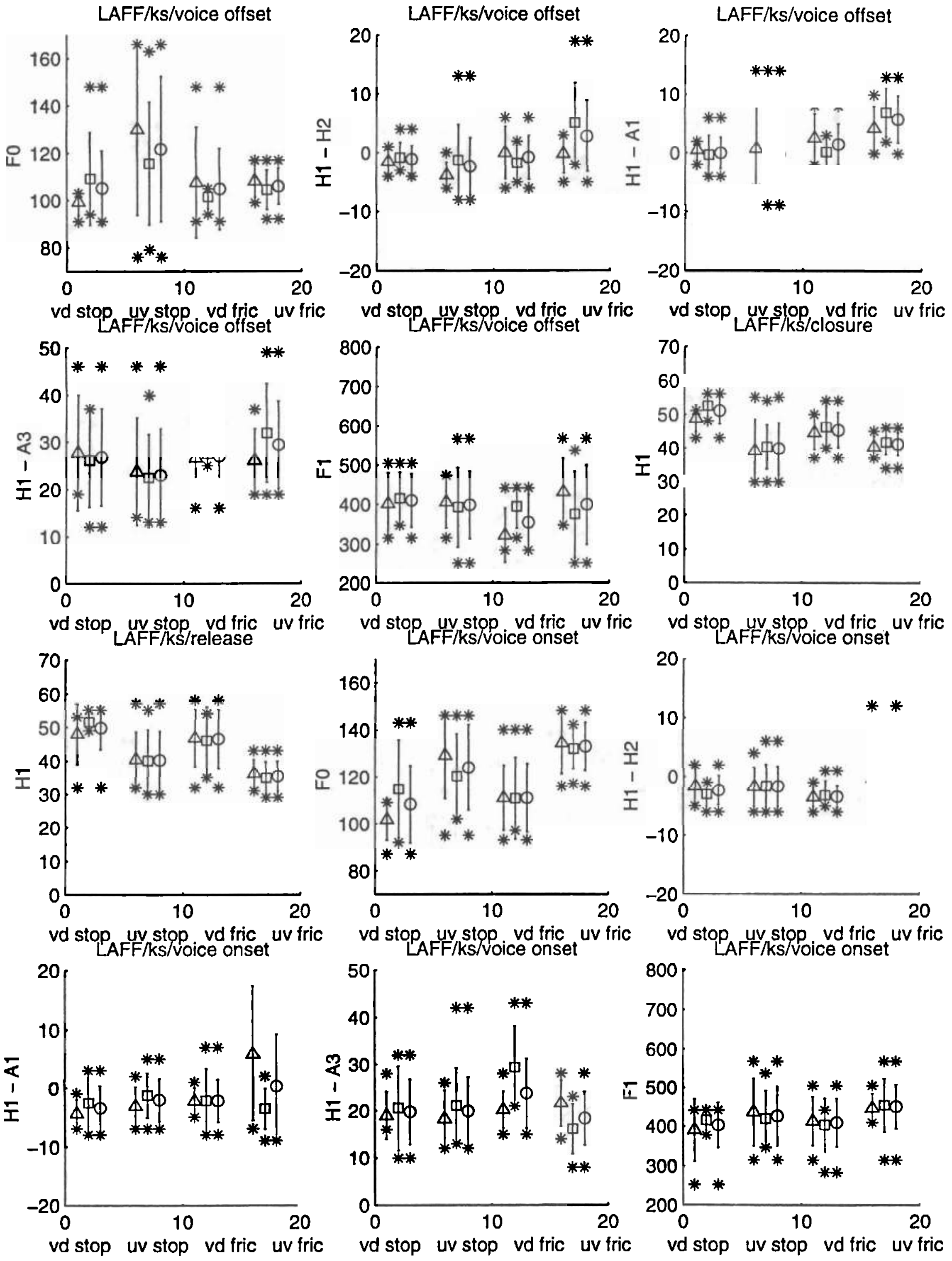


Figure 5.1: Distribution of consonant voicing measures for LAFF sentences for speaker ks

marked II. In the analysis in this chapter, stressed and nonreduced full vowels have been grouped as strong vowels, while reduced vowels are referred to as weak vowels.

The entire database contains 758 consonants, of which 374 are lexically voiced and 384 are unvoiced. A subset consisting of the first ten sentences for speaker ks and ss have been selected and measurements for detecting consonant voicing made by hand. There are 67 consonants in the subset selected, of which 25 are voiced and 42 are unvoiced. This set is further divided into two groups, sentences 1 through 5, and sentences 6 through 10, to study effects of disjoint groups in training and testing. The first set contains 30 consonants (13 voiced, 17 unvoiced) and the second set contains 37 consonants (12 voiced, 25 unvoiced). In the next chapter, a larger set consisting of the first 30 sentences for speakers ks and ss, are used in automatic classification experiments. This larger set contains 228 consonants, of which 112 are voiced and 116 are unvoiced.

5.2 Measurements

The measurements used in this chapter are the same as those used for the isolated utterances: H1 amplitudes are measured at +30ms after release and -30ms before closure, and F0, H-H2, H1-A1, H1-A3 and F1 frequency are found at -10ms before voice offset (at closure) and at +10ms after voiced onset (at release). These measurements were made manually for the first 10 sentences of the LAFF database, and the distributions are shown in Fig. 5.1 and Fig. 5.2. As in the previous chapter, the upper six plots show measurements at the closure and the lower are plots at the release. The distributions of the first five sentences are marked with a triangle for the mean values, and the means of measurements for sentences 6 through 10 are marked with a square. The overall means for all 10 sentences are marked with circles. Again, the short lines show the standard deviations and the stars show the ranges.

As shown in Fig. 5.1, the means are shifted closer together and the distributions are spread over a much wider range than for the isolated utterances in Fig. 4.5. The H1 amplitudes and F0 at voice onset seem to be most robust, although there is much

Chapter 5

Acoustic analysis and classification of consonant voicing in continuous speech

5.1 Description of the database

The method for detecting consonant voicing was tested on a database of continuous speech. The LAFF database consists of 100 grammatically correct sentences, spoken by four speakers. Of these only two speakers, ks and ss, are analyzed in this chapter. The speaker ks is the same as for the isolated utterances. Speaker ss is a different female speaker from speaker cb who produced the isolated utterances. The database contains about 200 words, of one to three syllables, and most contain no consonant cluster.

In order to assess the reliability of consonant voicing detection contexts, the sentences have been marked with information related to syllable structure and simple lexical stress. Each segment in a word is assigned to the onset, nucleus or coda; consonants may also be marked as ambisyllabic. The segments marked as nuclei (vowels) additionally have stress markings. Stressed vowels are assigned a value of I. Reduced vowels are marked III, and full vowels which are neither stressed nor reduced are higher for unvoiced consonants, suggesting a higher incidence of devoicing due to surrounding contexts than for modifications for voiced consonants.

how strongly cues are produced at each landmark.

4.5 Summary

In this chapter, a procedure for selecting measures for determining voicing and using these measures to classify consonants was described. The measures were selected by examining the distributions under different environments, such as landmark type and manner of production of the consonant. Measurements taken from various sets of isolated utterances were then used in classification experiments. Recognition rates ranged from 66 – 100% accuracy for various sets of training and test data. As expected, training and testing on the same set of data shows the best performance, but in certain cases, using a subset that may have better separation between classes results in similar or better performance. Using VCV utterances with ambisyllabic consonants preceded and followed by strong vowels as training sets produced the best results. Vowels with higher first formant frequencies such as /aa/ have less effect on the measurements of the lower harmonic frequencies and also result in better performance when used as training utterances. Therefore, pooling utterances spoken with different adjacent vowels seems to be possible, but may not necessary result in higher recognition rates. The results also show that cues for voicing may be asymmetrically distributed over the release and closure landmarks, and the degree to which each landmark is affected by surrounding context may be dependent on speaker style. For the utterances spoken by the male speaker examined in this chapter, the closure of the coda consonant in the CVC utterances showed more modification by the following unvoiced consonant than for the female speaker. Accordingly, a somewhat larger number of errors occurred at closures than in releases. Fricatives appeared less robust than stops, since the landmarks and the associated offset/onset of voicing may not be coincident, unlike stop consonants. These times may have to be placed separately in order to find measurements where the phonation and frication sources do not affect measurements that are characteristic of each source. Errors between voiced and unvoiced consonants appear to be more dependent on speaker style, but are slightly

training set	landmarks(ks)	segments(ks)	landmarks(cb)	segments(cb)
VCV/aa	2 (3.1)	0 (0)	8 (12.5)	2 (6.3)
VCV/eh	1 (1.6)	0 (0)	8 (12.5)	2 (6.3)
VCV/aa+eh	1 (1.6)	0 (0)	11 (17.2)	1 (3.1)
CVC/aa	15 (23.4)	6 (18.8)	8 (12.5)	4 (12.5)
CVC/eh	19 (29.7)	7 (21.9)	12 (18.8)	4 (12.5)
CVC/aa+eh	20 (31.3)	8 (25.0)	9 (14.1)	5 (15.6)
VCV+CVC/aa	12 (18.8)	4 (12.5)	7 (10.9)	3 (9.4)
VCV+CVC/eh	11 (17.2)	4 (12.5)	9 (14.1)	4 (12.5)
VCV+CVC/aa+eh	12 (18.8)	4 (12.5)	8 (12.5)	4 (12.5)

Table 4.5: Error rates of consolidating closure and release measurements of VCV utterances for speaker ks and speaker cb

closure and a release are present for a consonant, the cues at each landmark may be consolidated. Accordingly, voicing for the VCV utterances was found by averaging the voicing decision values obtained. Equal weight was given to the closure and the release decision values. The results are given in Table 4.5 for speaker ks (left) and cb (right). The number of landmark errors for testing all the VCV utterances for each training set are given in the second and fourth columns, and the number of errors after consolidation into segments are given in the third and last columns. The numbers in parentheses are the error rates in percent, where the total number of landmarks is 64, which corresponds to a total of 32 segments. The results show that overall, performance is increased by 1.6% (training set:VCV/eh or VCV/aa+eh) up to 7.8% (training set:CVC/eh) for speaker ks. Speaker cb shows improvement rates between -1.5% (training set: CVC/aa+eh) to 14.1% (training set:VCV/cb/aa+eh).

From the results, it can be seen that performance is mostly enhanced, but it is also possible for consolidation results to be worse than the landmark results. This is due to cases where one landmark was marginally correct, and the other was strongly indicative of the opposite value, so that consolidation resulted in a wrong decision. For speaker ks, the results from including CVC utterances in the training set showed the most improvement. This is an indication that consolidating measurements may be able to overcome a poor choice of training utterances. The results also show that the cues for a consonant may appear at both the closure and the release, and since each landmark is affected by its context, that the final decision may be dependent on

training set	closure	release	stop	fricative	voiced	unvoiced
VCV/aa	13 (1,3,4,5)	11 (2,2,4,3)	9 (1,0,4,4)	15 (2,5,4,4)	10 (3,1,3,3)	14 (0,4,5,5)
VCV/eh	12 (2,0,5,5)	11 (3,3,3,2)	4 (1,0,1,2)	19 (4,3,7,5)	21 (5,2,7,7)	2 (0,1,1,0)
VCV/all	10 (2,3,3,2)	11 (3,3,2,3)	7 (1,0,2,4)	14 (4,6,3,1)	12 (5,2,1,4)	9 (0,4,4,1)
CVC/aa	8 (2,1,3,2)	8 (3,2,2,1)	5 (1,0,2,2)	11 (4,3,3,1)	9 (5,1,1,2)	7 (0,2,4,1)
CVC/eh	11 (1,4,6,0)	12 (4,3,3,2)	8 (2,0,4,2)	15 (3,7,5,0)	8 (5,2,1,0)	15 (0,5,8,2)
CVC/all	11 (2,2,5,2)	10 (3,2,3,2)	7 (1,0,3,3)	14 (4,4,5,1)	9 (5,1,1,2)	12 (0,3,7,2)
all/aa	10 (1,1,4,4)	9 (3,2,2,2)	6 (1,0,2,3)	13 (3,3,4,3)	8 (4,1,1,2)	11 (0,2,5,4)
all/eh	12 (2,1,4,5)	9 (3,3,2,1)	5 (1,0,2,2)	16 (4,4,4,4)	15 (5,2,3,5)	6 (0,2,3,1)
cb all	8 (2,1,3,2)	8 (3,2,2,1)	5 (1,0,2,2)	11 (4,3,3,1)	9 (5,1,1,2)	7 (0,2,4,1)
totals	95	89	56	128	101	83

Table 4.4: Types of errors in classification results for isolated utterances spoken by speaker cb

there is a slight decrease in errors when both VCV and CVC utterances are used in the training set. In examining errors between stops and fricatives, there seem to be the same rates of improvements for both types, when CVC utterances are used or pooled with VCV training utterances. It is interesting to note that errors between voiced and unvoiced consonants show a larger discrepancy when the training set uses utterances with the vowel /eh/. Also, there are more errors in voiced consonants when the training set consists of VCV utterances, while more unvoiced consonants have errors when CVC utterances are used in training.

The even distribution of errors across closures and releases suggest that the CVC coda consonant closures were affected less by the following unvoiced consonant than for speaker ks. Therefore, the increase in the number of unvoiced consonants is not as great as in speaker ks. The differences in results for training and testing with utterances with the vowels /aa/ and /eh may be because of effects of stress. From informal perceptual listening, it was noted that the utterances containing the tense vowel /aa/ were produced with a stronger stress on the vowel than for CVC utterances with the lax vowel /eh/. The consonants adjacent to the weakly stressed vowel were produced less clearly, so that the cues were not as separable between the voiced and the unvoiced consonants, resulting in more errors for those utterances.

The results given above are for comparing the voicing decision at each landmark with the lexical voicing for the associated consonant. For the case where both a on the closure of the coda consonant, as discussed above. It may also explain the large increase in the number of stop errors when training with the CVC utterances, since voicing for stop closures are determined solely by residual H1 amplitude. The larger number of voiced consonants when VCV utterances are used as the training sets may also be related to the possible modification in the CVC coda consonant closures, since effectively devoiced consonants would be classified wrongly as unvoiced consonants. However, when the devoiced consonants are used in the training set, then contrastingly, the unvoiced consonants will be classified erroneously as voiced consonants.

Overall, there are also more errors for fricatives than for stops. This may be because detection of the landmarks and assisting voicing offset/onset times may be different for fricatives than stops. For the set of experiments conducted, the closure and voice offset times were considered to be the same for both fricatives and stops. However, there are cases where the landmark for the closure of the primary articulator may be displaced slightly from the offset of voicing. This is due to a gradual transition of the sound source from the larynx to the constriction in the oral tract, so that both sources may be observed in the signal for a short period at the closure. In these cases, the closure landmark was placed at the midpoint between the start of frication and the end of observable formant structure. The time difference ranges from about 5 to 20 ms, with more disparity observed in the female speaker cb. A more accurate placement would be to separately mark the times for the closure and the offset of voicing, so that effects of underlying consonant voicing may be observed at times where the sources do not appear simultaneously in the signal.

The different types of errors for classification results for speaker cb are shown in Table 4.4. There are more errors in fricatives than in stops, but unlike speaker ks, errors are more evenly distributed between closures and releases. Slightly more errors are found for the voiced consonants than for unvoiced consonants, but the difference is not large. A slight improvement seems to be observable when CVC utterances are used in the training set, more for closures than releases. There is an increase in the error rates for the test set containing CVC utterances with the vowel /aa/. Overall,

training set	closure	release	stop	fricative	voiced	unvoiced
VCV/aa	7 (0,1,3,3)	2 (0,1,1,0)	2 (0,0,1,1)	7 (0,2,3,2)	8 (0,1,4,3)	1 (0,1,0,0)
VCV/eh	8 (0,0,4,4)	2 (1,0,1,0)	0 (0,0,0,0)	10 (1,0,5,4)	9 (0,0,5,4)	1 (1,0,0,0)
VCV/all	9 (0,0,5,4)	2 (1,0,1,0)	2 (0,0,1,1)	9 (1,0,5,3)	10 (0,0,6,4)	1 (1,0,0,0)
CVC/aa	12 (3,5,2,2)	8 (5,2,0,1)	10 (5,4,1,0)	10 (3,3,1,3)	0 (0,0,0,0)	20 (8,7,2,3)
CVC/eh	20 (7,7,4,2)	6 (3,2,1,0)	10 (5,4,1,0)	16 (5,5,4,2)	3 (0,0,3,0)	23 (10,9,2,2)
CVC/all	20 (7,7,4,2)	8 (4,2,1,1)	10 (5,4,1,0)	18 (6,5,4,3)	2 (0,0,2,0)	26 (11,9,3,3)
all/aa	11 (6,4,1,0)	3 (1,1,1,0)	4 (2,2,0,0)	10 (5,3,2,0)	2 (0,0,2,0)	12 (7,5,0,0)
all/eh	13 (6,3,3,1)	3 (1,1,1,0)	5 (3,2,0,0)	11 (4,2,4,1)	5 (0,0,4,1)	11 (7,4,0,0)
ks all	12 (6,4,2,0)	3 (1,1,1,0)	5 (3,2,0,0)	10 (4,3,3,0)	3 (0,0,3,0)	12 (7,5,0,0)
totals	112	37	48	101	42	107

Table 4.3: Types of errors in classification results for isolated utterances spoken by speaker ks

column, and the number of errors are listed across the rows. The total number of errors is followed by the number occurrences in each of the four test sets – VCV/aa, VCV/eh, CVC/aa and CVC/eh – in parentheses. The last row contains the total number of errors for each column.

Table 4.3 shows that overall, errors were more prevalent in closures rather than releases, in fricatives rather than stops, and more in unvoiced consonants, for speaker ks. The number of errors in releases show a large increase when the CVC utterances are used as the training set, especially for the VCV test data. Including both VCV and CVC utterances reduces the number of errors in the releases of VCV utterances more than for the closures. A similar trend can be seen in the errors between the stops and the fricatives. Although the overall error rate for stops is smaller than that for fricatives, a larger increase in the number of errors in stops for VCV test data can be seen when the training data is CVC utterances, compared to the increase in errors for fricatives. Using both VCV and CVC utterances in the training set resulted in a large reduction in the number of errors in stops, but not for fricatives. Finally, errors for voiced consonants show a decrease in errors when CVC utterances are included in the training set, compared with training only with VCV utterances. However, this seems to occur at the expense of greater errors in the unvoiced consonants.

The larger number of errors for closures, especially when the training data are the CVC utterances may be because of the effect of the following unvoiced consonant

training set	VCV/aa	VCV/eh	VCV/all	CVC/aa	CVC/eh	CVC/all	cb all
VCV/aa	3	3	8	8	8	16	24
VCV/eh	5	5	8	8	7	15	23
VCV/all	5	6	11	5	5	10	21
CVC/aa	5	3	8	5	3	8	16
CVC/eh	5	7	12	9	2	11	23
CVC/all	5	4	9	8	4	12	21
all/aa	4	3	7	6	6	12	19
all/eh	5	4	9	6	6	12	21
cb all	5	3	8	5	3	8	16
totals	42	38	80	60	44	104	184

Table 4.2: Errors in classification results for isolated utterances spoken by speaker cb

in the training set (comparing vertically within a column). Thus, it may be inferred that the closure for the coda consonant in the CVC utterances was affected less by the following /t/. However, the overall recognition rate is worse than for speaker ks, due to the larger variability of the measures, as shown in Figure 4.8.

These results for speakers ks and cb show that, as expected, training and testing on the same set of utterances gives the best performance. However, the results also show that using a carefully chosen subset may provide similar performance. For speaker ks, using utterances with the vowel /aa/ in the training set gives better results than those with the vowel /eh/. Also, the utterances in the context VCV provide a better training set than the CVC utterances. The opposite seems to be the case with speaker cb, where the CVC utterances yielded better results. Overall, pooling across vowels seems to improve reliability for speaker cb, but this is not necessarily true for speaker ks. Considering the distributions of measurements for the different sets of utterances as discussed in Section 4.3.2, it can be seen that using the sets of data that exhibit clearer separation between classes in the training data results in a higher rate of correct classification. Therefore, pooling measurements from data that are less clearly separable may result in degrading the separability of the distributions.

The errors in classification were further analyzed according to landmark type (closure or release), consonant manner (stop or fricative), and voicing. The analysis for speaker ks is summarized in Table 4.3. The training sets are listed in the first with both the preceding and the adjacent vowels being strong vowels. The second consonant is in the context of /aa/ or /eh/ - C - /t/, with the stress pattern being strong - weak. Also, the first consonant comprises the onset of the syllable, while the second consonant is in the coda. In addition, the second consonant is released into a following unvoiced consonant. In the experiments, the CVC measurements for determining voicing are made at the release of the first consonant, which is similar to that in the VCV utterances, but also at the closure of the second consonant, which is in a different context. At the closure of the coda consonant, H1-H2, offset F1 and residual amplitude of H1 is measured for fricative, but only H1 amplitude is measured for stops. Therefore, if anticipation of laryngeal configuration for the following /t/ occurs during the closure interval, the amplitude of H1 may be decreased. This would result in unreliable estimates of H1 amplitude at the closure, and negatively affect recognition rates for closure landmarks. The effects of stress, syllable position and adjacent segments will be considered again in detail in continuous speech in the next chapter.

The general results obtained for speaker cb are somewhat different from those for speaker ks; the results of classification experiments for speaker cb are shown in Table 4.2. Training and testing on the same utterances do not provide perfect classification in both the VCV and CVC cases. Training with utterances with the vowel /aa/ gives better results in the VCV test data, but utterances in the context /eh/ improves performance for CVC test data. Pooling across vowels does not seem to show a consistent effect on performance, but using both VCV's and CVC's in training results in a slight improvement. Overall, the recognition rates ranged from 72% (training set: CVC/eh, test set: CVC/aa) to 94% (training and test set both CVC/eh).

Training with /eh/ utterances yield better results than for speaker ks, and this may be because the first formant frequency for the female speaker cb is higher in relation to the harmonic frequencies than for speaker ks, so that H1 and H2 measurements are affected less. The results show slightly less errors for the CVC test utterances (comparing results across Table 4.2), as well as when the CVC utterances are used

training set	VCV/aa	VCV/eh	VCV/all	CVC/aa	CVC/eh	CVC/all	ks all
VCV/aa	0	2	2	4	3	7	9
VCV/eh	1	0	1	5	4	9	10
VCV/all	1	0	1	6	4	10	11
CVC/aa	8	7	15	2	3	5	20
CVC/eh	10	9	19	5	2	7	26
CVC/all	11	9	20	5	3	8	28
all/aa	7	5	12	2	0	2	14
all/eh	7	4	11	4	1	5	16
ks all	7	5	12	3	0	3	15
totals	52	41	93	36	20	56	149

Table 4.1: Errors in classification results for isolated utterances spoken by speaker ks

of errors over the entire set of test utterances. Training and testing on the same set of CVC utterances do not eliminate errors, and pooling across different vowels does not increase performance. Using measurements from both VCV and CVC utterances results in error rates between that of only using VCV utterances or CVC utterances. The improvement occurs for both the VCV and CVC test sets.

It is interesting to note that using training sets of utterances in the context of the vowel /aa/ shows consistently better performance than using utterances with the vowel /eh/ or when the utterances are pooled. This is because the first formant for the vowel /aa/ is higher in relation to the first two harmonics than in the vowel /eh/, so that first formant frequency affects the H1 and H1-H2 measurements less at the offset/onset of voicing, resulting in more reliable measurements for those cues.

Using only VCV utterances as the training set gives better performance over using CVC utterances or using utterances from both contexts. Overall, classification performances ranged from 66% (training set: CVC/all, test set: VCV/aa) to 100% (same training and test set for VCV utterances; training set: all/aa or ks all, test set: CVC/all). The difference in performance may be due to the contexts of the utterance sets. The closure and release measurements are for the same ambisyllabic consonant in the case of VCV utterances, and both landmarks are adjacent to a strong vowel. On the other hand, CVC utterances are excised from the carrier sentence, "Say CVC today." The first consonant is in the context of /ey/ - C - /aa/ or /eh/, into account F0 frequency. For each individual measurement, a voiced decision is given a value of one, while an unvoiced decision has a value of zero. For both closures and releases, a decision resulting from the H1 amplitude is given a weight of two, while each of the other measures are weighted with a value of one. The decisions are then combined and the average value is taken as the overall voicing decision. The results of classification experiments using this scheme is described in the following section.

4.4.2 Results

Measurements from the isolated utterances were used to obtain the means described above, to assess the performance of the procedure described above. Classification results using utterances spoken by speaker ks are shown in Table 4.1. The subsets of the isolated utterances used to set the means are shown under the column listing the training sets. A training set may include utterances from VCV or CVC utterances, in the context of the vowels /aa/ or /eh/, or a pooled set across vowels or VCV/CVC contexts. The classification errors resulting from the trained measurements are shown for various test sets along each row. Each entry in columns 2, 3, 5 and 6 is the number of errors out of a possible 32, which is the voicing decision of a total of 16 consonants, each at the closure and release. These columns are marked in the form of VCV or CVC / aa or eh. The sums of errors combining results from two possible vowel contexts are tallied in columns 4 and 7, and the overall number of errors across the utterances for the speaker are shown in the last column. The last row contains the sums of errors for different subsets of the test utterances.

From Table 4.1, it can be seen that training and testing on the same utterances for the VCV data results in no errors, and testing on different utterances yields only slight decrease in performance. Pooling the utterances with different vowels also gives similar results. There is a noticeable increase in the number of errors when these measurements are used to classify the CVC utterances, and pooling of the utterances with both vowels does not decrease the error rate. However, if the CVC utterances are used as the training data, there is a much larger increase in the number appears to be a relatively good measure, unlike the measurements for speaker ks. The other quantities, F0 amplitude for both stops and fricatives, and H1-H2 for stops, at the onset of voicing remain relatively well separated.

From these results, it can be seen that a relatively small set of measurements at points near the closure and release may be sufficient to distinguish between voiced and unvoiced consonants. These measurements are: H1 amplitude at +30ms after closure and -30ms before release; F1 cutoff frequency and/or H1-H2 at the offset of voicing for fricatives; and H1-H2 for stops, and F0 frequency for both stops and fricatives at the onset of voicing. These measures have been used in classification experiments that will be described in the next section.

4.4 Classification experiments

4.4.1 Procedure

The measurements described above were used to determine the voicing of consonants in the VCV and CVC utterances. The means for the H1 amplitude at closure and release, and means for F0 frequency, H1-H2 and F1 frequency were found for voiced and unvoiced stops and fricatives. Measurements from test utterances were then compared with the means and classified with the closest group. The individual decisions from each measurement were then interpreted according to manner (i.e. stops or fricatives) and landmark type (i.e. closure or release) to produce a voicing decision for each closure and release. In combining the decisions, the decisions for measurements at the offset of voicing and after the closure are consolidated into a single decision at the closure, and a similar grouping is made for the release and voice onset measurements.

For stops, the H1 amplitude is used to detect voicing at the closure, while fricatives additionally consider H1-H2 and the cutoff F1 measurements. At the release, H1 amplitude is again used for both stops and fricatives. In addition, at the onset of voicing, stops consider measures for F0 frequency and H1-H2, but fricatives only take

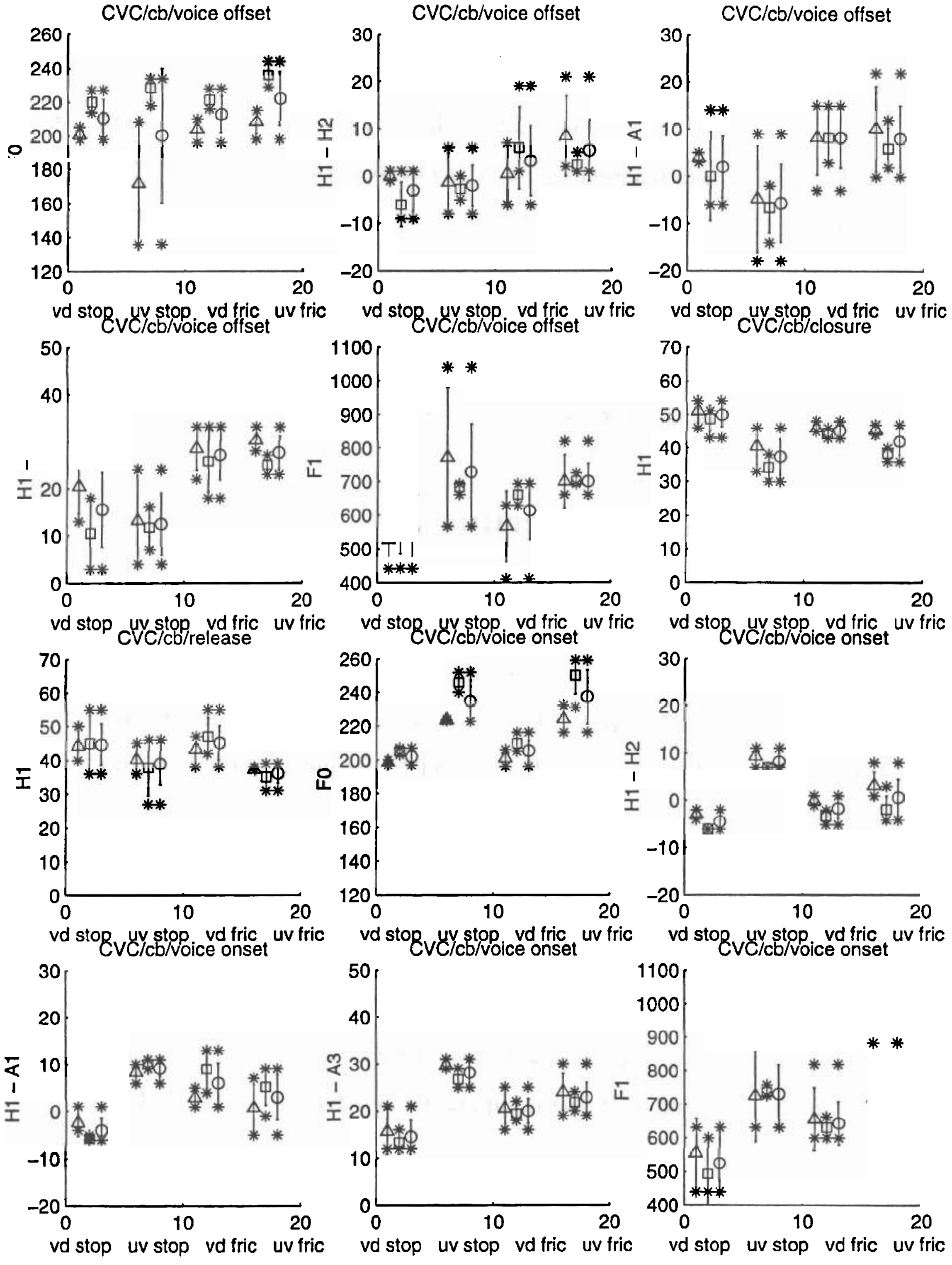


Figure 4.8: Measures for determining consonant voicing at closure and release for voiced and unvoiced consonants in CVC utterances for speaker cb

training set	lm (102)	- err/mod	- err/mod	sm (65)	- err/mod	- err/mod
		(12)	+ cor/mod		(8)	+ cor/mod
VCV/aa	26.5	16.7	18.6	18.5	12.3	18.5
VCV/eh	25.5	15.7	17.6	18.5	12.3	18.5
VCV/all	24.5	14.7	16.7	18.5	12.3	18.5
CVC/aa	40.2	32.4	36.3	29.2	23.1	29.2
CVC/eh	42.1	33.3	36.3	33.8	27.7	33.8
CVC/all	43.1	34.3	37.3	35.4	29.2	35.4
VCV+CVC/aa	32.4	23.5	26.5	26.2	20.0	26.2
VCV+CVC/eh	32.4	22.5	24.5	26.2	20.0	26.2
VCV+CVC/all	32.4	23.5	26.5	26.2	20.0	26.2
LAFF/1-5	21.6	11.8	13.7	16.9	10.7	16.9
LAFF/6-10	24.5	14.7	16.7	18.5	12.3	18.5
LAFF/1-10	22.5	12.7	14.7	18.5	12.3	18.5

Table 5.3: Classification results for LAFF utterances for speaker ks using isolated utterances for training: effect of combining closure and release measurements and evaluation with perceived voicing

geminates such as "ta*k*e *c*aution," and the /z//s/ sequence in "i*s* *s*omething." In both cases, further knowledge of the position of the consonant within the affiliated syllable is needed to recognize the presence of two segments and prevent consolidation of landmarks, but this was not carried out in this thesis. The landmark error rates and the error rates after combining in closure and release landmarks for consonants for the ten sentences are given in Table 5.3.

The total number of landmarks is 102, as before, and the number of segments is found to be 65. The total number of segments found is less by two from the actual number. This is due to counting two occurrences of geminates as single consonants. Sequential consonants differing in voicing were counted separately in evaluation, so that a single voicing decision found for this closure/release pair always produced one segmental error. The overall results show that in all cases, the error rate decreased – from 4% (train:LAFF/1-10) to 11% (train:CVC/aa). The larger improvements occurred for the CVC training sets, which were identified as poor training utterances previously.

Up to this point, evaluation of voicing detection was carried out by comparison with the underlying lexical voicing for each consonant. However, the sentences contained consonant whose actual realization resulted in modification from the lexical voicing description. An informal perception test showed 8 modified consonants (corresponding to 12 landmarks), of which 4 consonants were flapped /t/'s as in "city." In this case, the unvoiced segment effectively becomes a glide, which is nondistinctively voiced. The remaining 4 cases included devoicing (assimilation), as in "i*s* *s*omething." The results of discounting the errors corresponding to these cases are shown in columns 3 and 6 of Table 5.3. In all cases, error rates improved, by about 9% for the landmark results and 6% for the segmental results. However, the "improvements" also included cases where consonants that were actually modified were determined as showing the underlying lexical voicing. When these cases are counted as errors, the resulting error rates are higher, as shown in columns 4 and 7 for the landmarks and segments, respectively. The landmark error rates still show improvement over evaluation with the lexical voicing, but the results for the segments show no improvement. Thus, it may be concluded that consolidating measures for closures and releases overcome most of the voicing modifications that may have occurred in landmarks, if these modifications are not "strong." In other words, the modifications may still retain residual cues for the underlying voicing for one or both of the closure and release landmarks, so that when the measures are combined into segments, the underlying voicing may be recovered. However, for the case of "strong" modifications, such as flapped /t/'s, the underlying voicing is not recoverable. The flapped /t/'s account for about 8% of the landmarks and 6% for the segments.

The errors obtained for both sets of training speech were further analyzed according to landmark type, and the distributions are shown in Table 5.4. The types of errors are listed across the top row and the number beneath in parentheses is the total number of that type of landmark in the first 10 sentences of the LAFF database. The errors in the results when the training sets are the LAFF sentences are shown in the upper portion, and the errors when isolated utterances are used are shown in the lower portion. Overall, there is a more or less even distribution of errors across closures and releases, and also between stops and fricatives. However, there seems to be relatively more errors for unvoiced consonants than for voiced consonants, especially

training set	closure	release	stop	fricative	voiced	unvoiced
	(47)	(55)	(55)	(47)	(41)	(61)
LAFF/1-5	12	15	12	16	6	22
LAFF/6-10	15	15	14	16	11	19
LAFF/1-10	12	15	14	13	7	20
VCV/aa	12	15	14	13	8	19
VCV/eh	12	15	14	13	8	19
VCV/all	12	15	14	13	8	19
CVC/aa	17	24	23	18	2	39
CVC/eh	21	22	24	19	3	40
CVC/all	21	23	24	20	3	41
VCV+CVC/aa	13	20	15	18	3	30
VCV+CVC/eh	13	20	17	16	5	28
VCV+CVC/all	12	21	17	16	3	30

Table 5.4: Error analysis for classification results for speaker ks

when the isolated utterances are used as the training set. Interestingly, the number of errors for unvoiced consonants increases by a larger amount compared to the other types of errors when CVC utterances are used in training.

The errors were also broken down according to syllabic context and are shown in Table 5.5. The consonant may occupy the place in the onset or coda in the syllable, or may be ambisyllabic. Consonants in the onset may be followed by a strong or a weak vowel; consonants in the coda may be preceded by either a strong or weak vowel. Ambisyllabic consonants may have a strong preceding vowel and a weak following vowel, or vice versa. The numbers of landmarks in each category are listed under the types in parentheses. Overall, there are more consonants in the onset of the syllable than in the coda or in ambisyllabic position. Also, there are more consonants that are associated with strong syllables than weak ones – this may be due to the fact that the preponderance of words in the database are monosyllabic or disyllabic.

The results show that the most errors occur in ambisyllabic consonants preceded by a strong vowel and followed by a weak vowel. This is true for all sets of training data. When CVC isolated utterances are used as training data, there is also a larger number of errors for consonants associated with strong vowels, both in the onset and in the coda. Closer inspection of the errors for strong-ambisyllabic-weak consonants

training set	onset-st (29)	onset-wk (22)	st-ambi-wk (22)	wk-ambi-st (8)	st-coda (17)	wk-coda (4)
LAFF/1-5	5	7	12	0	2	2
LAFF/6-10	5	6	12	1	4	2
LAFF/1-10	5	6	12	0	3	1
VCV/aa	6	6	12	1	3	2
VCV/eh	4	6	12	0	3	2
VCV/all	4	6	12	0	3	2
CVC/aa	14	8	11	4	8	1
CVC/eh	11	10	13	3	9	2
CVC/all	11	10	13	3	9	2
VCV+CVC/aa	9	8	12	2	5	2
VCV+CVC/eh	7	8	12	2	6	3
VCV+CVC/all	8	8	12	2	6	2

Table 5.5: Analysis of classification errors into syllabic context for speaker ks

shows that a large number of these consonants undergo a reduction, such as flapping of an underlying segment /t/ into a flapped /t/. This accounts for 6 landmarks that are considered as errors. The other 2 landmarks for the remaining occurrence of a flap is include in the st-coda environment, from the sequence "wri*te* a." When these errors are accounted for, it can be seen that most other errors occur when the consonant is adjacent to a weak vowel. As more detailed examination shows that the release of an onset-wk consonant and the closure of a weak-coda consonant, i.e. the landmark closer to the affiliated weak vowel is more susceptible to being classified erroneously. This is also seen in the case of the strong-ambi-weak consonants. Discounting the flapped /t/'s, in which both the closure and the release are modified, the releases contained more than three times the errors than the closures. Since errors in the strong-ambi-weak environment accounts for a large portion of the total errors, this may be a reason for the larger occurrence of errors in releases in continuous speech than in the results for the isolated utterances discussed in the previous chapter.

The results for detecting consonant voicing for speaker ss are shown in Table 5.6. The results show a distribution of errors that is similar to that for speaker ks, but the error rates are slightly lower overall, around 22%. Results from combining the closure and release landmark measurements and considering the effects of perceived

training set	LAFF/1-5 (48)	LAFF/6-10 (54)	all (102)
LAFF/1-5	12 (25.0)	11 (20.4)	23 (22.5)
LAFF/6-10	12 (25.0)	10 (18.5)	22 (21.6)
LAFF/1-10	12 (25.0)	11 (20.4)	23 (22.5)
totals	36	32	68

Table 5.6: Classification results for LAFF sentences for speaker ss

training set	lm (102)	- err/mod	- err/mod	sm (65)	- err/mod	- err/mod
		(12)	+ cor/mod		(8)	+ cor/mod
LAFF/1-5	22.5	12.7	14.7	13.8	7.7	13.8
LAFF/6-10	21.6	11.8	13.7	15.4	9.2	15.4
LAFF/1-10	22.5	12.7	14.7	13.8	7.7	13.8

Table 5.7: Classification results for LAFF utterances for speaker ss using isolated utterances for training: effect of combining closure and release measurements and evaluation with perceived voicing

modifications are shown in Table 5.7. It is interesting to note that the perceived modifications occurred in identical environments as for speaker ks, i.e. flapped /t/'s preceded by a strong vowel and followed by a weak vowel, and voicing assimilation at voiced-unvoiced consonant boundaries. As for speaker ks, error rates decrease for all cases when landmarks are combined and perceived modifications are taken into account. Again, no further improvement is obtained by considering modification for the segmental results.

The error analyses for speaker ss are shown in Table 5.8 and Table 5.9, for landmark types and syllabic position, respectively. Again, the errors seem evenly spread between closures and releases, and between stops and fricatives, but are slightly more for unvoiced consonants than for voiced consonants. The majority of errors occur in the strong-ambisyllabic-weak environment as for speaker ks. Speaker ss has two instances of two landmarks (one closure and one release) that are marked as strong-ambisyllabic-weak which were marked as weak-ambisyllabic-strong in speaker ks. These occurred for the landmarks for the segment /s/ in the word "lasso." As for speaker ks, most errors occur adjacent to weak vowels, for all stress patterns and syllable positions.

training set	closure (47)	release (55)	stop (55)	fricative (47)	voiced (41)	unvoiced (61)
LAFF/1-5	15	11	14	12	8	18
LAFF/6-10	14	12	14	12	9	17
LAFF/1-10	14	12	14	12	8	18

Table 5.8: Error analysis for classification results for speaker ss

training set	onset-st (29)	onset-wk (22)	st-ambi-wk (24)	wk-ambi-st (6)	st-coda (17)	wk-coda (4)
LAFF/1-5	2	3	9	0	6	4
LAFF/6-10	4	3	9	0	6	4
LAFF/1-10	4	4	9	0	5	3
LAFF totals	10	10	27	0	17	11

Table 5.9: Analysis of classification errors into syllabic context for speaker ss

5.4 Summary

In this section, detection of consonant voicing was carried out for continuously spoken utterances. The procedure for detecting measures for consonant voicing was the same as for the isolated utterances. The performance was evaluated by comparing with both the underlying lexical voicing and with perceived voicing. Using the same training and test sentences predictably gave the best peformance, but using a selected set of isolated utterance for classifying the continuous speech also gave comparable results. The results also show that the detection scheme used in the experiments correctly classifies most of the modified voicing as different from the lexical voicing for each individual landmark. However, combining measurements from the closure and the release landmarks often resulted in recovering the underlying lexical voicing. This suggests that there may be a gradation in the modification of cues, so that voicing cues for consonants that are modified weakly may be recovered by combining all voicing measurements available. However, modifications such as flapping of underlying unvoiced /t/'s were consistently classified as voiced, and underlying lexical voicing was not recoverable.

The results were further analyzed according to landmark types and syllabic position for the consonants. Analysis of the errors showed a large number of errors being produced in consonants adjacent to weak vowels, for all stress patterns and positions within a syllable. The strong-ambisyllabic-weak environment, or onset-weak environment preceded by a strong syllable, commonly resulted in underlying unvoiced /t/'s to become flaps. It was also noted that for the closure and release of a consonant adjacent to a weak vowel, the landmark closer to the weak vowel was more susceptible to modification.

Chapter 6

Automatic detection of consonant voicing

In this chapter, the measurements developed up to this point are implemented automatically. The measurements thus obtained are used in various training and testing combinations,. to assess the performance of the procedure. In addition to the utterances tested in previous chapters, a larger set of continuous speech utterances is included.

6.1 Description of the utterances

Four sets of data are examined in this chapter. The first two include hand measurements for the isolated and continuous speech utterances, obtained previously. The next set is the automatic measurements for the isolated utterances for speaker ks and cb. The final set comprises automatic measurements from the first ten sentences of the LAFF database, and from an additional twenty sentences. This last set of measurements include 346 landmarks, corresponding to releases and closures from 228 underlying consonants.

6.2 Procedure

In order to obtain automatic measurements, each utterance must first be labeled with landmarks. These labels were located manually. Times for the closures and releases, as well as the voice onset/offsets were marked. At those times, the manner of production of the underlying consonant was also marked. These quantities were entered into a set of *.label* files.

The spectrogram and formant tracks for each utterance were generated using the *sgram* and *formant* utility programs in *xwaves* [11], which yield the *.sgram*, *.f0* and *.fb* files. The .sgram file contains the spectral amplitudes of the signal extracted at 10ms intervals. The .f0 file contains a track of the fundamental frequency, and the .fb file includes tracks of the formant frequencies up to the fifth formant, also at 10ms intervals.

The measurements for determining consonant voicing were extracted from these files by the following procedures. The fundamental frequency and the first and third formant frequency were extracted directly from the .f0 and .fb files, respectively, at each frame corresponding to times 10ms after that marked as a voice onset or 10ms just prior to a voice offset in the .label file. The amplitude of the first harmonic at this time (and also at times for closure or release) was determined as the spectral amplitude of the frequency closest to that of the fundamental frequency obtained above. Likewise, the amplitude of the second harmonic was determined as the amplitude of the frequency closest to twice that of the estimated fundamental frequency. The amplitudes of the first formant was also determined in a similar fashion.

These measurements were called the automatic measurements, in contrast with the hand measurements for the same utterances. Classification experiments using these measurements are described next.

6.3 Classification results

The distribution of automatic measurements for VCV and CVC utterances spoken by speaker ks are shown in Fig. 6.1 and Fig. 6.2, respectively. As before, the triangles show the means for utterances with the vowel /aa/, the squares for those with the vowel /eh/, and the circles denote means over both vowels. The short lines show the standard deviations, and the stars show the range of values. From the figures, it can be seen that the measures follow similar trends from those for the hand measurements. The H1 amplitudes at closure and release show good separability, as well as F0 and H1-H2 at the release. However, there is a greater variability in the measurements, as can be seen clearly for the F1 cutoff frequency for unvoiced fricatives in the VCV utterances. Also, it must be noted that the scales for the relative amplitude measurements are different. For example, the average mean of residual amplitude of H1 at the closure for voiced stops is about 40dB in Fig. 6.1, but the same mean is measured to be about 50dB in Fig. 4.5 in Chapter 4. Hand measurements were made using the *xkl* spectral analysis program, and the amplitude scales appear to be offset by about 10dB from *xwaves*. Therefore, measurements related to absolute amplitude are expected to be unreliable. Relative measures such as H1-H2 show more similar values. The measurements related to extracting the fundamental frequency and the first formant frequency show similar means, but more variability for the automatic measurements.

The continuous speech data are examined next. The distributions for measurements from the first 30 sentences of the LAFF database are shown in Fig. 6.3 and Fig. 6.4, for speaker ks and ss, respectively. Again, the overall positions of the distributions for voiced and unvoiced consonants are similar to those of the hand measurements, but there is a much greater variability, as was noted for the automatic measurements for the isolated utterances. The larger variability is most easily seen in measurements related to extracting the fundamental frequency. This is true for both speaker ks and speaker ss. The distributions for speaker ks also show that the amplitude scales have an offset of about 10dB. From these and previous figures, it

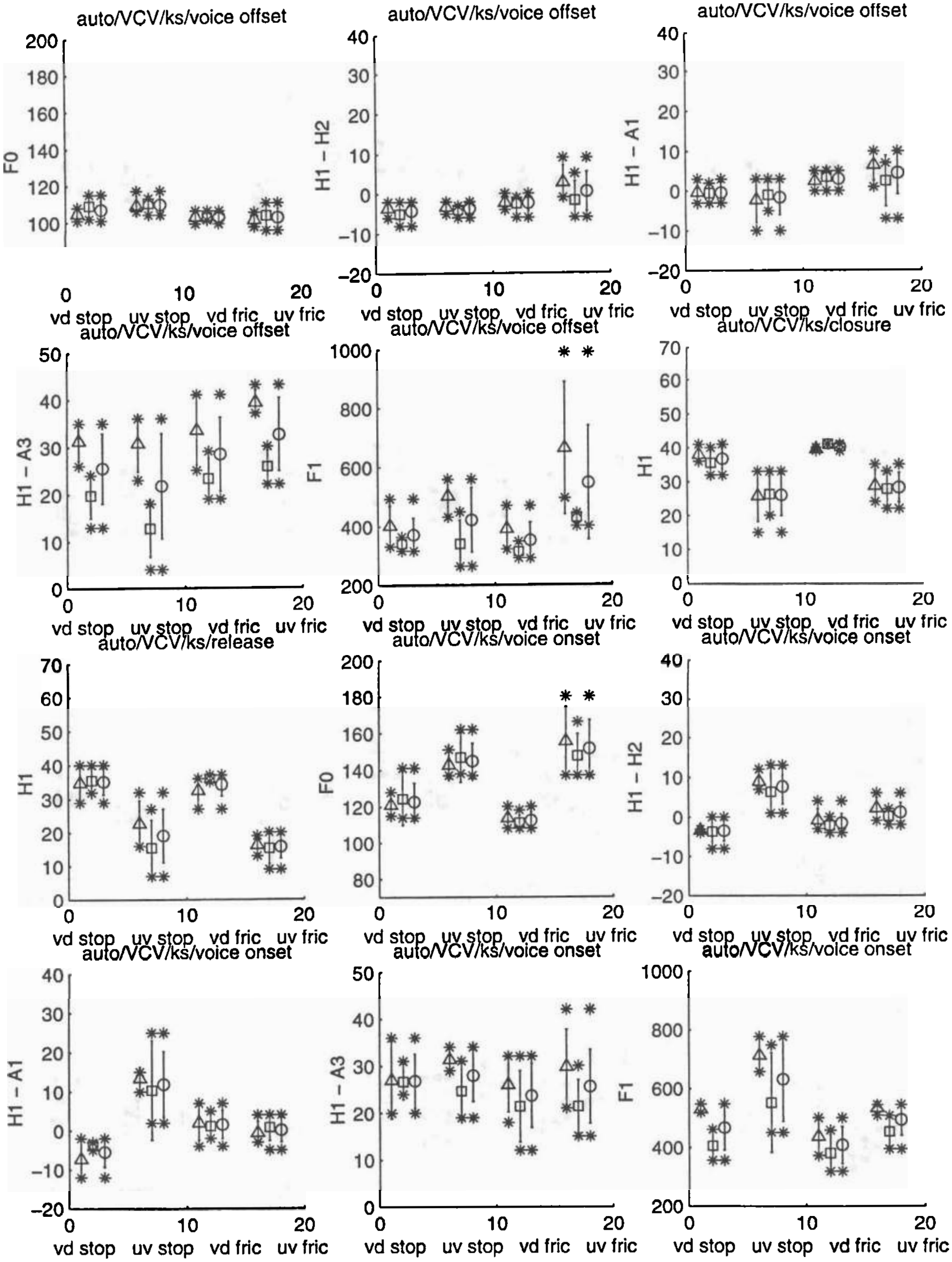


Figure 6.1: Distribution of automatic measures for consonant voicing in VCV utterances for speaker ks

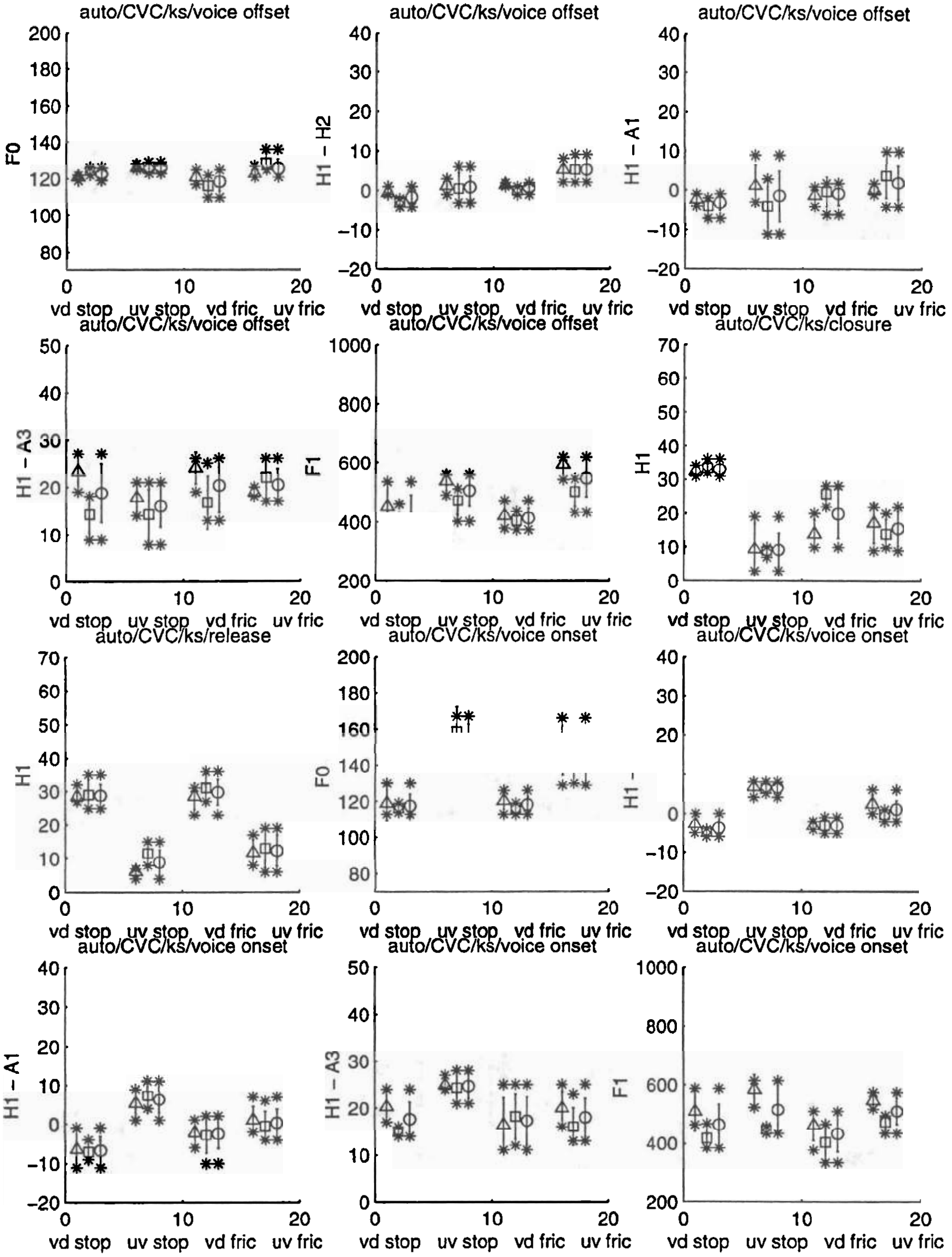


Figure 6.2: Distribution of automatic measures for consonant voicing in CVC utterances for speaker ks

may be concluded that determining voicing from automatic measures is expected to be worse than using hand measurements, particularly for continuous speech.

The results of training and testing with various combinations of data are examined next. Table 6.1 shows results for training on hand and automatic measurements of isolated utterances and testing on automatic isolated utterances. The number in parentheses on the top of the second column show the total number of landmarks tested. The number of errors are listed below, along with the percentage of error in parentheses. The results for training with the hand measurements show very large error rates, about 30%. Training with the automatic measurements improves error rates, but the performance is worse than for the hand measurements in Chapter 4. The error rates for training and testing on hand measurements of isolated utterances showed error rates of less than 10%. It is interesting to note that training with the CVC utterances still result in more errors, as for the hand measurements.

The offset in the scale of the spectral amplitude between the hand and automatic measurements appear to be the largest factor in the mismatch between the two sets of data. This is because the amplitude of H1 during the closure interval is used as a key measurement in deciding voicing. Previously, this measurement has been the most separable. Moreover, this measurement is assigned a weight of two, so that it has a greater effect on the final voicing decision than the other measurements. However, when the automatic measurements are used for training, there is no mismatch between the amplitude scales, and as a result, performance is closer to that obtained for training and testing with hand measurements.

The results for combining the measures for the closure and release landmarks for the VCV utterances are shown in Table 6.2. The number of errors and the error rates (in parentheses) are listed for training with the hand measurements on the left, and training with automatic measurements on the right. The total number of landmarks is 64, which corresponds to 32 segments, as shown across the top row. When the hand measurements for VCV utterances were used as the training sets, the performance decreased when the segments were consolidated. As discussed above, this indicates that the voicing decisions for either or both of the landmarks were strongly indicative

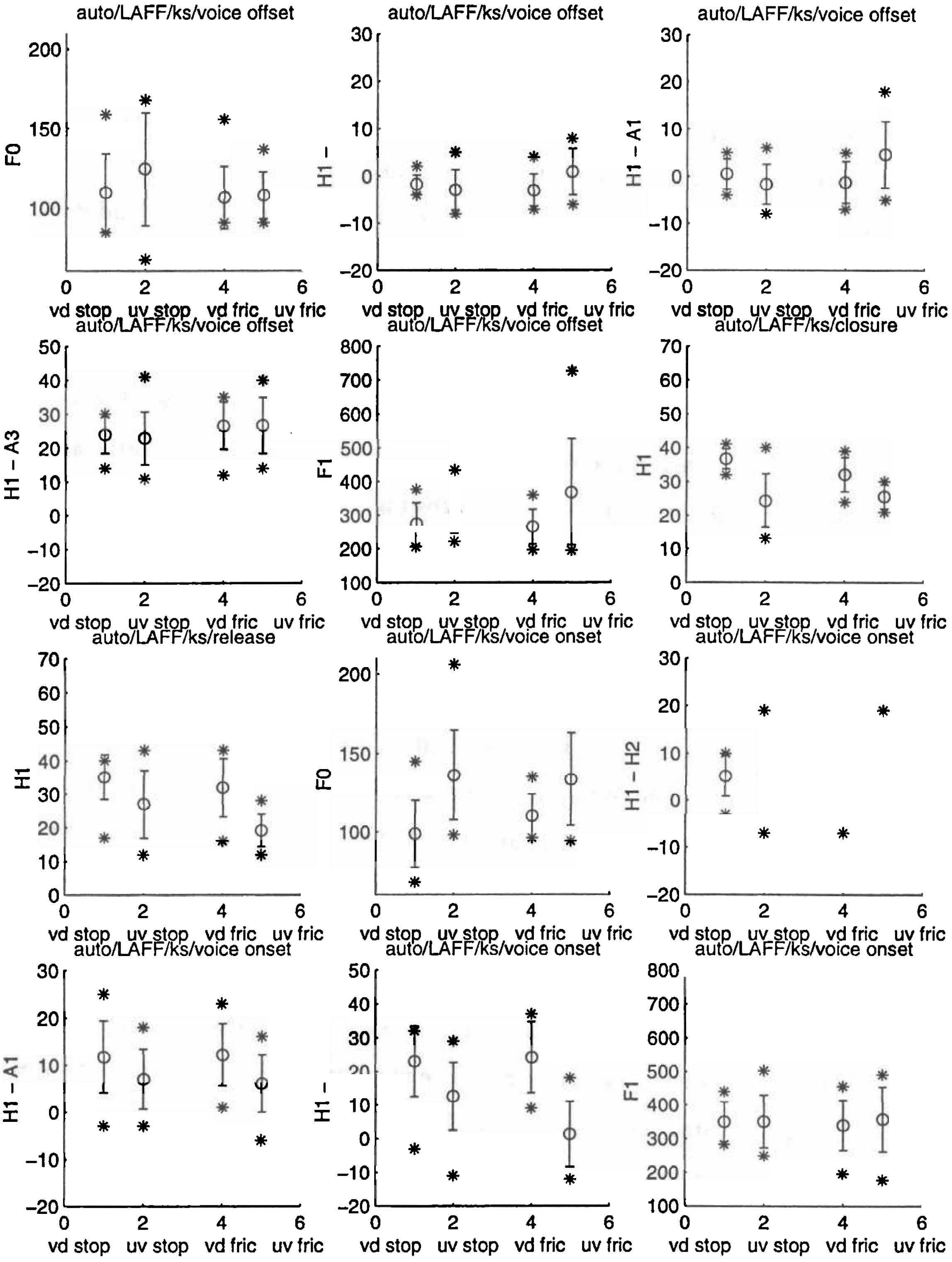


Figure 6.3: Distribution of automatic measures for consonant voicing for LAFF sentences spoken by speaker ks

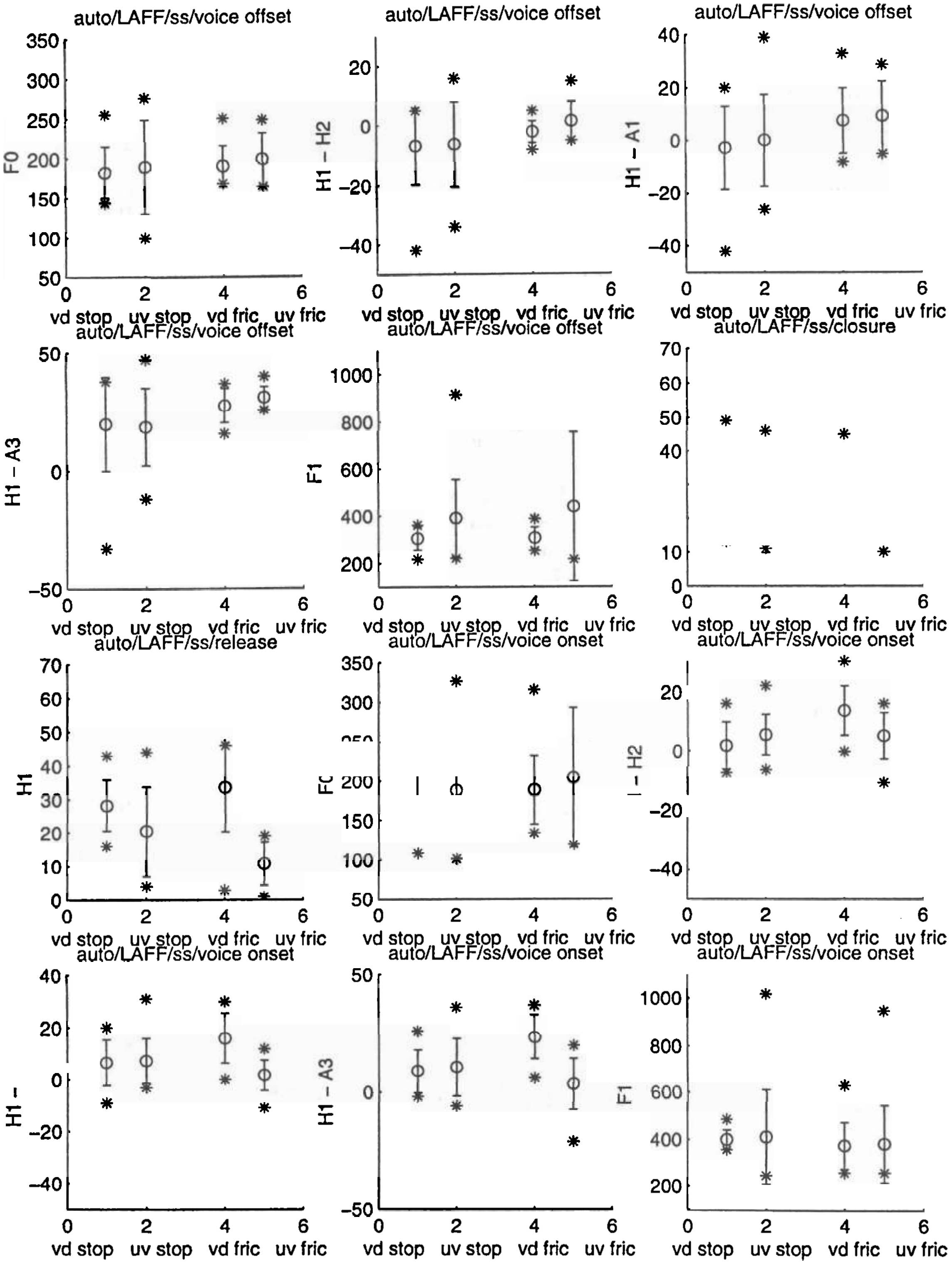


Figure 6.4: Distribution of automatic measures for consonant voicing for LAFF sentences spoken by speaker ss

training set	auto/isol (128)	training set	auto/isol (128)
VCV/ks/aa	45 (35.2)	auto/VCV/ks/aa	14 (11.0)
VCV/ks/eh	51 (39.8)	auto/VCV/ks/eh	14 (11.0)
VCV/ks/aa+eh	47 (36.7)	auto/VCV/ks/aa+eh	16 (12.5)
CVC/ks/aa	41 (32.0)	auto/CVC/ks/aa	19 (14.8)
CVC/ks/eh	35 (27.3)	auto/CVC/ks/eh	21 (16.4)
CVC/ks/aa+eh	35 (27.3)	auto/CVC/ks/aa+eh	21 (16.4)
VCV+CVC/ks/aa	42 (32.8)	auto/VCV+CVC/ks/aa	16 (12.5)
VCV+CVC/ks/eh	41 (32.0)	auto/VCV+CVC/ks/eh	17 (13.2)
VCV+CVC/ks/aa+eh	39 (30.5)	auto/VCV+CVC/ks/aa+eh	17 (13.2)

Table 6.1: Classification results for training and testing on automatic measurements of the isolated utterances for speaker ks

training set	lm (64)	sm (32)	training set	lm (64)	sm (32)
VCV/ks/aa	24 (37.5)	16 (50.0)	auto/VCV/ks/aa	7 (10.9)	2 (6.3)
VCV/ks/eh	25 (39.1)	16 (50.0)	auto/VCV/ks/eh	5 (7.8)	2 (6.3)
VCV/ks/aa+eh	23 (35.9)	16 (50.0)	auto/VCV/ks/aa+eh	7 (10.9)	2 (6.3)
CVC/ks/aa	16 (25.0)	6 (18.8)	auto/CVC/ks/aa	14 (21.9)	6 (18.8)
CVC/ks/eh	17 (26.6)	1 (3.1)	auto/CVC/ks/eh	17 (26.6)	9 (28.1)
CVC/ks/aa+eh	17 (26.6)	2 (6.3)	auto/CVC/ks/aa+eh	18 (28.1)	9 (28.1)
VCV+CVC/ks/aa	22 (34.4)	8 (25.0)	auto/VCV+CVC/ks/aa	14 (21.9)	8 (25.0)
VCV+CVC/ks/eh	20 (31.3)	14 (43.8)	auto/VCV+CVC/ks/eh	12 (18.8)	6 (18.8)
VCV+CVC/ks/aa+eh	20 (31.3)	10 (31.3)	auto/VCV+CVC/ks/aa+eh	14 (21.9)	7 (21.9)

Table 6.2: Results for consolidating automatic measurements at the closures and releases of VCV utterances for speaker ks

of voicing that was different from the lexical value. In contrast, if the CVC utterances were used, consolidation led to a sharp decrease in errors. This shows that if one of the landmarks were marginally incorrect, the voicing decision for the other was strong enough to overcome the error. However, this does not necessarily indicate that the CVC hand measurements are a good set of training utterances for recognizing the automatically measured utterances. It is probably the case that the lower residual H1 amplitude at the closures (due to partial devoicing from the following /t/) resulted in a better correspondence with the automatic measurements.

Table 6.3 shows the results from training on the hand or automatic measurements of the isolated utterances, and testing on the automatic measurements of the first 30 sentences of the LAFF database. The numbers in parentheses at the top of columns 2 and 4 are the total number of landmarks examined. The number of errors are shown, along with the error rates in parentheses. Overall, the performance is again

training set	auto/LAFF/1-30 (346)	training set	auto/LAFF/1-30 (346)
VCV/ks/aa	144 (41.6)	auto/VCV/ks/aa	86 (24.3)
VCV/ks/eh	141 (40.8)	auto/VCV/ks/eh	81 (23.4)
VCV/ks/aa+eh	149 (43.0)	auto/VCV/ks/aa+eh	80 (23.2)
CVC/ks/aa	107 (30.9)	auto/CVC/ks/aa	132 (38.2)
CVC/ks/eh	120 (34.6)	auto/CVC/ks/eh	129 (37.2)
CVC/ks/aa+eh	106 (30.6)	auto/CVC/ks/aa+eh	131 (37.8)
VCV+CVC/ks/aa	137 (39.6)	auto/VCV+CVC/ks/aa	101 (29.2)
VCV+CVC/ks/eh	137 (39.6)	auto/CVC+CVC/ks/eh	97 (28.0)
VCV+CVC/ks/aa+eh	140 (40.4)	auto/CVC+CVC/ks/aa+eh	99 (28.6)

Table 6.3: Classification results for training on hand measurements and automatic measurements of isolated utterances and testing on continuous speech for speaker ks

much worse than for the hand measurements. There is again a large improvement when the automatic measurements are used in training, over that of using the hand measurements. The error rates for using the CVC hand measurements show better results than using the VCV measurements or both. This is in contrast with the results for the automatic measurements, where the CVC error rates are higher. This is because the hand measurements for the CVC utterances have lower values of means for the residual H1 amplitude at the closure and release than the VCV utterances (see Fig. 4.6), and is consequentially closer to the means for the automatic utterances, which are also about 10dB lower. However, when the automatic measurements are used, the VCV utterances are shown to provide a better training set.

The results in Table 6.4 are obtained from training with hand and automatic measurements from the continuous sentences, and testing on the first 30 LAFF sentences. The results for evaluating the voicing decisions for landmarks are given on the left, and results for measurements combined into segments are given on the right. The number of errors and error rates (in parentheses) evaluated by direct comparison with underlying lexical voicing are given in columns 2 and 5. Error rates obtained after discounting modified landmarks and segments are given in columns 3 and 6, respectively. The test sentences contained 33 modified landmarks corresponding to 23 consonants. The remaining two columns are the error rates when failure to recognize modified consonants was counted as errors. It is again clear that training on the automatic measurements yields better recognition rates than the hand measure-

training set	lm	-err/mod	- err/mod	sm	-err/mod	- err/mod
	(346)	(33)	+ cor/mod	(222)	(23)	+ cor/mod
LAFF/ks/1-5	160 (46.2)	42.8	48.8	87 (39.2)	38.7	48.6
LAFF/ks/6-10	145 (41.9)	39.6	46.8	81 (36.5)	36.5	46.8
LAFF/ks/1-10	145 (41.9)	39.6	46.8	83 (37.4)	37.4	47.7
auto/LAFF/ks/1-5	76 (22.0)	15.0	17.6	36 (16.2)	12.6	19.4
auto/LAFF/ks/6-10	74 (21.4)	14.7	17.6	36 (16.2)	12.6	19.4
auto/LAFF/ks/1-10	69 (20.0)	13.6	16.8	35 (15.8)	12.2	18.9
auto/LAFF/ks/11-30	70 (20.2)	14.2	17.6	34 (15.3)	11.7	18.5
auto/LAFF/ks/1-30	71 (20.6)	11.7	17.3	35 (15.8)	12.2	18.9

Table 6.4: Classification results for training on hand and automatic measurements and testing on automatic measurements of the first 30 sentences of the LAFF database for speaker ks. Landmark errors are on the left, and consolidated segment errors are on the right.

ments. Even for training with the automatic measurements, the error rates are worse compared to training and testing on hand measurements, as discussed in the previous chapter. It should be noted that using a larger portion of the test set as the training set leads to better results, as expected. However, the difference is on the order of a few percent. Combining measures for closures and releases led to improved performance in all cases. Evaluation by comparison with perceived voicing yielded better landmark error rates, but decreased performance when closure and release measurements were combined. As discussed in the previous chapter, this suggests that marginal modifications in voicing may be recoverable when segments are consolidated.

The classification results for training on hand and automatic measurements of continuous speech and testing on the first 30 sentences of the LAFF database are given in Table 6.5 for speaker ss. The error rates are slightly higher than speaker ks, which is different from the results for the hand measurements. Examination of the distributions of automatic measurements shows a much wider variation for speaker ss than for speaker ks, especially in measurements for fundamental and first formant frequency, which may be a reason for the lower recognition rate. In general, consolidated segment errors are less than the landmark errors, and comparison with perceived voicing lowers error rates for landmarks, but not for segments.

The errors that resulted from testing the automatic measurements for continuous speech with the various training sets are examined next in further detail. Ta-

training set	lm (346)	-err/mod	- err/mod	sm	-err/mod	- err/mod
	(346)	(33)	+ cor/mod	(222)	(23)	+ cor/mod
LAFF/ss/1-5	146 (42.2)	36.1	39.6	81 (36.5)	33.3	40.5
LAFF/ss/6-10	145 (41.9)	34.7	37.0	79 (35.6)	32.9	40.5
LAFF/ss/1-10	143 (41.3)	35.3	38.7	79 (35.6)	32.4	39.6
auto/LAFF/ss/1-5	123 (35.5)	30.9	35.8	57 (25.7)	22.1	28.8
auto/LAFF/ss/6-10	90 (26.0)	19.1	21.7	42 (18.9)	14.9	21.2
auto/LAFF/ss/1-10	89 (25.7)	18.8	21.4	40 (18.0)	14.0	20.3
auto/LAFF/ss/11-30	85 (24.6)	17.3	19.7	37 (16.7)	12.6	18.9
auto/LAFF/ss/1-30	81 (23.4)	16.2	18.5	32 (14.4)	10.4	16.7

Table 6.5: Classification results for training on hand and automatic measurements and testing on automatic measurements of the first 30 sentences of the LAFF database for speaker ss. Landmark errors are on the left, and consolidated segment errors are on the right.

ble 6.6 shows the results from training with hand measurements of the isolated utterances and testing with the automatic measurements of the first 30 sentences of the LAFF database. The errors are divided into closure/releases, stops/fricatives, and voiced/unvoiced groupings. The number in parentheses at the top of each column is the total number of landmarks corresponding to that category. There are slightly more releases than closures, and approximately the same number of landmarks for voiced and unvoiced consonants, but a much larger number of stop landmarks than fricative landmarks. From the results, it can be seen that there is an even distribution of errors between closures and releases, and the proportion of errors in stops is just slightly lower than in fricatives. However, there are more than twice the number of errors for voiced consonants than for unvoiced consonants. This is in contrast to the results from the previous chapter, where most errors occurred for the unvoiced cases. This effect is due to the difference in amplitude scales. The amplitude of H1 during the closure interval for voiced stops have lower values for the automatic measurements, and are thus classified as unvoiced, when compared with distributions from the hand measurements.

Table 6.7 shows the results when the automatic measurements for the isolated utterances were used as the training set. In this case, there is a proportionately larger number of errors in stops than in fricatives, unlike the results from Table 6.6. Also, the majority of errors are now in the unvoiced consonants. This trend is most

training set	cl(163)	rl(183)	stop(206)	fric(140)	vd(175)	uv(171)
VCV/ks/aa	63	81	77	67	112	32
VCV/ks/eh	62	79	75	66	118	23
VCV/ks/aa+eh	66	83	79	70	115	34
CVC/ks/aa	54	53	66	41	55	52
CVC/ks/eh	46	74	62	58	71	49
CVC/ks/aa+eh	45	61	61	45	57	49
VCV+CVC/ks/aa	62	75	77	60	95	42
VCV+CVC/ks/eh	64	73	71	66	104	33
VCV+CVC/ks/aa+eh	65	75	77	63	100	40
totals	527	654	645	536	827	354

Table 6.6: Error analysis for training on hand measurements of isolated utterances and testing on automatic measurements of continuous speech for speaker ks

easily seen for the cases where CVC utterances were used as the training utterances. In these results, the VCV utterances are seen to be better training sets, as in the previous chapter. The large number of errors in unvoiced consonants is due to actual modifications of the consonants. The modifications include 20 landmarks corresponding to flapped /t/'s, and 3 other landmarks for devoiced consonants. On the other hand, only 7 voiced landmarks were perceived as modified. Another reason is due to prosodic boundary effects. The fundamental frequency of phonated segments were found to decrease at the end of sentences. Therefore, the F0 measurements at phonation regions adjacent to consonants were measured to be lower than at an earlier point in the sentence. This results in F0 measurements at voice offsets and onsets for unvoiced stops approaching the mean values for voiced stops, leading to erroneous classification.

Similar analyses for training on manual and automatic measurements of the LAFF database and testing on the automatic LAFF measurements are shown in Table 6.8 and Table 6.9 for speakers ks and ss, respectively. Again, for both speakers, errors are distributed more or less evenly between closures and releases, and between stops and fricatives, but occur more in voiced consonants when trained with hand measurements and in unvoiced consonants when trained with the automatic measurements. This is again due to a mismatch between the hand measurements and the automatic measurements, as discussed previously. Training with the continuous utterances results

training set	cl(163)	rl(183)	stop(206)	fric(140)	vd(175)	uv(171)
auto/VCV/ks/aa	32	52	49	35	27	57
auto/VCV/ks/eh	27	54	48	33	28	53
auto/VCV/ks/aa+eh	27	53	50	30	28	52
auto/CVC/ks/aa	66	66	94	38	13	119
auto/CVC/ks/eh	66	63	89	40	13	116
auto/CVC/ks/aa+eh	67	64	92	39	9	122
auto/VCV+CVC/ks/aa	41	60	62	39	14	87
auto/VCV+CVC/ks/eh	39	58	62	35	13	84
auto/VCV+CVC/ks/aa+eh	40	59	63	36	14	85
totals	405	529	609	325	159	741

Table 6.7: Error analysis for training on automatic measurements of isolated utterances and testing on automatic measurements of continuous speech for speaker ks

training set	cl(163)	rl(183)	stop(206)	fric(140)	vd(175)	uv(171)
LAFF/ks/1-5	67	93	90	70	136	24
LAFF/ks/6-10	64	81	80	65	128	17
LAFF/ks/1-10	65	80	77	68	128	17
auto/LAFF/ks/1-5	26	50	44	32	22	54
auto/LAFF/ks/6-10	29	45	42	32	28	46
auto/LAFF/ks/1-10	22	47	40	29	22	47
auto/LAFF/ks/11-30	25	45	40	30	23	47
auto/LAFF/ks/1-30	25	46	40	31	24	47

Table 6.8: Error analysis for training on manual and automatic measurements of continuous speech and testing on automatic measurements of continuous speech for speaker ks

in decreases in overall error rates over that of training with the isolated utterances. Nevertheless, comparison of the results in Table 6.7 and Table 6.8 show that using the VCV utterances as training sets results in performance that is just slightly worse than using the LAFF sentences for speaker ks. This is in accordance with the results of the previous chapter.

The results for speaker ss show error rates that are somewhat higher than for speaker ks. There are also more errors in voiced consonants relative to unvoiced consonants. Examination of the distributions of measurements show a larger increase in range of F0 values for speaker ss. The range for unvoiced consonants is larger, and the overall mean is lower. Training on this distribution results in a higher number of voiced consonants being classified as closer to the unvoiced mean.

training set	cl(163)	rl(183)	stop(206)	fric(140)	vd(175)	uv(171)
LAFF/ss/1-5	71	75	82	64	104	42
LAFF/ss/6-10	74	71	79	66	112	33
LAFF/ss/1-10	71	72	80	63	100	43
auto/LAFF/ss/1-5	36	87	83	40	51	72
auto/LAFF/ss/6-10	39	51	50	40	41	49
auto/LAFF/ss/1-10	39	50	49	40	39	50
auto/LAFF/ss/11-30	33	52	49	36	39	46
auto/LAFF/ss/1-30	30	51	48	33	37	44

Table 6.9: Error analysis for training on manual and automatic measurements of continuous speech and testing on automatic measurements of continuous speech for speaker ss

Next, the errors have been analyzed according to the position of the consonant within the syllable. Table 6.10 shows results from training with hand measurements of isolated utterances and testing on the first 30 sentences of the LAFF database. The numbers in parentheses across the first row are the total numbers of landmarks corresponding to each category. The first two marked os and ow denote onsets followed by strong and weak vowels, respectively. The category marked sw denoted ambisyllabic consonants preceded by a strong vowel and followed by a weak vowel, and vice versa for the case of ws. The last two entries are consonants in the coda, preceded respectively by a strong or weak vowel. There are more errors in the consonants in onsets followed by weak vowels than by strong vowels, and also for consonants in codas preceded by weak vowels than by strong vowels. Also, proportionately more errors lie in ambisyllabic consonants in the weak-strong environment than in the strong-weak environment; these characteristics are in contrast with the results in the previous chapter.

Table 6.11 shows the analysis of results from training with the automatic measurements for the isolated utterances and testing with the automatic LAFF measurements. The results are similar, except more errors now occur in the strong-coda environment than the weak-coda environment. Also, there is a large decrease in errors for onset consonants followed by weak vowels, so that the number of errors is more comparable to that of the onset-strong environment.

The increase in the number of consonants in the onset-strong environment is

training set	os(90)	ow(78)	sw(62)	ws(32)	sc(44)	wc(40)
VCV/ks/aa	30	36	32	15	10	21
VCV/ks/eh	25	38	32	12	12	22
VCV/ks/aa+eh	30	36	33	15	13	22
CVC/ks/aa	25	29	20	13	10	10
CVC/ks/eh	25	34	25	11	11	14
CVC/ks/aa+eh	23	28	23	12	11	9
VCV+CVC/ks/aa	30	37	27	13	13	17
VCV+CVC/ks/eh	27	38	28	11	12	21
VCV+CVC/ks/aa+eh	29	37	27	13	13	20
totals	244	313	248	115	105	156

Table 6.10: Analysis of errors according to position within a syllable for results from training on hand measurements of isolated speech and testing on automatic measurements of continuous speech for speaker ks

training set	os(90)	ow(78)	sw(62)	ws(32)	sc(44)	wc(40)
auto/VCV/ks/aa	17	24	14	8	12	9
auto/VCV/ks/eh	17	20	15	5	13	11
auto/VCV/ks/aa+eh	15	24	14	6	12	9
auto/CVC/ks/aa	36	32	17	10	25	12
auto/CVC/ks/eh	34	34	18	11	23	9
auto/CVC/ks/aa+eh	37	30	18	11	25	10
auto/VCV+CVC/ks/aa	26	29	16	8	14	8
auto/VCV+CVC/ks/eh	24	28	15	8	14	8
auto/VCV+CVC/ks/aa+eh	25	29	15	8	14	8
totals	231	250	142	75	152	84

Table 6.11: Analysis of errors according to position within a syllable for results from training on automatic measurements of isolated speech and testing on automatic measurements of continuous speech for speaker ks

training set	os(90)	ow(78)	sw(62)	ws(32)	sc(44)	wc(40)
LAFF/ks/1-5	40	41	30	15	10	24
LAFF/ks/6-10	34	39	31	14	9	18
LAFF/ks/1-10	33	37	32	14	9	20
auto/LAFF/ks/1-5	16	19	15	3	14	9
auto/LAFF/ks/6-10	18	21	13	5	9	8
auto/LAFF/ks/1-10	15	21	13	3	9	8
auto/LAFF/ks/11-30	18	20	12	4	9	7
auto/LAFF/ks/1-30	18	20	13	4	8	8

Table 6.12: Analysis of errors according to position within a syllable for results from training on manual and automatic measurements of continuous speech and testing on automatic measurements of continuous speech for speaker ks

in large measure due to prosodic boundary effects, where the drop in fundamental frequency at the end of a sentence leads to unvoiced consonants being classified as voiced. At the same time, the /dh/ at the onset of the word "they" at the start of a sentence was often classified as an unvoiced consonant. This segment is produced often produced similar to a stop, and with very little prevoicing before the release. This modification results in a lower value of H1 amplitude before the release, and leads to a wrong classification.

Next, the results from training on the manual and automatic measurements of the LAFF sentences and testing on the automatic LAFF measurements are shown in Table 6.12 and Table 6.13 for speaker ks and ss, respectively. Overall, the distributions of errors are similar to the results above, where the errors between onset consonants followed by strong and weak vowels are comparable, with somewhat more occurring in the onset-weak environment. The number of errors decreases more for coda consonants preceded by weak vowels than for those preceded by strong vowels when the training set is changed from the hand measurements to the automatic measurements. This is also the case for the ambisyllabic consonants in the weak-strong versus the strong-weak cases.

Again, the greater number of errors in the strong-coda and onset-strong environment is mostly due to prosodic effects at the end of a sentence, where the fundamental frequency decreases. When these errors are accounted for, the results show that landmarks adjacent to weak vowels are more susceptible to modification, as in the previous

training set	os(90)	ow(78)	sw(62)	ws(32)	sc(44)	wc(40)
LAFF/ss/1-5	34	30	29	16	14	23
LAFF/ss/6-10	27	29	34	12	15	28
LAFF/ss/1-10	33	28	29	16	14	23
auto/LAFF/ss/1-5	16	29	18	13	9	15
auto/LAFF/ss/6-10	18	25	17	6	8	15
auto/LAFF/ss/1-10	15	24	17	6	9	15
auto/LAFF/ss/11-30	18	19	19	6	10	16
auto/LAFF/ss/1-30	18	19	14	4	11	16

Table 6.13: Analysis of errors according to position within a syllable for results from training on manual and automatic measurements of continuous speech and testing on automatic measurements of continuous speech for speaker ss

chapter. Among the modifications, six flaps occurred in a strong-ambisyllabic-weak environment, or in strong-coda environments followed by a weak vowel. The remaining two flaps, one in an onset and one in a coda, occurred next to weak vowels. Other modifications included five cases in the weak-coda environment, and four cases in the onset-weak environment. One modification was noted for the lexically strong-coda environment for “i*t* *b*egan,” but the first vowel in this sequence seems to undergo reduction into a weak vowel.

Overall, the results show that training with the hand measurements provides an ill match for testing the automatic measurements. Among the hand measurements, using the CVC utterances results in better performance than including the VCV utterances. When the automatic measurements are used, the performance is improved, and in this case, the VCV utterances are better training sets. Using the isolated utterances as training data results in error rates that are lower than those obtained using hand measurements of continuous speech. However, when the automatic measurements of the continuous speech are used, the lowest error rates are obtained. These rates are comparable to those obtained by training and testing on the hand measurements of the continuous speech, as described in the previous chapter.

6.4 Summary

In this chapter, the measures for determining consonant voicing were extracted using an automatic algorithm. The measurements were found at times in the signal corresponding to consonant landmarks. These times, along with the manner of the underlying consonant, were determined manually. Results show that large error rates are obtained when hand measurements are used in training. This was largely due to a difference in the scale of spectral amplitude in the automatic measurements, resulting in unreliable measurements for amplitude of H1 during the closure interval. Relative measurements such as H1-H2 were not affected as greatly.

The distributions for the measurements, especially fundamental and first formant frequency, also showed a much greater variability than the distributions for the hand measurements. As a result, training and testing on the automatic measurements also yielded performance that was slightly lower than training and testing on the hand measurements. Using continuous speech in training resulted in better performance for tests with the LAFF sentences. However, selecting a suitable set of isolated utterances for training gave error rates that were only slightly higher.

Consolidating closure and release landmarks led to improved recognition rates in all cases, to about 16% for speaker ks and 18% for speaker ss. Comparison with perceived modifications showed that in many cases, combining measures for both landmarks resulted in recovering underlying lexical voicing.

Analysis of errors according to landmark type showed greater errors for unvoiced consonants. A large portion of these errors were due to flapping of underlying /t/ segments. Of the remaining perceived modifications, most were classified as voicing assimilation of underlying unvoiced consonants. In addition, most errors occurred in landmarks adjacent to weak vowels. Of those that occurred adjacent to strong vowels, most were identified as due to prosodic boundary effects, or within a strong-consonant-weak environment that encouraged flapping.

The results show that using hand and automatic measurements in training and testing requires a good match between the amplitude scales adopted by the measurement procedures. However, since it is also possible that recording conditions may differ for utterances that are to be used in training and testing, relative measures, such as decrease in amplitude of H1 in the closure interval from the adjacent vowel, may be a better solution. When the amplitude scales are well matched for the training and test sets, the results are similar to those for the hand measurements discussed in the previous section. However, it can be seen that the ranges of distributions of the automatic measurements are larger. This is most easily observed for estimates of fundamental and first formant frequency. Since all measures used for voicing decisions in this thesis are related to estimation of fundamental frequency, it is expected that independent measures, such as duration information, may be needed to improve reliability.

Chapter 7

Summary and discussions

In this thesis, an overview of a hierarchical speech recognition system based on knowledge about representation of speech has been described. As an example of implementation of a component in this system, a module for detection of consonant voicing has been designed, and results of classification experiments have been analyzed. A summary of the work described in this thesis and directions for further study will be discussed in the following sections.

7.1 Design of the hierarchical speech recognition system

The recognition system proposed in this thesis comprises several levels of representation of speech information extracted from the signal. The information is found by a number of component modules that examine the acoustic signal for acoustic cues to infer the values of the underlying features. The signal is first examined to determine the landmarks, and further processing around the landmarks leads to values for the underlying features. These landmarks and their features are then consolidated into segments and their features. The sequence of segments thus obtained is then compared with items in the working lexicon (derived from the canonical lexicon through phonological rules), to yield possible word matches. Higher-level information needed to guide individual modules is expressed in terms of features, or other linguistic units.

For the consonant voicing module, necessary information includes consonant landmarks and landmark types. Consonant landmark types, such as fricative or stop, can be expressed in terms of articulator-free features. For example, a fricative landmark is described as [+consonantal, -sonorant, +continuant], and [+strident] or [-strident], depending on place. In this thesis, determination of stridency is not needed in finding consonant voicing. In other words, place information was not used. From experimental results, it may be suggested that specifying higher level context, such as stress patterns and syllable position may be useful in refining the consonant voicing procedure to include context-specific processing. These types of information are effectively represented within the hierarchical recognition system.

The acoustic cues used to infer the values of underlying features were initially selected by studying the production mechanisms involved in producing a segment described by the features. The measurements are then refined by examination of spectral representations corresponding to those utterances. For consonant voicing, this procedure resulted in selection of residual first harmonic amplitude during the closure interval, and fundamental frequency, relative amplitude of the first two harmonics and first formant cutoff frequency at phonated intervals immediately adjacent to the consonant landmarks. Measurements that contribute to voicing decision also depended on landmark type (e.g. H1-H2 at voice onset for stops, but not fricatives).

The acoustic cues thus found may then be further interpreted into classification decisions for categories that are expressed by features corresponding to those acoustic cues. This process is language dependent. In English, the voice/unvoiced distinction in consonants is expressed primarily by the features [stiff, slack], with the features [spread, constr] providing secondary information. The measurements show that the distributions of acoustic cues for these features are indicative of the underlying voiced/unvoiced classes. Thus, measurements of the acoustic cues corresponding to the four laryngeal features were used to determine consonant voicing.

Consonant landmarks and their associated features are then consolidated in the matcher. In this thesis, a closure-release pair with the same manner and place were used as the criterion in combining landmarks. A simple method of equal weighting of the two landmarks was used to find the resulting feature values. For consonant voicing, the results showed that this procedure yielded better estimates of the underlying voiced/unvoiced category than estimates for individual landmarks, which may suggest that acoustic cues for features may be unequally realized at the closure and release landmarks. In fact, the results suggest that this is affected by higher-level linguistic units, such as stress patterns and syllable position.

The segments and their feature values may then be used in the matcher to find lexical items consistent with the segment sequence found. This and higher level processing was not discussed in this thesis.

7.2 Discussions and further work in implementation of the consonant voicing module

Chapter 3 presented a discussion of the production of consonant voicing in cases where adjacent sounds were assumed to be produced with a phonation source. However, it is also possible that adjacent sounds may have other sources, i.e. frication or aspiration, as in the sequence "lifts heavy." Additionally, although nasal segments have a phonation source, the output signal is modified by the vocal tract very differently from a vowel or a glide where the nasal tract is closed off. In these cases, the relative amplitudes of the harmonic components cannot be compared with that in the vowels and glides. The effect of non-phonation sources and phonation sources with coupling of the nasal tract in the environment of voiced and unvoiced consonants needs to be further examined in order to determine suitable acoustic cues for consonant voicing in these cases.

The measurements that are used in this thesis do not consider any duration information. Traditionally, voice onset time (the time between the release and the onset of voicing) has been used as a good criterion for distinguishing between voiced and unvoiced stop consonants. Instead of this measure, the measure H1-H2 has been used to characterize the spread of the glottis, which is viewed as the mechanism affecting the degree and duration of aspiration present between the release and the onset of voicing. However, under adverse conditions, such as presence of noise that masks the low frequency harmonic structure, duration information may become a primary source of information in determining consonant voicing. Also, duration information is less dependent on accurate location of the landmarks. Other durational measures such as length of the closure interval for the consonant and/or duration of adjacent segments may also be used. For example, voiced consonants tend to be shorter in duration in general than unvoiced consonants. In order to determine the underlying mechanisms that result in differences in adjacent vowel durations, the relation between the timing of the closure and release of the primary articulators and that of the offset and onset of phonation for voiced and unvoiced consonants needs to be studied further.

The measures for detecting consonant voicing are related to the laryngeal configuration at the times of voice offset and onset and during the closure interval. These measures are absolute measures, in that they are not compared with any other measure. However, the results suggest that using relative measures may lead to improved performance. For example, instead of using the amplitude of H1 singly, a measure such as the difference between the amplitude of H1 at the preceding vowel and the amplitude of H1 at the offset of voicing of that vowel may be used. Using such relative measures may provide a more reliable measure than using absolute measures, since local perturbations such as change in overall loudness or prosodic effects may have less effect on the measurement. However, relative measures require knowledge of information from other (adjacent) segments. As a result, if the measures or procedures for extracting acoustic cues from other segments (i.e., finding the landmark for the adjacent vowel, in this example) are not reliable, the relative measure will be unreliable as well. In addition to measures that are found relative to different times in the signal, rates of change of the measurements at a point in time may also be included.

The utterances examined in this thesis were spoken by 3 different speakers (1 male and 2 female). In order to ascertain that observations made in this study are relevant in general, data from more speakers need to examined in the future. Utterances from the three speakers were used in speaker-dependent experiments in this thesis, i.e. utterances in the training and test sets were from the same speaker. Using relative measures, as discussed above, may make it possible to implement a speaker-independent procedure, since average measures that are related to physical dimensions of the vocal tract of the speaker may be discounted.

In this thesis, measurements were made at each point in the signal corresponding to the closure and release of a consonant, and the offset and onset of voicing of adjacent vowels. At each time, decisions were made as to whether individual measurements, such as amplitude of H1 during the closure interval, was characteristic of an underlying voiced or unvoiced consonant. These decisions were then summed, with a weight of two placed on the amplitude of H1 during the closure interval, to obtain a voicing decision for that landmark. The weights placed on the individual measurements were motivated by the description of the voiced and unvoiced classes of consonants in English by the four laryngeal features ([stiff, slack, spread, constr]), as well as from examination of the data.

The measures of onset F0, H1-H2, and cutoff F1 each present information that is used to infer the value of one feature, namely [stiff], [spread], and [constr], respectively. As a result, each measure was given a weight of one. Within the closure interval, a large amplitude of H1 signals a voiced consonant, with slack vocal folds which are neither spread nor constricted ([+slack, -spread, -constr]). In other words, this measure can be seen to present information for three features. However, the last two features cannot be considered independently, since values of [+spread] and [+constr] are not simultaneously possible, but must be considered as a pair. It can then be stated that the amplitude of H1 during the closure gives information for both the [stiff, slack] pair, as well as for the [spread, constr] pair. Accordingly, a weight of two was assigned to this measure.

Examination of the data shows that amplitude of H1 during the closure interval is more robust than the other measures, so that in terms of reliability, a greater weight on this measure does not degrade performance. It must be noted, however, that this weighting scheme may not be applicable across all cases, particularly if an acoustic cue appears more prominently in certain contexts, and not in others. Moreover, this scheme for assigning weights to acoustic cues for certain features is inherently dependent on language, since the acoustic cues used to identify features may be used differently to describe classes of consonants for different languages.

Another possible method of extracting acoustic cues from the signal relates to graded measurements. In this thesis, a voicing decision was made for each individual measurement and the results were combined for each landmark. It is also possible to assign a graded value for each measurement, and those values used to determine the voicing decision for the landmark. This scheme would allow explicit used of knowledge where the acoustic cues give less reliable or ambiguous information in determining the values for the underlying features.

The performance of the scheme proposed in this thesis was assessed by comparing the results of the voicing decisions with the lexical definition of voicing for each consonant, and also with perceived voicing. It must be noted that the perceived voicing was found by informal listening of the speech signal by the author, who is a non-native speaker of English. The experimental results show that voicing decisions counted as errors (when compared with lexical voicing) involved most of the cases where voicing was perceived as modified, but a more accurate perception test involving several native speakers of English is needed for a more rigorous analysis. Alternatively, hand analysis of each consonant to determine the actual realization of consonant voicing may also be conducted.

The classification experiments presented in this thesis include cases where the training and test sets overlap, as well as cases where the two sets are disjoint. Results obtained from training and testing on the same set give an indication of the separability of that data set, and are usually better than training and testing on different sets. For a more rigorous comparison between the different cases, it is necessary to conduct experiments where the test data is not included in the training set. This may be accomplished by using multiple repetitions of utterances in the same context by the same speaker. Alternatively, testing each utterance of a set while training with the remaining utterances may be possible.

In addition, it may be possible to refine the simple classification scheme used in this thesis to obtain the voicing decisions. For example, the distance to the means of the distributions for voiced and unvoiced consonants for each measure was used as to determine the voicing decision. Other methods may employ a weighted distance measure that takes into account the variations of the distributions. Another possible procedure would used the median values rather than the means of the distributions.

The voicing module proposed in this thesis receives information such as times in the signal corresponding to consonant landmarks, and specification of the landmark types. Such information has been found manually, and used in classification schemes involving both manual and automatic measurements. The results of experiments in this thesis suggest that other higher level information, such as position of the consonant within a syllable and stress position, may be useful. For example, less weight may be placed on landmarks for consonants affiliated with weak syllables, when landmark pairs are converted into segments. Higher level information is also needed to make use of phonotactic knowledge. For example, the sequence /sb/ is possible only at syllable boundaries in English, as in "baseball," while the sequence /sp/ is possible in all contexts, as in "spot," "display," and "grasp." In this thesis, consolidation of voicing decisions for each consonant landmark into a voicing decision for a consonantal segment uses a simple averaging scheme which weighs each landmark equally. Using contextual information, different weights could be given to either the closure or release.

Bibliography

[1] F. Bell-Berti. Control of pharyngeal cavity size for English voiced and voiceless stops. *Journal of the Acoustical Society of America*, 57:456 – 461, 1975.

[2] W. N. Campbell and S. D. Isard. Segment durations in a syllable frame. *Journal of Phonetics*, 19:37 – 47, 1991.

[3] A. Ni Chasaide and C. Gobl. Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, 36(2,3):303 – 330, 1993.

[4] J. Y. Choi, E. Chuang, D. Gow, K. Kwong, S. Shattuck-Hufnagel, K. Stevens, and Y. Zhang. Labeling a speech database with landmarks and features (4aSC6). In *Program of the 134th Meeting of the Acoustical Society of America*, page 3163, December 1997.

[5] N. Chomsky and M. Halle. *The Sound Pattern of English.* Harper and Row, New York, 1968.

[6] C. S. Crowther and V. Mann. Native language factors affecting use of vocalic cues to final consonant voicing in English. *Journal of the Acoustical Society of America*, 92:711 – 722, 1992.

[7] T. H. Crystal and A. S. House. Segmental durations in connected-speech signals: current results. *Journal of the Acoustical Society of America*, 83:1553 – 1573, 1988.

[8] P. C. Delattre, A. M. Liberman, and F. S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27:769 – 774, 1955.

[9] P. B. Denes and E. N. Pinson. *The Speech Chain.* Anchor Press/Doubleday, Garden City, New York, 1973.

[10] T. J. Edwards. Multiple features analysis of intervocalic English plosives. *Journal of the Acoustical Society of America*, 69:535 – 545, 1981.

[11] Entropic Research Laboratory, Inc. xwaves+ *Manual*, March 1996.

[12] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1970.

[13] B. D. Fear, A. Cutler, and S. Butterfield. The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, 97:1893 – 1904, 1995.

[14] O. Fujimura. Analysis of nasal consonants. *Journal of the Acoustical Society of America*, 34:1865 – 1875, 1962.

[15] H. Hanson. *Glottal characteristics of female speakers.* PhD thesis, Harvard University, Cambridge, MA, 1995.

[16] R. E. Hillman, E. Oesterle, and L. L. Feth. Characteristics of the glottal turbulent noise source. *Journal of the Acoustical Society of America*, 74:691 –694, 1983.

[17] N. Hiraoka, Y. Kitazoe, and H. Ueta. Harmonic-intensity analysis of normal and hoarse voices. *Journal of the Acoustical Society of America*, 76:1648 – 1651, 1984.

[18] D. Kewley-Port. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, 72:379 – 389, 1982.

[19] S. K. Keyser and K. N. Stevens. Feature geometry and the vocal tract. *Phonology*, 11:207 – 236, 1994.

[20] D. H. Klatt. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59:1208 – 1221, 1976.

[21] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820 – 857, 1990.

[22] K. Kohler. F0 in the perception of lenis and fortis plosives. *Journal of the Acoustical Society of America*, 78:21 – 32, 1985.

[23] I. Lehiste and G. E. Peterson. Transitions, glides and diphthongs. *Journal of the Acoustical Society of America*, 33:268 – 277, 1961.

[24] L. Lisker. Minimal cues for separating /w,r,l,y/ in intervocalic position. *Word*, 13:256 – 267, 1957.

[25] L. Lisker and A. S. Abramson. A cross-language study of voicing in initial stops: acoustical measurements. *Word*, 20:324 – 422, 1964.

[26] S. A. Liu. *Landmark detection for distinctive feature-based speech recognition.* PhD thesis, MIT, Cambridge, MA, May 1995.

[27] A. Lofqvist, L. L. Koenig, and R. S. McGowan. Vocal tract aerodynamics in /aCa/ utterances: Measurements. *Speech Communication*, 16:49 – 66, 1995.

[28] A. Lofqvist and H. Yoshioka. Intrasegmental timing: laryngeal-oral coordination in voiceless consonant production. *Speech Communication*, 3:279 – 289, 1984.

[29] J. C. Lucero. A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. *Journal of the Acoustical Society of America*, 105:423 –431, 1999.

[30] S. Y. Manuel and K. N. Stevens. Formant transitions: Teasing apart consonant and vowel contributions. In *Proceedings of the International Congress of Phonetic Sciences*, volume 4, pages 436 – 439, 1995.

[31] B. T. Oshika, V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach. The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):104 – 112, 1975.

[32] G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175 – 185, 1952.

[33] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE Acoustics, Speech and Signal Processing Magazine*, pages 4–16, January 1986.

[34] K. N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17:3 – 45, 1989.

[35] K. N. Stevens. Modelling affricate consonants. *Speech Communication*, 13:33 – 43, 1993.

[36] K. N. Stevens. Models for the production and acoustics of stop consonants. *Speech Communication*, 13:367 – 375, 1993.

[37] K. N. Stevens. *Acoustic Phonetics.* MIT Press, Cambridge, MA, 1999.

[38] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. Acoustic and perceptual characteristics of voicing in fricative and fricative clusters. *Journal of the Acoustical Society of America*, 91:2979 – 3000, 1992.

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

[40] W. Van Summers. F1 structure provides information for final-consonant voicing. *Journal of the Acoustical Society of America*, 84:485 – 492, 1988.

[41] W. Sun. Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition. Master's thesis, MIT, Cambridge, MA, May 1996.

[42] I. R. Titze. Phonation threshold pressure: A missing link in glottal aerodynamics. *Journal of the Acoustical Society of America*, 91:2926 – 2935, 1992.

[43] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody. F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America*, 93:2152 – 2159, 1993.

[44] C. G. Wolf. Voicing cues in English final stops. *Journal of Phonetics*, 6:299 – 309, 1978.

[45] H. Yoshioka, A. Lofqvist, and H. Hirose. Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America*, 70:1615 – 1623, 1981.

[46] Y. Zhang. Toward implementation of a feature-based lexical access system. Master's thesis, MIT, Cambridge, MA, May 1998.