Consonant Landmark Detection for Speech Recognition

by

Chiyoun Park

B.S., Korea Advanced Institute of Science and Technology (2002)
S.M., Massachusetts Institute of Technology (2004)

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 2008

Author . . . .
Department of Electrical Engineering and Computer Science
May, 2008

Certified by . . . .
Kenneth N. Stevens
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Consonant Landmark Detection for Speech Recognition

by

Chiyoun Park

Submitted to the Department of Electrical Engineering and Computer Science
on May, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis focuses on the detection of abrupt acoustic discontinuities in the speech signal, which constitute landmarks for consonant sounds. Because a large amount of phonetic information is concentrated near acoustic discontinuities, more focused speech analysis and recognition can be performed based on the landmarks. Three types of consonant landmarks are defined according to its characteristics—glottal vibration, turbulence noise, and sonorant consonant—so that the appropriate analysis method for each landmark point can be determined.

A probabilistic knowledge-based algorithm is developed in three steps. First, landmark candidates are detected and their landmark types are classified based on changes in spectral amplitude. Next, a bigram model describing the physiologically-feasible sequences of consonant landmarks is proposed, so that the most likely landmark sequence among the candidates can be found. Finally, it has been observed that certain landmarks are ambiguous in certain sets of phonetic and prosodic contexts, while they can be reliably detected in other contexts. A method to represent the regions where the landmarks are reliably detected versus where they are ambiguous is presented.

On TIMIT test set, 91% of all the consonant landmarks and 95% of obstruent landmarks are located as landmark candidates. The bigram-based process for determining the most likely landmark sequences yields 12% deletion and substitution rates and a 15% insertion rate. An alternative representation that distinguishes reliable and ambiguous regions can detect 92% of the landmarks and 40% of the landmarks are judged to be reliable. The deletion rate within reliable regions is as low as 5%.

The resulting landmark sequences form a basis for a knowledge-based speech recognition system since the landmarks imply broad phonetic classes of the speech signal and indicate the points of focus for estimating detailed phonetic information. In addition, because the reliable regions generally correspond to lexical stresses and word boundaries, it is expected that the landmarks can guide the focus of attention not only at the phoneme-level, but at the phrase-level as well.

Thesis Supervisor: Kenneth N. Stevens
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First of all, I would like to thank my supervisor, Ken Stevens, for his guidance. He introduced me to the world of speech, encouraged me to expand my imagination about his theory of landmark, and inspired me to form these ideas into shape. His insights and advice helped me overcome difficulties in every step of the research.

I also give thanks to other members of the thesis committee, Janet Slifka and Louis Braida, for reading the thesis draft and giving a lot of helpful suggestions for improving it. I especially thank Janet for all the advice and encouragement during the time when I first started working on this topic.

Thanks to all the members of the Speech Communication Group:

Special thanks to Nancy Chen and Youngsook Jung for their help in shaping my coarse and unsettled idea into a working model. They have given me the motivation and passion for the research, and every bit of discussion with them was both helpful and fun. If it were not for them, it would not have been possible to pursue this work to where it is now.

I am also deeply indebted to Stefanie Shattuck-Hufnagel for her great interest in this work and her support. She have shown me different aspects and usages of landmarks besides speech recognition and provided me valuable knowledge about speech, which greatly broadened my view.

Thanks to Caroline Huang for organizing the lunch discussion group. The biweekly meeting really helped me organize my thought. I also thank Xuemin Chi, who is submitting her thesis at the same time as me, for helping me with preparation of the thesis and defense. And thanks to Arlene Wint for making sure that I am not missing any social events around the lab.

Finally, I would like to thank my friends and my family for their love and support all through the years.

This research was supported by NIH DC02978.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Non-uniform Distribution of Speech Information

In a speech signal, phonetic information is not distributed uniformly across the whole utterance. For example, the silent region at around 900ms in Figure 1-1 does not provide any phonetically-related cues about the utterance, although it may indicate a possible word or phrase boundary. On the other hand, a lot of information can be found in the vicinity of the abrupt spectral discontinuity near the 400ms time-point [51]. For example, low-frequency energy below 1kHz is persistent across the transition but the higher frequency amplitude suddenly increases, indicating a possible transition from a nasal sound into a vowel. Changes in the first formant frequency and bandwidth provide additional evidences that support this assumption. The first (F1), second (F2) and third (F3) formant frequencies slightly increase near the transition, which suggests that the consonant before the vowel is likely to be labial. The F2 frequency on the right-hand side of the transition is as high as 2kHz, which is a characteristic of a front vowel. Thus, transition regions in the acoustic signal are particularly rich in cues to feature contrasts.

Speech information is also distributed in non-transitional periods as well. For example, in the vowel sound which spans through the region between 1000ms and

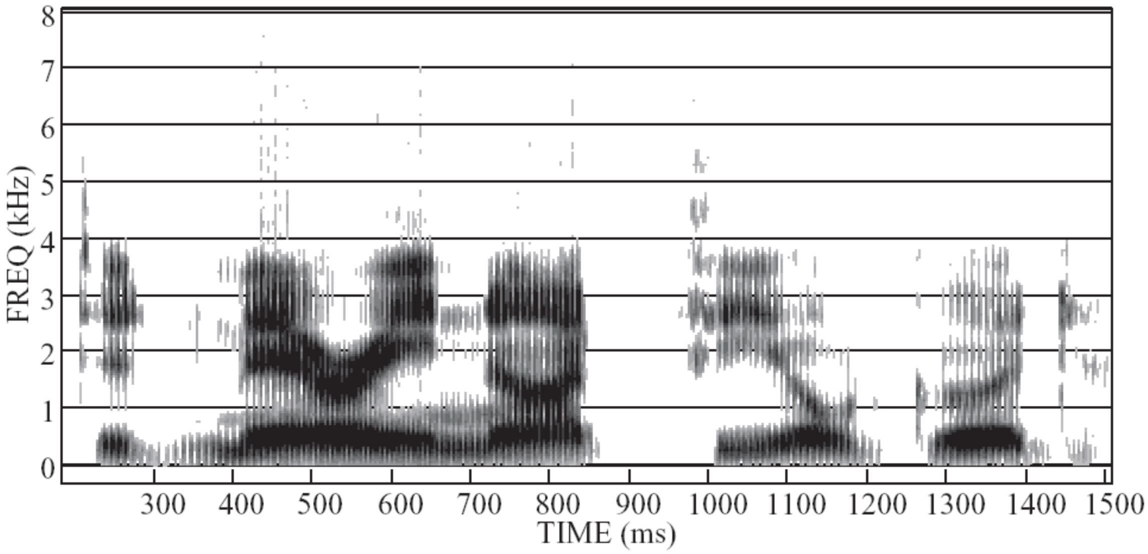


Figure 1-1: A spectrogram of an utterance “Did Mary not feel good?”

1100ms, the low F1 formant frequency and high F2 frequency show that this is a high vowel as well as a front vowel. The information that can be found in this region is not as rich as that found near consonant-vowel transitions, and the spectrum is almost steady during the 100ms period, which implies that different points in the vowel region contain similar characteristics. This property can be helpful in that the information can be estimated reliably over a long period of time, due to the steadiness of the signal. Moreover, additional characteristics besides phonetic information, such as vowel quality and clues for speaker identification, may be found in this region.

Non-abrupt spectral changes are another indication of phonetically important events. For example, at around the 530ms time-point of Figure 1-1, the second and third formant frequencies decrease and then increase slowly over a 150ms period. Such a change normally occurs due to a glide or liquid sound or an off-glide of a vowel. In this example, F3 formant frequency decreased significantly, which is a characteristic of an /r/ sound.

Therefore, a number of advantages can be gained by first locating and classifying these points with different acoustic distributional characteristics. That is, a more focused and detailed analysis can be performed at abrupt transitions where a large amount of phonetic information is likely to be concentrated; cues can be estimated with higher confidence during a steady-state region; and less abrupt changes can also be highlighted and handled according to their characteristics. These points of acoustic importance are called landmarks.

1.2 Landmarks and Speech Analysis

1.2.1 Landmarks as Focal Points

Due to the non-uniform distribution of speech information in the signal, a listener does not have to listen to all the time-points of a speech signal equally carefully in order to understand it. Instead, the focus can be placed on the instances where more information is located. Perceptual experiments support the hypothesis that humans focus on the regions of abrupt change, where a large amount of information is concentrated.

Jenkins *et al.* [27] performed an experiment that supports the assumption that the information near consonant-vowel boundaries plays an important role in the classification of vowels. This experiment compares the perception of vowels in /b/+vowel+/b/ syllables presented in three types: when each stimulus is presented as a whole syllable, when the steady-state portions are replaced with silence, and when the center portions are given without the transitional periods. The three types of stimuli are illustrated in Figure 1-2.

The result of this experiment shows that the perception error does not change by much (less than 1 percent point difference) when the steady-state portion of the signal is omitted, but when the transitional periods are deleted, the error rate increases by 6%. From this experiment, it can be concluded that the information in the transitional periods at the onset and offset of a vowel is more important than that provided by the steady-state portion of the syllable in the identification of a vowel sound.

Furui [13] made a similar observation on Japanese syllables by performing perceptual experiments on truncated speech signals. By cutting off the initial part or the

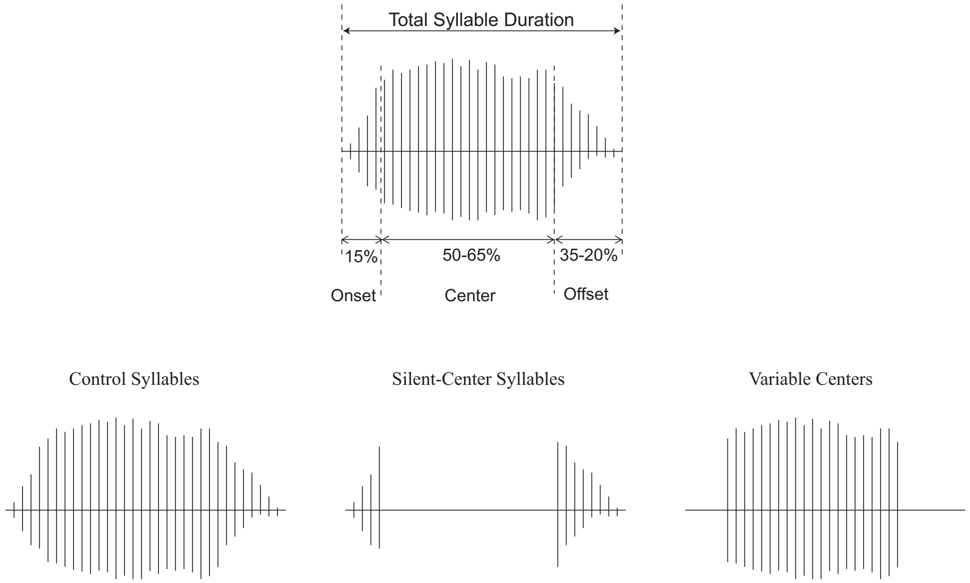


Figure 1-2: Three types of stimuli in the vowelless perception experiment (from Jenkins *et al.* [27])

final part of a syllable at different time points, he found out that there is a critical point at which the identification rate rapidly decreases, and the perceptual critical points for initial and final truncations are separated by only 10ms. This result provides evidence for the assumption that the information in a speech signal is highly concentrated near certain time points.

Stevens [49] carried out perceptual experiments based on synthesized syllables with varying parameters, and also observed that the distinction between /s/ and /θ/, and between /s/ and /ʃ/, depend not only on the spectral shape and amplitude of frication noise itself, but also on the amplitude change at the acoustic boundary between the fricative and the vowel.

Therefore, it can be reasonably assumed that by first locating the abrupt acoustic changes in a signal, and then carrying out a focused analysis around the landmark, important perceptual information can be extracted from the signal. As was discussed in the previous section, landmarks provide the time-points where speech events with acoustical prominence occur, and the type of a landmark—abruptness of a consonant, steady-state period of a vowel, and non-abrupt transition of a glides—implies some of the feature characteristics of the event and also specifies the typical acoustic cues that can be found nearby. Therefore, a more knowledge-based method of speech analysis and recognition, which takes into account different characteristics of signal, can be initiated by first recognizing landmark locations and types.

1.2.2 Landmarks as Boundaries

While an individual landmark pinpoints the location where speech-related information is concentrated, a pair of landmarks marks the boundaries of a region where a certain phonetic cue spans throughout. By locating the extent where one particular phonetic feature has a dominant influence, the feature value can be estimated more reliably.

A perceptual experiment by Jongman [28] supports this claim by showing that at least a 50ms interval is needed to determine the place of articulation features of non-strident fricatives with 80% confidence, and when the whole extent of the frication is provided, the correct identification of the place of articulation approaches 100%.

In addition, the length of the regions divided by landmarks provide temporal cues to the speech events. An experiment by Jenkins *et al.* [27] also shows that when both onset and offset parts of a vowel segment were presented without the steady-state portion of the vowel, the listeners classified the vowel with significantly greater accuracy than in the case when only one of the two boundaries was given. This result suggests that the durational cues presented by a sequence of landmarks can be helpful in the perception of vowel sounds.

1.3 Thesis Overview

The main object of this thesis is to implement an automatic algorithm that locates the consonant landmarks of a speech signal. Because the consonant landmarks not only correspond to quantal acoustic changes that occur during closures and releases of consonants, but also correpond to the time-points where phonetic information is highly concentrated, this can be applied as the initial step of a speech analysis and recognition system.

The next three chapters provide preliminary knowledge used in this thesis. In Chapter 2, the definition of landmarks is clarified, and three different consonant landmark types are defined. The landmark types represent different characteristics of consonants, and so the articulatory and acoustic characteristics of each landmark type are examined as well. In addition, the list of distinctive features used in this thesis is explained in this chapter, and the close relationship between the landmark types and the distinctive features are explored.

In Chapter 3, the landmark-based speech recognition system proposed by Stevens [52] is described, and the current status of the development of this system is reviewed. The landmark detection algorithm developed in this thesis aims to be applied as the first step of this speech recognition system, which both provides information about certain features and defines the focus of attention for the subsequent processes.

Chapter 4 explains the preliminary steps in the development of landmark detection algorithm. First, a landmark detection algorithm previously developed by Liu [36] is briefly reviewed. From this background, possible avenues for improvement of the algorithm are discussed and the general characteristics of the new system are determined. This chapter also discusses the characteristics of the database and the data preparation method, including an algorithm that maps phonetic transcriptions to landmarks.

The implementation of the landmark detection algorithm is described in the following three chapters, each of which deals with different aspect of the algorithm. Chapter 5 provides a probabilistic algorithm that detects different types of landmarks individually without considering the relationship between adjacent landmarks. This procedure detects possible landmark locations for different landmark types, and then calculates their probabilities. This process aims to locate as many acoustic discontinuities as possible, so that only a small number of true landmarks are missing.

In Chapter 6, the constraints in landmark sequencing are investigated, and a bigram model is used to represent the restrictions in the transition between different types of landmarks. An algorithm that determines the most likely sequence of landmarks based on the bigram transition model in addition to individual landmark probability is developed in this chapter. Some of the contexts in which the landmarks are likely to be detected with more ambiguity are analyzed as well.

Chapter 7 presents an alternative algorithm that selects a possible landmark sequence from the previously detected landmark candidates. A representation that distinguishes the regions within which the landmarks are reliably detected, from the regions where the landmarks are ambiguous, is discussed, and an algorithm that creates such a representation is developed. This algorithm makes it possible to focus on the regions where the utterance is produced more clearly.

The last chapter summarizes the results of this thesis, discusses possible applications within the landmark-based speech recognition system, and suggests further improvements of the landmark detection algorithm as well as possible applications of the landmark detector beyond a speech recognition system.

Chapter 2

Landmarks and Distinctive Features

2.1 Landmarks

2.1.1 Definition of Landmarks

As was mentioned in the previous chapter, landmarks are defined as the time-points of acoustic events that are consistently correlated to major articulatory movements. Landmarks correspond to abrupt discontinuities in the spectrum (consonants), locally stable points of the spectrum (vowels), or slow movements of the formant frequencies (glides). Five different types of landmarks—three for consonant, one for vowel, and one for glide—are defined according to their acoustic characteristics. The list of five landmarks and their descriptions are shown in Table 2.1, and the details of these landmark types will be discussed later in this section.

Because the landmark types are closely related to the manner of articulation, it is expected that the type of information that can be best estimated near each landmark can be predicted by distinguishing the landmark types. In addition, the study of Huttenlocher and Zue [24] shows that the number of possible word candidates can be reduced significantly by knowing the manner features.

Segment Type	Landmark		Description
Consonant	g	glottis	Vocal fold vibration
	s	sonorant	Velopharyngeal port opening
	b	burst	Turbulence noise source
Vowel	V	vowel	Vowel
Glide	G	glide	Glide

Table 2.1: Five different landmarks and related events

Consonant Landmarks

Consonants are usually produced by abrupt formation and release of a constriction in the mouth. This articulatory movement affects the acoustics so that the spectrum of the speech signal changes abruptly at the time-point where the consonant closure or release is produced [36, 35].

Three types of consonant landmarks have been proposed by Liu [35]: g (glottis), b (burst), and s (sonorant). Examples of each type of consonant landmarks in an actual utterance are illustrated in Figure 2-1. The spectrogram is extracted from the utterance of the sentence “Did Mary not feel good?” recorded by a male speaker. The lines represent the expected consonant landmarks of this signal. Dotted lines are located at b-landmarks, thick solid line at s-landmarks, and thin solid line at g-landmarks.

A g-landmark pinpoints a time when vocal folds start vibrating freely or when the vibration ends or gets suppressed due to increased intraoral pressure. The voice bar after the closure of a voiced stop consonant and the voicing during a voiced fricative consonant are examples of suppressed vocal fold vibrations. Therefore, the g-landmarks distinguish obstruent consonants or silence from vowels or sonorant consonants.

A b-landmark corresponds to the existence of turbulence noise during obstruent regions. Therefore, it is set at the boundary between a silent interval and a frication noise of a stop or affricate burst. The silent interval is usually caused by a complete closure inside the oral cavity, but the silence can also occur due to the wide opening of

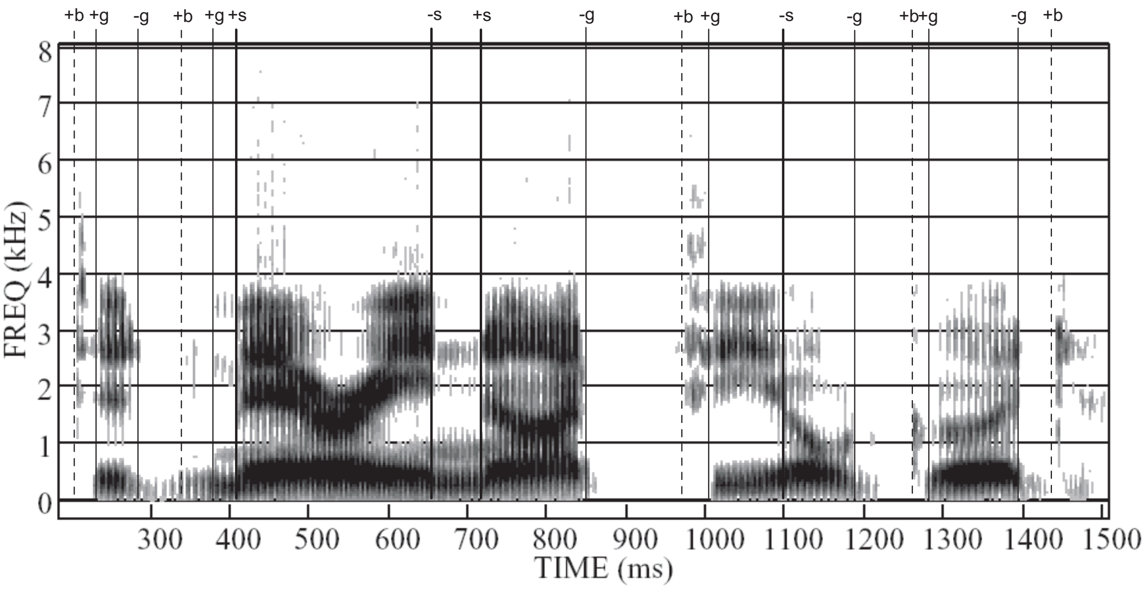


Figure 2-1: Examples of consonantal landmarks from an utterance "Did May not feel good?"

the vocal folds at the start or end of an utterance or at a pause within an utterance.

An s-landmark mostly corresponds to opening or closing of the velopharyngeal port during a sonorant sound. Phonetically, it is located at the boundary between a vowel or glide and a sonorant consonant. Although an /l/ sound is not pronounced with a velopharyngeal port opening, abrupt /l/ is included in the class of sonorant consonants because the acoustic characteristic of abrupt /l/ sound is similar to that of nasals, which will be explained in the following sections in more detail. The segments /r/ and non-abrupt /l/, which is produced without making clear contact between the tongue tip and the roof of the mouth, are considered glides in landmark type because they do not accompany acoustic discontinuities.

The three types of consonantal landmarks are further classified depending on the increase or decrease of energy in the vicinity of the landmark. When the energy increases around a landmark, the landmark is classified as '+', and if it decreases, the landmark is classified as '−'. For example, at 650ms time-point of the spectrum in Figure 2-1, which is the transition from a vowel to the following nasal at the word boundary of 'Mary not', the high-frequency energy decreases due to opening of the velopharyngeal port. Therefore, the acoustic discontinuity is classified as a −s landmark. On the other hand, the 720ms time-point, which corresponds to the transition from /n/ to /ɑ/ of the word 'not', is classified as a +s landmark due to the rise in energy amplitude.

Because g-landmarks are at the boundaries of sonorant-obstruent sounds, and b- and s-landmark types are defined only for obstruent and sonorant regions, respectively, the sequence of landmarks have intrinsic constraints; for example, a b-landmark can be found only between a −g landmark and the next +g landmark, and an s-landmark can be found only between a +g landmark and the next −g landmark.

Vowel and Glide Landmarks

Vowels and glides are produced without creating an oral constriction narrow enough to produce turbulence noise or silence. Therefore, these segments do not create abrupt discontinuities in the spectrum, and so it is difficult to set a clear boundary between two vowels, or between a glide and a vowel. Thus, instead of locating the boundaries of vowel or glide segments, landmarks for these classes of speech sounds are defined to be the position where the effect of the vowel or glide on the acoustic signal is the most dominant.

Figure 2-2 shows examples of vowel and glide landmarks extracted from the utterance of the sentence "Which year were you lazy?" recorded by a female speaker. The lines represent the expected vowel and glide landmarks of this signal. Dashed lines indicate the position of G landmarks, and the solid lines are located at V-landmarks.

A vowel is produced with a maximum opening in the vocal tract, and when a closure or narrowing is made in the oral cavity, the first formant frequency goes down and its bandwidth widens. Thus, a vowel landmark can be located in the vowel where the first formant frequency is the highest, or where there is maximum amplitude in the range of the first formant frequency [23, 50].

On the other hand, a glide can be identified with minimum amplitude in the low frequency range, and slow movements of the formants [50]. Therefore, according to Sun [55], this landmark can be determined by a combination of low F1 frequency,

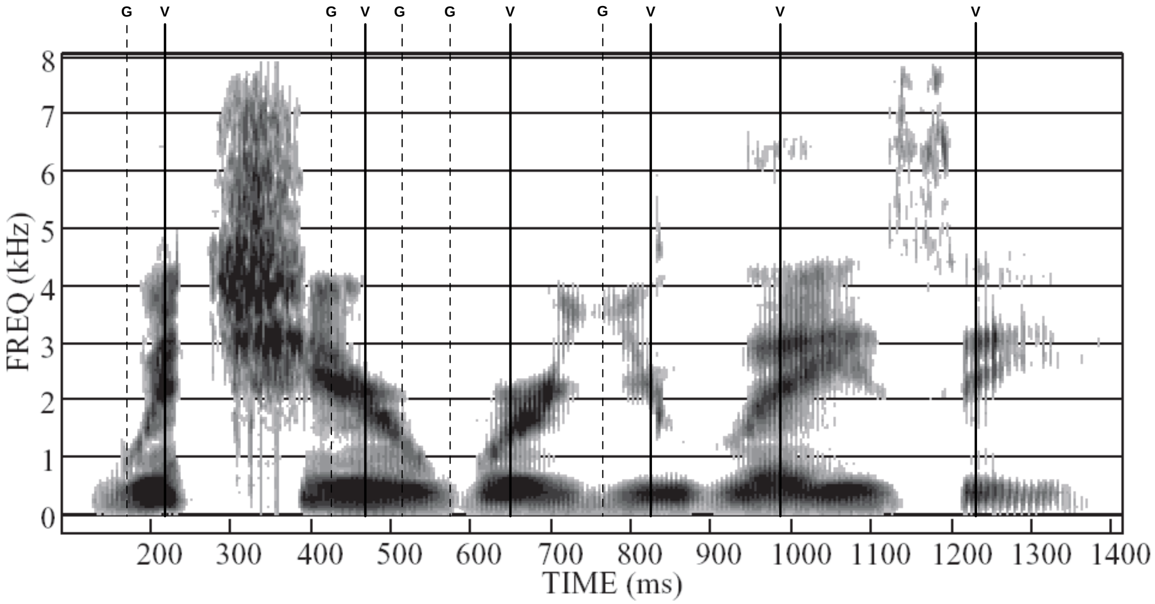


Figure 2-2: Examples of vowel and glide landmarks from an utterance "Which year were you lazy?" The symbol V indicates vowel landmarks and G indicates glide landmarks.

maximum F1 rate of change, along with low mean-square amplitude in energy and a maximum in its rate of change.

2.1.2 Articulatory Characteristics of Consonant Landmarks

As was discussed in the previous section, information that can be found in a speech signal can be located by three different classes of landmarks: abrupt discontinuities by consonant landmarks, steady-state periods by vowel landmarks, and non-abrupt transitions by glide landmarks. This thesis will focus on the consonant landmarks alone.

A consonant is pronounced by making a complete closure or a significant narrowing in the oral tract. The constriction can be formed with one of the following three articulators: lips, tongue blade and tongue body. Such articulatory movement modifies the overall shape of the vocal tract, and this change results in an acoustic discontinuity in the speech signal.

However, not all articulatory movements related to consonants result in abrupt

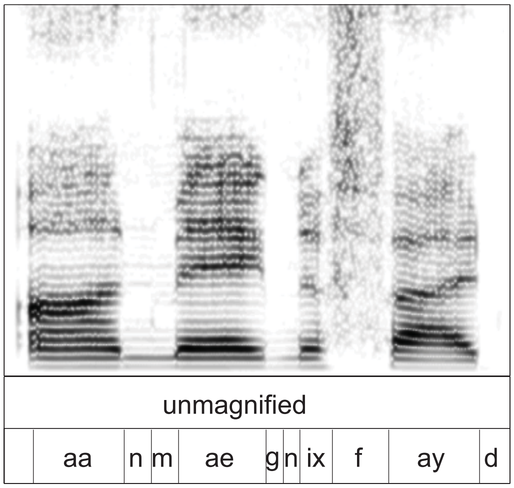


Figure 2-3: An example of the change in place of articulation

changes. For example, the change of the tongue position in the transition from /m/ to /n/ in the word 'alumnus' or 'remnant' reduces the size of oral cavity and the resonance pattern changes accordingly, but as can be seen in Figure 2-3, its acoustic discontinuity between /n/ and /m/ sounds is not as prominent as the one in the transition from a vowel into /m/. Change in the place of articulation does not affect the overall shape of the vocal tract, and the corresponding acoustic change is generally not consistent or prominent.

Therefore, it can be hypothesized that the acoustic abruptness is caused only by a significant change in the general configuration of the vocal tract shape. The vocal tract configurations can be roughly classified into one of the four types shown in Figure 2-4: A vowel is pronounced without making a constriction in the oral tract (*open*), a stop consonant is preceded by a complete closure in the oral tract, resulting in a silent period before the burst release (*closure*), a fricative consonant is made with a constriction which is narrow enough to generate turbulence noise across the constriction (*turbulence*), and a nasal consonant is made with a complete closure in the mouth but the nasal passage is open (*side-branch*). The acoustic abruptness arises when the articulatory movement changes the vocal tract shape among these four types of configurations.

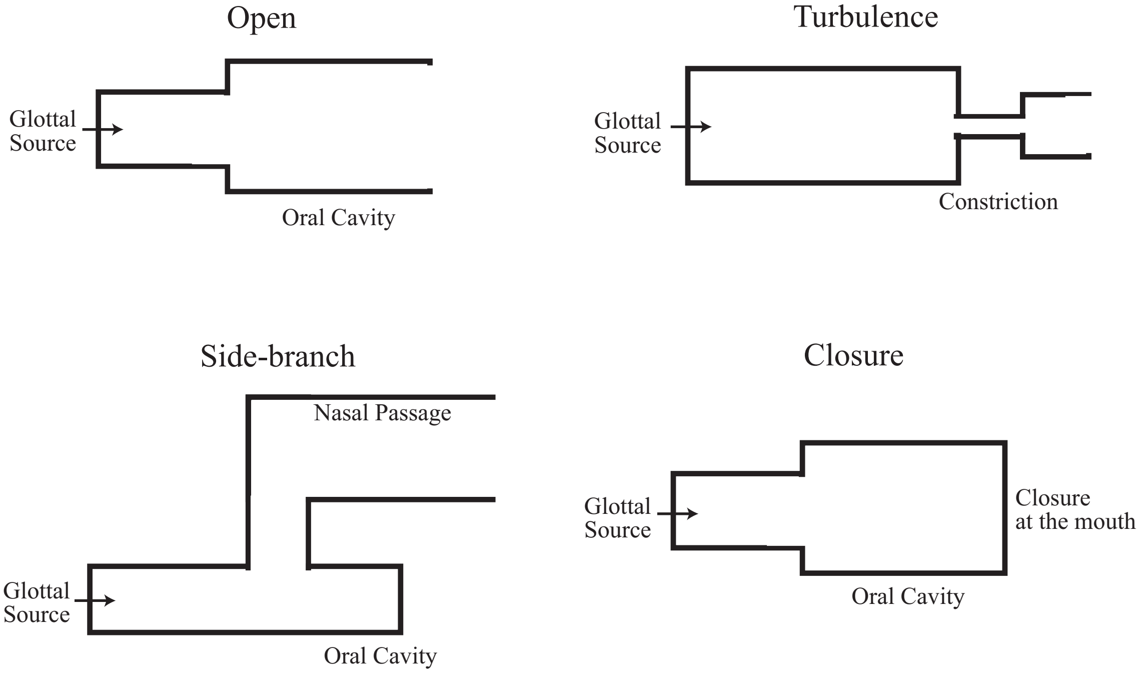


Figure 2-4: Four different types of vocal tract configurations

The characteristic of the acoustic discontinuity differs according to the manner of the closure. For example, a /p/ sound is pronounced with a complete closure, which increases the intraoral pressure and stops excitation of the resonance of the vocal tract. On the other hand, an /m/ sound makes the same closure with the same articulator, but because the nasal passage is still open, the intraoral pressure does not increase as much and excitation of the resonance persists throughout the closure, although it may be less prominent in the high-frequency region due to the introduction of pole-zero pairs by the oral cavity. More detailed correlation between the vocal tract shapes and their acoustic realizations are explained in Section 2.1.3.

Theoretically, there can be twelve different types of transitions between the four configurations. These consonantal discontinuities can be classified according to the acoustic characteristics that can be found nearby. The glottal vibration changes at the transitions between the configurations with relatively open vocal tract (open and side-branch configurations) and the configurations with a constriction (closure and turbulence configurations). When there is a constriction in the vocal tract, intraoral pressure increases, which reduces the amount of pressure drop across the glottis. Thus, the vibration of the vocal folds is stopped or suppressed during these types of configurations. On the other hand, when the vocal tract is open, the intraoral pressure does not increase and the vocal folds can vibrate freely.

Other transitions are related to certain acoustic characteristics as well. The transition between closure and turbulence configuration identifies an onset of a burst noise or an offset of a frication into a silence, and the transition between open and side-branch configurations indicates the acoustic change due to the pole-zero pairs introduced by the side-branch.

2.1.3 Acoustic Correlates of Consonant Landmarks

As was mentioned in the previous section, the acoustic discontinuities arise when the articulatory movement significantly changes the overall configuration of the vocal tract shape. An example of a word which contains all four types of vocal tract configurations is shown in Figure 2-5. The vowels are pronounced with open configurations, the nasal /m/ with side-branch, /z/ with turbulence, and /d/ closure with closure configuration. At each boundary between these segments, an abrupt change can be observed. These discontinuities can be further classified according to their characteristics: spontaneous vibration of vocal folds, existence of additional poles and zeroes due to a side branch in the vocal tract, and turbulence noise source. Acoustic properties of each type of abruptness are discussed in this section.

Glottal Vibration

Fundamental frequency of vocal fold vibration typically ranges from 80Hz to 210Hz for male speakers, 150Hz to 320Hz for typical female speakers, and 300Hz and higher for children [1]. Therefore, it is expected that the low-frequency energy at about 0-400Hz frequency band directly relates to vocal fold vibrations.

The 0–400Hz frequency band energies in each type of vocal tract configuration are measured from 5,000 segments from the TIMIT database (see Section 4.2.1 for more

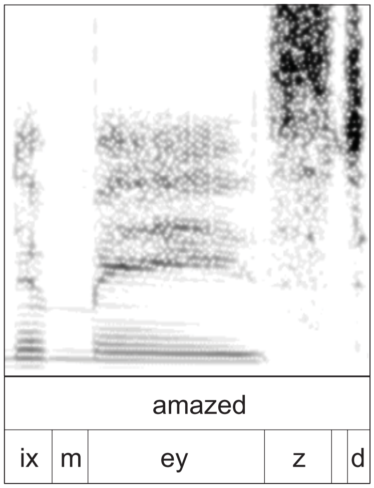


Figure 2-5: A speech signal containing all four types of vocal tract configurations

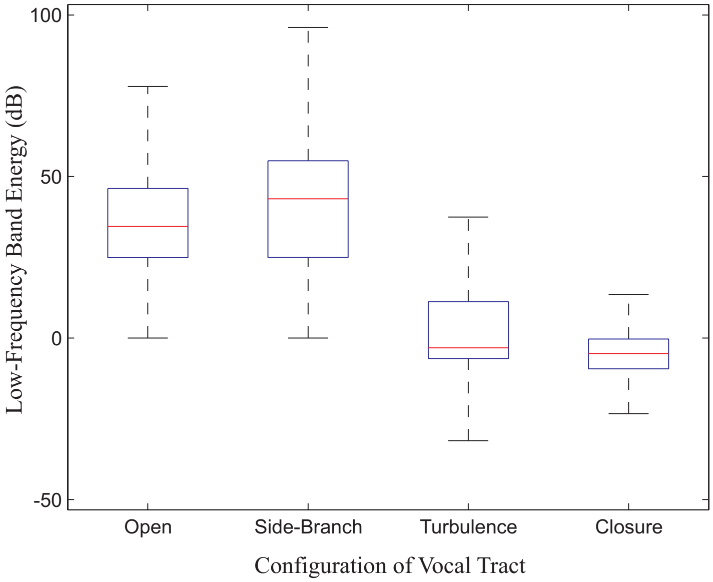


Figure 2-6: A box-plot of low-frequency (0–400Hz) band energies in different types of vocal tract configurations

detail), and the box-plot of the result is shown in Figure 2-6. The measurements are made in the middle of the segment, and their dB values are calculated relative to the energy level within the silent region estimated from the first 30ms of the utterance. Each box of the plot spans between the first and the third quartiles of the measured data, and the line in the middle is located at the median value. It can be observed that most of the utterances pronounced with open or side-branch configuration have more than 25dB band energy, whereas the ones with narrow or complete constriction have less than 25dB energy.

The side-branch configuration has even higher low-frequency energy on average than the open configuration. This may be because a nasal resonance is introduced near 250Hz, which is the lowest natural resonance of the whole vocal tract including the nasal passage, and the added nasal-pole reinforces the fundamental or the second harmonic by about 5dB amplitude [22, 8]. The low mean energy in vowels may be caused by a large number of schwas as well.

It can be noted that the low-frequency energy in turbulence configuration spans almost as high as that of open or side-branch configuration. This is due to the fact that the vocal folds may keep vibrating during a voiced fricative sound like /v/ or /z/. However, because the intraoral pressure must be increased to make turbulence noise at the constriction, the glottal source amplitude is reduced at least by 9dB relative to that of the neighboring vowel [35]. Therefore, the voicing in a fricative or the voice bar of a stop consonant does not have as high glottal vibration energy as that of an open configuration.

The glottal vibration not only increases the low-frequency energy, but also excites high-frequency ranges. However, because nasals usually have less prominent high-frequency energy, and strident fricatives are pronounced with turbulence noise whose high-frequency amplitude is greater than that of the neighboring vowel, the high-frequency band energy cannot be used as a consistent measure of glottal vibration.

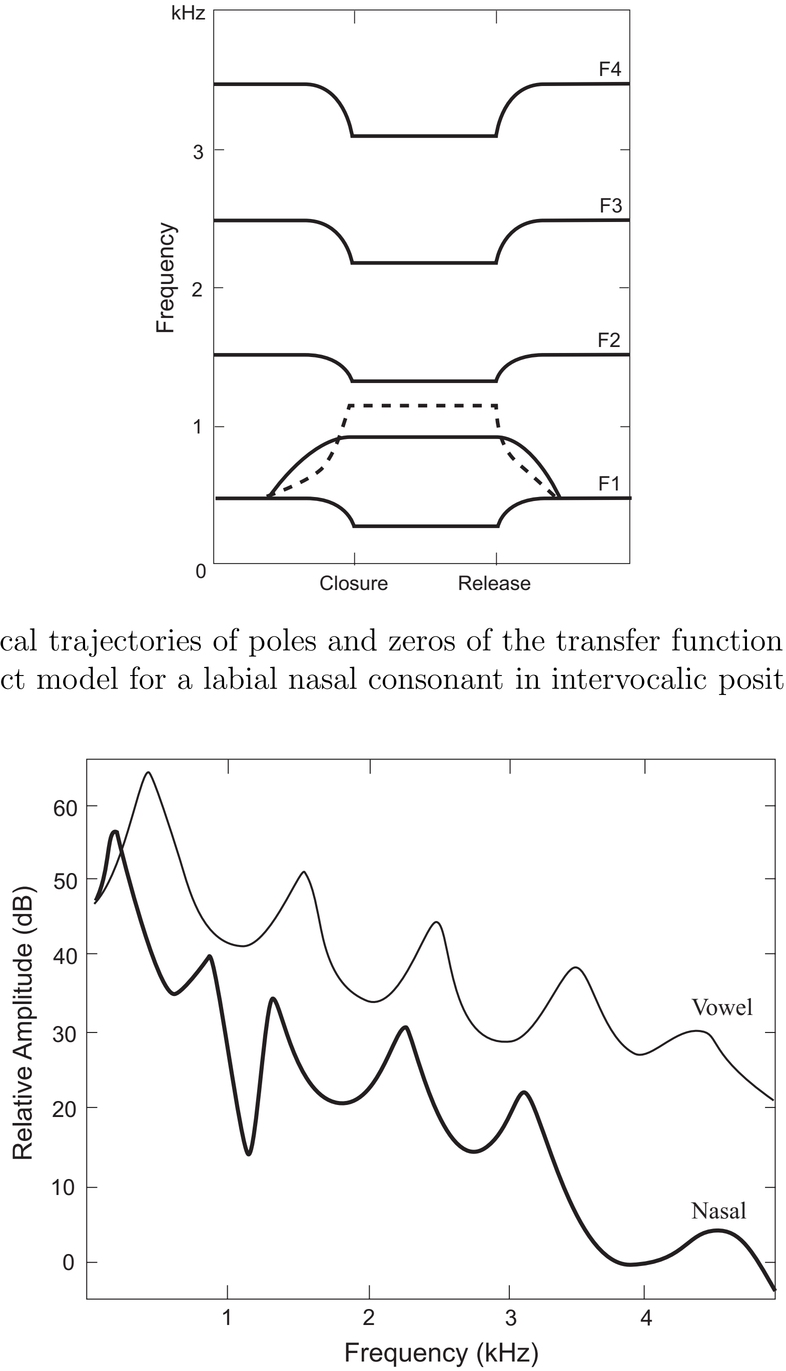
Nasal Passage

As discussed above, the low-frequency energy does not change significantly in the boundary from a vowel to a nasal, except that it may show a weak low-frequency prominence at 250–400Hz range [8]. On the other hand, the high-frequency range shows a notable difference. This change in the high-frequency range is caused by nasal pole-zero pairs introduced by the side branch between the velopharyngeal opening and the closure in the mouth.

The theoretical trajectories of poles and zeros derived from a vocal tract model for a labial nasal consonant are plotted in Figure 2-7(a) [50]. The estimated lowest zero frequency of a labial nasal is near 1000Hz, and because the nasal zero frequency is inversely proportion to the length of the side-branch, alveolar and velar nasals have zeros at higher frequencies, which is in the range of 1600–1900Hz. The introduction of a zero reduces the spectrum amplitude in the second and third formant frequency range. This effect can be observed in Figure 2-7(b), in which the spectrum slices in the middle of a nasal and its adjacent vowel are compared.

Liquids also accompany extreme types of formant frequency movements. A retroflex liquid is produced by creating a small cavity under the raised tongue, which creates additional zeros that causes significant drop of the second and third formant frequencies. On the other hand, a lateral is produced dividing the oral tract into two small branches with the tongue tip, and this change of vocal tract shape results in a heightened F3 frequency.

Lateral sounds do not always have slow transitions. For example, when it is used in a word-initial position such as in the word ‘let’ or ‘lion’, the tongue-tip makes a clear contact with the roof of the mouth, and in many cases, the release of the tongue creates an abrupt discontinuity in the speech signal. On the other extreme, when the /l/ sound is pronounced as a syllable-final consonant following a back vowel, such as in the word ‘ball’ or ‘small’, occasionally the tongue tip does not touch the roof of the mouth and no abrupt transition is made. Instead, a slight change in formant frequency can be observed due to backing of the tongue. Sproat and Fujimura [48]



(a) Theoretical trajectories of poles and zeros of the transfer function of the vocal tract model for a labial nasal consonant in intervocalic position

(b) Spectrum envelopes at the center of the nasal (*thick*) and in the vowel (*thin*)

Figure 2-7: Illustrations of the effect of nasal passage

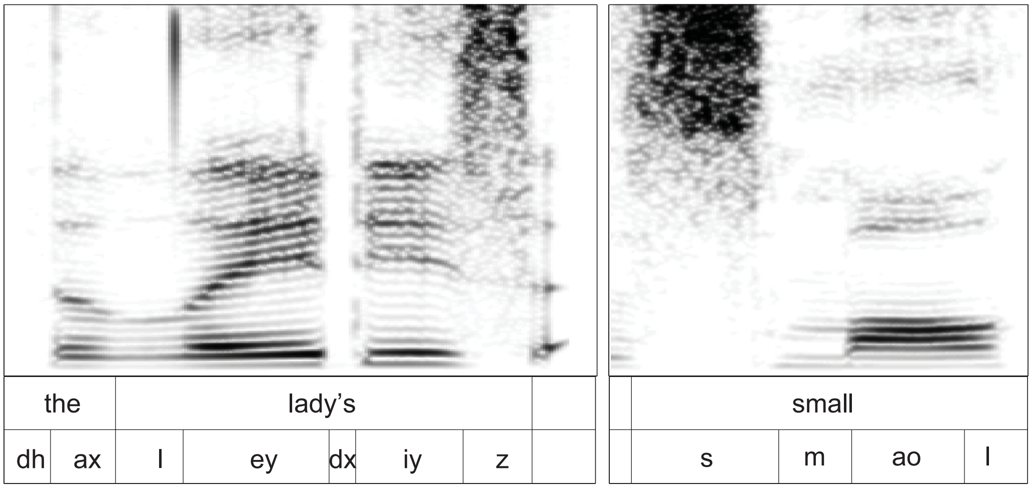


Figure 2-8: Examples of two extreme cases of /l/

explain this acoustic difference by an asynchrony between dorsal retraction and apical movement. Two of the extreme cases of the /l/ sound are shown in Figure 2-8. The /l/ sound in the syllable-initial position of the word ‘lady’ shows a clear diminishment in the spectral energy throughout all frequency range, whereas the /l/ following a back vowel /ɔ/ in the word ‘small’ does not show any significant spectral changes.

Turbulence Noise Source

There are generally two cases that create turbulence noise: stop or affricate burst and fricative.

A stop burst is preceded by a complete closure in the oral tract, which builds up the intraoral pressure and stops vocal fold vibration. When the closure is released, a rapid airflow is generated through the previous closure point, during which the cross-sectional area is increasing. This sudden burst of airflow causes a turbulence noise which excites the region between the constriction and the lips.

The burst noise can affect all frequency ranges, although the spectral shape may be affected by the shape of the front cavity, which depends on the place of articulation of the stop consonant and the frontness of the adjacent vowel [53]. A labial stop burst is the weakest among the three because there is no frontal cavity to be excited by the turbulence noise, a velar stop burst excites the frequency range near 2kHz due to the long frontal cavity length in front of the closure, and an alveolar stop burst excites higher frequency range because it has a shorter frontal cavity.

The burst of a stop consonant is produced with a rapid release of the stop closure and the burst noise takes place for a short time during the release—the burst noise of a stop consonant takes place for a short time during the release—about 5–20ms on average. On the other hand, a fricative is produced by making a narrow constriction, and so the duration of a fricative consonant is relatively longer than that of stop bursts [28]. The duration depends on individual phonemes and various phonetic contexts [33].

As was the case for stop consonants, the spectral shape and amplitude of fricative consonants depend on the place of articulation. While some alveolar fricatives such as /s/ and /z/ mostly affect the region higher than 3.5kHz, labial and dental fricatives such as /f/ and /θ/ affect the overall frequency range, although with less prominence in energy.

Some allophones of weak voiced fricatives are pronounced without turbulence noise. For example, the /v/ sounds in intervocalic position, such as in the word ‘every’, are sometimes pronounced without enough narrowing to make turbulence noise across the constriction, and is produced with characteristics similar to those of a glide /w/. Similarly, the voiced fricative /ð/ following a nasal, such as in the word sequence ‘in the’, is often produced without completely closing the velopharyngeal port during the /ð/ sound. Therefore, the intraoral pressure is not raised enough to make a frication noise, and the sound takes characteristics similar to those of a nasal /n/.

2.2 Distinctive Features

2.2.1 Introduction

Distinctive features are binary-valued characteristics of a sound that can distinguish one phoneme from another. The presence or absence of voicing, abrupt or transient onset of an obstruent consonant, and the realization of a consonant sound from a nasal tract or an oral tract are some of the examples of the correlates of distinctive features.

The concept of binary distinctive feature was suggested by Jakobson, Fant and Halle [25]. Distinctive feature theory is based on the assumption that the speech signal can be segmented in time, and each segment can be described with a set of discrete properties instead of by continuous-valued measures. Some observations assert that those features are of a discrete nature; for example, a study by Lulich *et al.* [37] shows that the discontinuity of the second formant frequency when it passes through the second subglottal resonance influences the perception of the backness in the vowel.

In addition, a perception experiment by Miller and Nicely [41] also provide evidence to the assumption that listeners focus on the acoustic evidence that distinguishes certain distinctive features. The experiment was performed by asking listeners to distinguish 16 different consonant sounds in nonsense syllables with varying degrees of noise levels and with different cut-off frequencies. The result of this experiment showed that in noisy or frequency-restricted cases the confusion occurred mostly across certain distinctive features, and that each of the distinctive features for consonants—voicing, nasality, continuant, and place of articulation—is relatively independent of the perception of other features.

Chomsky and Halle [10] suggested a universal set of distinctive features. Because most of the features are defined according to the articulatory gesture that humans make during speech, these features can be applied across different languages. Each bundle of the distinctive features can specify a speech sound. Stevens [50, 52] derived a subset of about twenty distinctive features from the universal feature set, so that it can contrast among all the sounds in English.

2.2.2 Two Classes of Distinctive Features

The set of distinctive features used in this thesis is based on the ones defined by Stevens [52]. Stevens adopted Ladefoged and Halle's [34] classification of features into two broad categories: articulator-free features and articulator-bound features. Articulator-free features are the ones that are not associated with any particular articulator, and articulator-bound features describe the active articulators and their movement.

Articulator-Free Features

An articulator-free feature does not relate to any specific articulator, but represents the general manner of articulation. For example, the feature [sonorant] is defined to be the sound produced with a vocal tract configuration that enables a free vibration of the vocal folds, and the feature [consonant] represents the sound pronounced with an abrupt constriction inside the vocal tract.

Stevens first defines three features that specify the broad classes of segments: consonants, vowels, and glides. Consonantal segments can be further classified with additional distinctive features: sonorant, continuant, and strident. A sonorant feature contrasts the sounds that are produced with spontaneous vibration of the vocal folds versus the ones with suppressed vibration or without vibration. A continuant feature distinguishes the speech sound produced with a complete closure inside the oral tract from the sound produced with a narrow constriction, which results in turbulence noise during the sound. A strident feature contrasts the continuant non-sonorant sounds based on the amplitude of the high-frequency. When the cavities and obstacles around the constriction are positioned in a way that the spectrum amplitude in the high-frequency region is higher than that of adjacent vowel, it is called to be strident.

When a speech sound has the characteristic defined by the feature, the feature is represented with a + sign in front of it, and when a speech sound lacks the characteristic, the feature is represented with a − sign. The values of the articulator free features for some speech sounds are shown in Table 2.2. Note that not all the features

	æ, ɪ	w, j	p, d	z, ʃ	ð, f	m, n
Consonant			+	+	+	+
Vowel	+					
Glide		+				
Sonorant			−	−	−	+
Continuant			−	+	+	−
Strident				+	−	

Table 2.2: Articulator-free features for some speech sounds in English

are assigned a value. Redundant features or the ones that cannot be defined in certain contexts are not marked with a sign. For example, because the consonant, vowel and glide features are mutually exclusive, only +values are marked. Similarly, because all the [−consonant] sounds are pronounced with [+sonorant] feature, that is, without significant increase in the intraoral pressure, the sonorant feature is marked only for [+consonant] sounds. On the other hand, the strident feature is marked only for [−sonorant, +continuant] sounds because the feature is defined for fricatives alone.

Articulator-Bound Features

While articulator-free features describe the manner of articulation of a speech sound, articulator-bound features describe the specific movement of articulators controlled in the production of the sound.

Some of the articulator-bound features specify the articulators in use. For example, the sound /p/ is pronounced with a complete closure made with lips, therefore, the /p/ sound has the feature [+lips]. There are three place-of-articulation features: lips, tongue body and tongue blade. These features discriminate among the sounds /b/, /g/ and /d/, respectively. Other articulator-bound features describe the movement of certain articulators. For example, high and low features can differentiate between different heights of the tongue body. Table 2.3 gives a complete list of the articulator-bound distinctive features used in this work.

A list of the values for the articulator-free and articulator-bound features for some speech sounds are tabulated in Table 2.4. As was the case with articulator-

Articulator	Feature	Description
Vocal Folds	stiff vocal folds	Voicing vs. unvoicing
Glottis	spread glottis	Introduction of aspiration
Pharynx	advanced tongue root	Tense vs. lax
Soft Palate	nasal	Nasal vs. oral
Tongue Body	body	Place of articulation
	high	Tongue body position
	low	Tongue body position
	back	Tongue body position
Tongue Blade	blade	Place of articulation
	rhotic	Retroflexed tongue tip
	lateral	Oral tract is separated into two paths
	anterior	Constriction nearer to the lips
Lips	lips	Place of articulation
	round	Lips are rounded

Table 2.3: List of articulator-bound features that distinguishes English sounds

free features, not all the articulator-bound features need to be assigned a value to distinguish a speech sound. The features that are not available or that cannot be defined in the context of other features are not specified. For example, the place of articulation features are available only for [−vowel] sounds because a vowel is pronounced without a constriction in the oral tract. Similarly, because the anterior feature describes the location of the tongue blade, the feature is specified only when the sound has [+blade] feature.

In addition, the features that are not distinctive are not assigned a value as well. For example, the /b/ sound in the word ‘bee’ is pronounced with fronted tongue body, while the same sound in the word ‘boo’ is pronounced with the tongue position in the back. Because the tongue body positions are not distinctive in the production of a consonant sound, the features that specify the tongue body position—high, low, back—are not given a value for [+consonant] sounds. The feature [round] is another example of such features. While the sound /ɑ/ and /ɔ/ are distinguished by the rounding of the lips, the sound /i/ does not have such counterpart. Because in English, all the front vowels are [−round], and the feature [round] is distinctive only

	i	ɪ	ɑ	ɔ	w	p	g	z	ʃ	θ	m
Consonant						+	+	+	+	+	+
Vowel	+	+	+	+							
Glide					+						
Sonorant						−	−	−	−	−	+
Continuant						−	−	+	+	+	−
Strident								+	+	−	
High	+	+	−	−	+						
Low	−	−	+	+	−						
Back	−	−	+	+	+						
Round			−	+	+						
Adv. Tongue Root	+	−			+						
Lips					+	+					+
Blade								+	+	+	
Body							+				
Anterior								+	−	+	
Stiff Vocal Folds						+	−	−	+	+	

Table 2.4: Feature bundle representation for some speech sounds in English

for back vowels, the round features are specified only when the segments have [+back] property.

Due to these relationships between features, each speech sound can be distinguished from others by specifying five to seven features, instead of estimating all the values of the twenty one distinctive features. The distinctive feature bundle representation of the complete set of English phonemes is shown in Appendix A.

2.2.3 Landmarks and Distinctive Features

Landmarks and the distinctive features are closely related in that the landmarks can determine some of the articulator-free features, and also in that the landmark types can restrict the type of articulator-bound features that can be estimated near the landmark position to identify the speech sound.

+g-Landmark	
+consonant −sonorant *or* Silence	+sonorant

+b-Landmark	
Silence	+consonant
	−sonorant

+s-Landmark	
+consonant	−consonant
+sonorant	+sonorant
−continuant	

Table 2.5: Articulator-free features that can be identified from individual landmark types

Landmarks and Articulator-Free Features

It can be understood from the definition that most of the consonant landmarks correspond to the time-points where articulator-free features change from one sign to another. The articulator-free features that can be determined from individual landmark types are shown in Table 2.5. The distinctive features written on the left of the vertical line are the expected feature values of the sound that precedes the landmark, and the ones on the right are the expected feature values of the sound that follows the landmark.

The g-landmarks are defined as the onset or offset of spontaneous vocal fold vibration and the feature [sonorant] is defined to be the speech sound produced with freely vibrating vocal folds. Therefore, g-landmarks indicate the locations where the sign of sonorant feature changes.

A b-landmark, on the other hand, does not correspond to the change of a particular distinctive feature, because a b-landmark marks the boundary between turbulence noise and silence (or closure of a stop consonant), and the features of a silent region re not defined. Although b-landmark does not indicate the change of a feature, it does provide information about some features of its adjacent segment. The turbulence noise is made with a constriction in the mouth which increases the intraoral pressure. Therefore, the speech sound adjacent to a b-landmark is expected to be [+consonant, −sonorant].

An s-landmark marks an abrupt change during a voiced region. Therefore, an s-landmark not only asserts that the adjacent segments have [+sonorant] feature, but it also indicate the change of a [consonant] feature value at the landmark position. Since the sonorant consonant is made with a complete closure in the oral tract, the consonantal sound has [+consonant, −continuant] feature as well.

Because vowel and glide landmarks are not located at the boundary but at the time-point where the acoustic characteristic is the most prominent, these landmarks do not indicate the transition, but the existence of a segment with [+vowel] or [+glide] feature at the position.

Although the landmarks correspond to most of the articulator-free features, it should be understood that the landmark types do not completely describe the articulator-free features of the adjacent segments. For example, a strident feature cannot be determined from the landmark information, and the continuant feature can be decided only in some particular cases.

As was stated before, all the sonorant consonants are non-continuant in English, and so the continuant feature need to be distinguished only within obstruent consonants. Because a non-continuant obstruent is made with a complete closure inside the oral tract, there must be a region of complete closure before the release of a burst noise, and this acoustic change is marked with a +b landmark. Continuant consonants, however, are sometimes produced with a pause before the frication noise, during which the intraoral pressure builds up high enough to make turbulence noise through the oral constriction. Therefore, the continuant features need to be distinguished only for the sounds that accompany +b landmarks. Otherwise, it can be assumed that the obstruent consonants are [−continuant].

As for continuant features, landmark pairs are more effective than individual landmark types. Table 2.6 lists all four types of landmark pairs that determine the bound-

Landmark Pair	Expected Features
(−g, +g)	+consonant −sonorant +continuant *or* Silence
(−g, −b)	+consonant −sonorant +continuant
(+b, +g)	+consonant −sonorant (?)continuant
(+b, −b)	+consonant −sonorant (?)continuant

Table 2.6: Landmark pairs that specify obstruent consonants, and the articulator-free features expected from the landmark types

aries of obstruent consonants, and the distinctive features of the segment expected from the surrounding landmark types are specified for each landmark pair. The question mark (?) is used when the feature cannot be determined from the landmark type only. Note that only the pairs that starts with +b landmark has ambiguous continuant feature. Most of the continuant feature can be identified by the distance between the two landmarks at the boundaries, because the burst release of a stop consonant is usually abrupt and short.

Strident features are defined only for [+continuant] sounds. Therefore, this feature can be estimated after the continuant feature value is known.

Landmarks and Articulator-Bound Features

While landmarks are closely related to articulator-free features, they do not exactly correspond to any of the articulator bound features. However, because the types of articulator-bound features that are available and distinctive are limited by the values of articulator-free features, the landmark types can provide information about the articulator-bound features that can be evaluated and the locations where the acoustic

+consonant ↓ sonorant continuant	−consonant ↓ high low back	−vowel ↓ lips blade body
+sonorant ↓ nasal	−sonorant ↓ stiff vocal folds	+continuant ↓ strident
+back ↓ round	−low ↓ adv. tongue root	+blade ↓ anterior

Table 2.7: List of features that are distinctive under certain contexts

cues that correspond to each feature can be estimated. Table 2.7 gives some examples of the features that are distinctive in certain contexts. For example, the features for tongue body position—high, low and back—are distinctive only for [−consonant] sounds, whereas the place of articulation features—lips, blade and body—are distinctive only for [−vowel] sounds.

Due to the relationship between different landmarks, the number of articulator-bound features that are needed to identify a speech sound does not exceed more than five in most cases. For example, to identify a [+vowel] sound in English, the values of at most five features need to be determined: high, low, back, round, and advanced tongue root. Moreover, not all five features are always required because round feature is distinctive only for a back vowel, and advanced tongue root is distinctive only for a non-low vowel. A sonorant consonant, which is typically located by s-landmarks, can be identified by specifying no more than two articulator-bound features, nasal and the place of articulation.

2.3 Summary

In this section, three types of consonant landmarks were defined and their articulatory and acoustic characteristics have been examined. In addition, a set of distinctive features that is used in this thesis has been defined, and the relationships between landmarks and distinctive features have been discussed.

It is speculated that the landmarks can be an effective starting point for analyzing the speech signal, not only because they point to the information-rich locations, but also because the landmark types provide information about the broad classification of the signal. This information provides two advantages to speech analysis system. It indicates the distributional characteristics of acoustic cues that can be estimated near the landmark position, and it can also highly reduce the number of the distinctive features that can or should be determined to identify speech sounds.

Chapter 3

Landmark-based Speech Recognition

3.1 Introduction

As was discussed in the previous chapters, the landmarks can be an adequate starting point for a speech analysis. Stevens [52] has proposed a knowledge-based speech recognition system which utilizes the knowledge about the acoustic landmarks and distinctive features. The recognition system aims to retrieve the word sequence from an utterance of a sentence, by first locating the acoustic landmarks and then estimating the sequence of distinctive feature bundles at the detected landmark positions.

A simple block diagram of the model is in Figure 3-1. First, the landmarks in a given speech signal are detected based on the acoustic cues described in Section 2.1. The landmark detection include not only locating the time-points that are information-rich, but also identifying the types of the landmarks, which provide in-

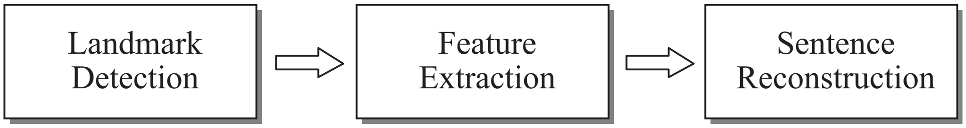


Figure 3-1: A block diagram of a landmark-based speech recognition system

formation about the broad-class of the signal adjacent to it. Then, the acoustic cues that are appropriate for the detected landmark types are measured near the landmarks. Based on the acoustic measurements, the values of distinctive features can be estimated at each of the landmark locations. As was explained in Section 2.2, the landmark types highly restrict the number of distinctive features that need to be evaluated, and the coarse syllable structure of the utterance can also be predicted from the landmark sequence. After enough information about the feature bundles are collected, it can be used to access lexicon and the original text can be reconstructed from it.

3.2 Landmark Detection Process

The first part of the speech recognition system, the landmark detection process, locates acoustic landmarks and classifies them according to their characteristics. By knowing the landmark locations, the system can focus on certain information-rich locations instead of distributing the attention uniformly throughout the signal, and apply different method of analysis to different distribution of the acoustic cues such as abrupt discontinuities and steady-state signal.

3.2.1 Segment-based Approach

This approach is different from typical frame-based models, which sample appropriate sets of cues at uniformly separated fixed-width windows called frames and reconstruct the original sentence by comparing each frame with the distribution of cues for different phonemes using method such as hidden Markov models (HMMs) [43, 4] or graphical models [3, 61] along with reasonable language models, such as n-gram models [2].

One of the limitations of the frame-based speech recognition system is that it treats each frame with equal importance, and that it does not make use of the dependency between adjacent frames. However, the sounds of phonemes are highly variable depending on adjacent sounds, while the frame-based model treats each frame independent from one another. In addition, perceptual evidence shows that the acoustic cues in the vicinity of the transition between different speech sounds play a critical role in the perception of a syllable, compared to the stationary part of the signal [13].

Landmark-based approach can provide a way to overcome this problem by first locating the points where transitional periods exist, and then applying appropriate measure to analyze the information. Some statistical recognition systems have incorporated such transitional information. For example, Ghitza and Sondhi [14] created a HMM system that recognizes the speech signal using a diphone model instead of recognizing individual phonemes separately, so that the variation due to adjacent phonetic context can be taken into account.

While such systems are only interested in incorporating the information of the phonetic transitions into traditional speech recognition system, the SUMMIT system [59, 58] locates the acoustic discontinuities based on a spectral distance measure [15] as a starting point of a segment-based speech recognition system. This system only focuses on the position of the segment boundaries, not on the acoustic nature of the discontinuities.

On the other hand, there are some knowledge-based recognition systems that segment a speech signal based on the broad-class classification. Juneja and Espy-Wilson [29, 30], and Jansen and Niyogi [26] classify the speech into manner classes, such as vowel, sonorant, fricative, stop, and silence, based on acoustic phonetic parameters. However, the goal of the segmentation is different from landmark detection, since the broad-class segmentation algorithm aims to be extended to a phonetic classification process by incorporating additional binary distinctive features that distinguish speech sounds [12, 42], instead of providing focus points to extract acoustic cues of phonetic importance.

Hasegawa-Johnson *et al.* [20, 21] introduce a landmark-based speech recognition system as well. This system uses a computationally-intensive landmark detection process based on a large number of acoustic observations extracted every 5ms. The observations include energy, spectral tilt, Mel-frequency cepstrum coefficients (MFCCs), formant frequencies and amplitudes, and other acoustic-phonetic cues. Instead of being the focus of attention, the detected landmarks are used as one of the cues for speech recognition, along with broad-class segmentations and distinctive feature classifications.

On the other hand, Liu [35] developed a consonant landmark detection process that aims to be the focus point for the subsequent distinctive feature processing [52], and the three types of consonant landmark defined by Liu not only enable us to locate the time-points of acoustic discontinuities, but also estimate the value of articulator-free features of the adjacent regions, as was explained in Section 2.1. The vowel and glide landmarks also locate the places where the acoustic information is the most prominent even though the vowels and glides do not accompany acoustic discontinuities.

3.2.2 Current Status

Automatic detection algorithms for all three classes of acoustic landmarks (i.e., consonants, vowels and glides) have been developed to some extent. A consonant landmark detector was created by Liu [36], which detects the locations of acoustic discontinuities, and classifies them into three types of landmarks as described in Section 2.1. An automatic vowel landmark detection algorithm was built by Howitt [23] and has been improved by Slifka [47]. A primitive algorithm for distinguishing glides from vowels and liquids was proposed by Sun [55] and an algorithm that automatically locates glide landmarks in a speech signal still need to be developed.

The three classes of landmarks are closely related to each other. For example, the vowels, which is a sonorant sound, must come after a +g landmark and before a −g landmark, and the glides and liquids are always adjacent to a vowel sound. Therefore, when the three landmark detection algorithms are integrated together, a better detection performance is expected.

3.3 Feature Estimation Process

After the landmarks are located and their types are identified, the distinctive features that describe the speech sound are estimated near each landmark position, or if available, near an appropriate group of landmarks where the acoustic cues of the same feature can be found.

3.3.1 Distinctive Features

Many speech recognition systems rely on the identification of phonemes based on the statistical distribution of a set of acoustic cues. However, because the acoustic realization of phonemes are highly variable depending not only on the context but also on speaker's gender, dialect and other characteristics, the performance of automatic phoneme identification is not high. To compensate for this problem, the automatic speech recognizers include possible allophones of the phonemes and use higher-level information such as diphone or triphone model and word or language models, but this requires a large amount of training due to a large number of tokens to be trained, and the language models may not be appropriate when there is a word that is not represented in the lexicon.

On the other hand, distinctive feature representation is less sensitive to such problems, because the acoustic realization of the sound is usually limited to a couple of features and the value of each feature can be estimated relatively independent from other features. These variation can be easily represented based on the distinctive feature representations—e.g., palatalization of the second /d/ sound in the word sequence "did you" can be represented by the change in the anterior feature from [+anterior] to [−anterior], and nasalization of /ð/ sound in "in the" can be represented by the change in sonorant feature from [−sonorant] to [+sonorant]. Thus, it is possible to construct a knowledge-based pronunciation model utilizing the distinctive feature in order to compensate for the phonetic variation.

Therefore, our system represents the lexical items in terms of the sequence of feature bundles, instead of a sequence of phonemes, and the recognition of speech sounds are performed by estimating the values of distinctive features. The articulator-free features can be estimated based on the landmark types, and once the articulator-free features are identified, only at most five articulator-bound features need to be estimated for the feature bundle to be able to identify a speech sound.

Some other knowledge-based speech recognition systems adopt the distinctive feature representation as well. Bitar and Espy-Wilson [5] use a decision tree based on the distinctive features to identify the sound of a segment, and Hasegawa-Johnson *et al.* [20] also categorize the utterance using a number of place classifiers that identifies the features palatal, labial, voiced, high, front, etc.

In our system, not only the features that distinguish phonemes are estimated, but the features that accounts for supra-segmental events, such as word boundary, syllable affiliation of consonants, and lexical stresses are also considered. These features does not distinguish a phoneme from another, but they can be useful in lexical access because the knowledge of lexical stress reduces the size of possible word candidates significantly even when only a partial information is known about the signal [24], and the word boundary and syllable affiliation information can resolve uncertain situations that cannot be distinguished by phonemic information alone, such as the distinction between “lay style” and “lace tile” or between “baby” and “bay bee”.

3.3.2 Current Status

Most of the articulator-free features can be identified from the landmark types without additional measure, and feature estimation modules for a large number of articulator-bound features and prosodic features—e.g., nasality, place of articulation, stridency, tongue position for vowel, voicing of obstruent consonant, lexical stress, etc.—have been developed as well.

Most of these feature detection modules assume that the places to measure the acoustic cues are already provided specifically. For example, a nasal detection module developed by Chen [9] assumes that the onset and offset boundaries of a sonorant sound is known, and the automatic detection algorithm that identifies the place of articulation for stop consonant [54] is based on the fact that the stop consonants are already identified, and the related landmark positions—i.e., closure, burst release, and the onset of the following vowel—are already known.

The locations that are needed for these feature detection modules correspond to the landmark positions. The landmark detectors that have been developed only identify individual landmark position of certain type, but they are not responsible for the grouping of the detected landmarks that represent the acoustic events of the same speech sound. For example, the three time-points that should be located in order to estimate the place of articulation features of a stop consonant correspond to (−g, +b, +g) landmarks. However, if the detected landmark sequence turns out to be (−g, −b, +b, +g) or (−g, +b, −b, +g), it is not clear which of the landmarks should be used as inputs for the feature detection module. Therefore, an intermediate procedure that groups the related landmarks according to landmark types needs to be developed.

3.4 Sentence Reconstruction Process

After all the features are estimated and grouped in appropriate feature bundles, this information can be directly used to access the lexicon, which is also represented in the form of sequences of feature bundles.

3.4.1 Difficulties in Lexical Access

Lexical access and sentence recognition based on the distinctive feature representation are different from those based on traditional statistical-feature representation, because the feature values are binary, and the number of features that are needed to specify a speech sound is different from one sound to another, and so not all the features are identified during the feature extraction process.

In addition, the distinctive features that are estimated from the realized acoustic signal may be different from the lexical representation of distinctive features, due to the overlapping of gestures and some other effects. For example, the place of articulation of the /n/ sound in the word sequence “in my”, may be changed to

[+lips] instead of [+blade] due to the adjacent /m/ sound, the /t/ sound in the word 'habitual' may be pronounced with either [+anterior] or [−anterior]. However, because most of the features remain the same and only some features changes within limited contexts, appropriate linguistic rules may be applied to compensate for these effects.

The incomplete performance of the landmark detection also introduces problems to the lexical access. When a landmark is not detected, all the features that are related to the landmark cannot be identified, and when a landmark is falsely detected, an additional feature bundle with redundant information might be included in the input sequence, and the lexical access system has to figure out the locations where the feature bundles are lost or inserted.

3.4.2 Current Status

The process of sentence reconstruction is not fully developed yet, but there have been several studies concerned with the difficulties due to the binary distinctive feature representation. Zhang [57] proposed a simple template matching algorithm which allows some modification according to a set of linguistic rules represented in the distinctive features, and Maldonado [38] improved the speed of Zhang's algorithm using the hierarchical characteristics of distinctive features. Kim [32] has developed an algorithm that compensates for falsely detected segments and other variants that can be introduced with imperfect performance of the landmark detection process.

Each of these matching algorithms tackles some aspect of the problems, but they are based on the assumption that most of the features are specified, with a few regular errors allowed by linguistic rules. However, as the experiment of Huttenlocher and Zue [24] shows, the lexical access can be performed without knowing all the distinctive features, and it can still reduce the size of the cohort significantly. Therefore, it might be more efficient to access the lexicon in the earlier stage of the speech recognition, than to wait for all the features to be estimated.

3.5 Research Scope

This thesis focuses on detection of the consonant landmarks, because the consonant landmarks represent the boundaries of the speech sounds. The locations of other classes of landmarks can be hypothesized by the consonant landmarks because vowels are always restricted to be between +g and −g landmarks, and glides are always next to vowels.

In Chapter 4, an overview of the previous landmark detection model proposed by Liu [36] is given, and possible improvement of the detection algorithm is discussed. Chapter 5 demonstrates a probabilistic algorithm that detects landmarks and classifies their types individually without considering the relationship among different landmark types.

A bigram model that restricts possible landmark sequences and an algorithm to determine the most likely landmark sequence based on the bigram model is proposed in Chapter 6. Some of the possible contexts that create ambiguous recognition of landmarks are discussed as well. Chapter 7 provides a method to represent multiple possibilities for the ambiguously detected regions.

Chapter 4

Preliminaries to the Implementation of Landmark Detector

4.1 Desired Properties of the Landmark Detector

Liu [35] has developed an algorithm that detects the acoustic landmarks by utilizing linguistic knowledge. Liu's approach is based on a deterministic algorithm that uses a series of decision processes, each of which represents a piece of acoustic knowledge about speech signal. The algorithm is briefly reviewed in Appendix B.

However, the algorithm used strict thresholds in each decision module which did not allow unclearly realized landmarks to be detected and not adequately account for individual variation between different speakers, and the speech-knowledge was built in the structure of the decision tree so that additional speech knowledge cannot be easily integrated into the system without changing the whole system. In this section, such disadvantages are reviewed and the desired properties for a new consonant landmark detector are discussed.

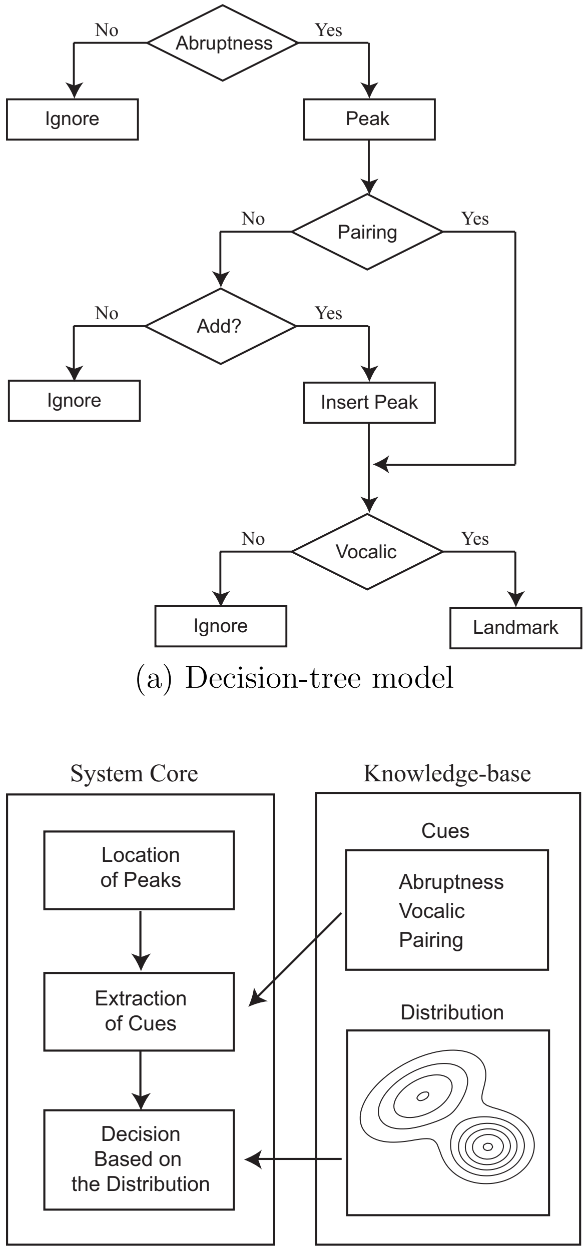
4.1.1 Separation of Knowledge-base

By making use of linguistic knowledge and its acoustic correlates, Liu's landmark detection algorithm achieves a high performance of almost 80% detection rate without any supervised training. However, the speech knowledge is embedded in the system structure itself, which means that the whole system should be modified whenever a change occurs in the criteria. To avoid this problem and to make the system more flexible to change, it is desirable to separate the knowledge-base from the system's core structure.

In Liu's consonant landmark detector, each criterion has been embedded in the system as a branch of a decision tree. These elements can be separated from the system by representing each criterion as a cue, and the decision algorithm as distribution of the cues, as is illustrated in Figure 4-1. To incorporate additional knowledge to the decision-tree model in Figure (a), one should change the connections among the decision modules completely. However, the revised approach introduced in this thesis, of which the schematic diagram is shown in Figure (b), separates the knowledge base from the system's core structure by considering each decision process as a cue, and represent the connections between decision modules as the distribution of the cues. In this cue-distribution model, the update of the speech knowledge can be done without affecting the whole system, just by adding additional cues relevant to the speech knowledge.

4.1.2 Reduction of Deletions

Another weak point of Liu's algorithm is that it uses hard thresholds in the peak picking process and in the decision processes. Because the operation of the landmark detection algorithm is based solely on the maximal transitions in energies in different frequency bands that are detected, when some landmarks fail to be recognized as transitions due to their less prominent properties, they will never be processed in the subsequent steps. Similarly, when some of the energy transitions are rejected in one of the knowledge-based decision steps, they will not be retained and the subsequent



(a) Decision-tree model

(b) Cue-distribution model

Figure 4-1: Comparison between (a) decision-tree based approach and (b) cue distribution approach

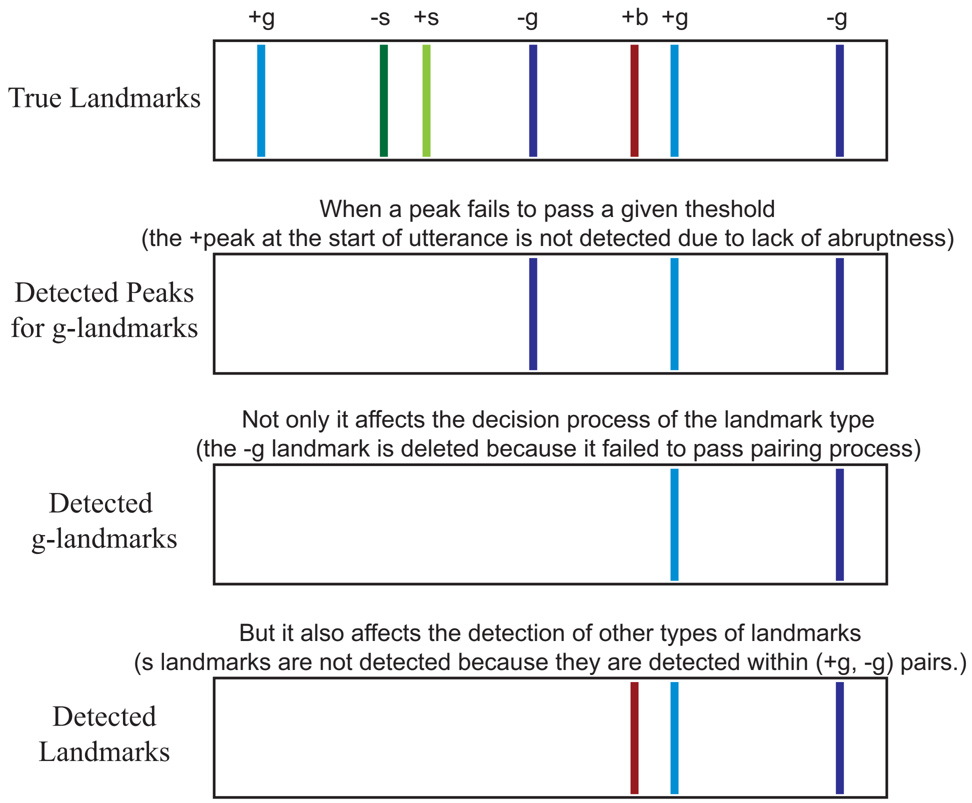

+g
-s
+s
-g
+b
+g
-g
True Landmarks
When a peak fails to pass a given theshold
(the +peak at the start of utterance is not detected due to lack of abruptness)
Detected Peaks for g-landmarks
Not only it affects the decision process of the landmark type
(the -g landmark is deleted because it failed to pass pairing process)
Detected g-landmarks
But it also affects the detection of other types of landmarks
(s landmarks are not detected because they are detected within (+g, -g) pairs.)
Detected Landmarks

Figure 4-2: An example of error propagation. When a +peak for a g-landmark fails to pass the threshold, the pairing criterion will delete its corresponding -g landmark as well, which also results in the deletion of s-landmarks between the deleted g-landmark pair.

decisions will be affected by the deletions.

The propagation of deletion not only affects one type of landmark, but also affects other types of landmarks. An example of such a domino effect is illustrated in Figure 4-2. In Liu's landmark detection algorithm, g-landmarks are first detected with the criterion that every +g landmark most have a pairing −g landmark, and then s-landmarks are located only between each (+g, −g) landmark pair. Therefore, if one of the +g landmarks is not detected in the peak-picking process (detected g-landmark peaks are illustrated in the second tier as blue lines), the corresponding −g landmark cannot be located due to the strict g-landmark pairing restriction (the detected g-landmarks are illustrated in the third tier). Therefore, the region between deleted g-landmarks is not searched for the existence of s-landmarks, and so the s-landmarks within the region will not be detected. As a result, four out of the seven landmarks may be deleted due to the omission of a single landmark peak.

Deletion of landmarks causes more problems than insertion, because the landmarks locate the time-points that need to be further investigated for additional cues. Given a system that relies upon the detected landmarks, an omission of a landmark means that no cues will be measured at the landmark point and that only higher-level processes such as phonotactics and lexical access will be able to retrieve the lost information. On the other hand, an insertion of a false landmark only results in estimation of information at the false alarm position where no relevant articulatory events occurs, and this insertion can be easily removed after verifying that the extracted information does not indicate important events.

Therefore, in the newly developed process, the thresholds are lowered in the peak-picking process to avoid losing true landmarks, despite the fact that it will introduce a large number of false alarms.

4.1.3 Probabilistic Process

Because the lowering of threshold values increases the number of insertion errors, a measure of likelihood is introduced to compensate for the side-effect. The probability calculation is merely an extension of the conversion from decision-tree approach to cue distribution approach, because the probability of each landmark candidate can be calculated from the distribution of the cues.

The probability measure of the landmark candidates can be used to distinguish reliable landmarks from false alarms. In addition, it can also be utilized to locate ambiguous landmark candidates and to extract additional information from them by postulating possible phonetic contexts in which the landmarks may be produced with less prominence.

4.2 Data Preparation

4.2.1 TIMIT Database

Classification of Speakers

The algorithm for automatic detection of landmarks is trained and tested on the TIMIT database [60]. This database consists of 6,300 utterances recorded by 630 speakers who are categorized into 8 different dialect regions based on geographical area. The number of male speakers outnumbers that of female speakers, but there are at least ten speakers per dialect region.

Because the landmarks correlate to the movement of speech organs, it can be hypothesized that the landmarks are robust to different genders, dialects and languages. Since each speaker class determined by gender and dialect contains at least 100 utterances or approximately 2,000 expected landmarks, it is possible to reliably estimate the performance of the algorithm for different dialects and genders on this database.

Recorded Texts

Ten utterances are recorded by each speaker. Two sentences are recorded by all the speakers to compare the effect of different dialect. Five of the utterances are from a set of 450 phonetically-compact sentences, which are designed to include most of the phone pairs and other phonetically interesting contexts. Each phonetically-compact sentence is recorded by seven speakers. The other three utterances recorded by each speaker are drawn from a set of 1,890 phonetically-diverse sentences, which are usually longer and contain various allophonic contexts.

The wide coverage of phone pairs is an advantage of using the TIMIT database because a landmark denotes the place where there is a transition from one segment to another. Thus, it is of some importance that the broad sampling of contexts allows us to observe how allophonic variations affect the landmarks in a relatively comprehensive manner. The two universally recorded sentences are not included in

Symbols	Description
dx	/t/ flap or /d/ flap
nx	nasal flap
q	glottal stop or irregular pitch periods
hv	voiced /h/
ux	fronted /u/
ax-h	devoiced schwa
el	syllabic /l/
en, em, eng	syllabic nasals

Table 4.1: Allophones used for the phonetic transcription in TIMIT database

the development and testing of the landmark detection algorithm, because that would overly emphasize only a small set of phonetic contexts.

Recording Quality and Phonetic Transcription

The utterances in the TIMIT database were read as isolated sentences, and recorded in a quiet environment at a 16kHz sampling rate. Each utterance is handlabeled with phones including certain types of allophones.

Some of the phones are defined ambiguously, which may cause a problem in automatic mapping of expected landmarks. For example, the [q] symbol is marked when there is a glottal stop, which is considered an offset of glottal vibration, but it also labels any occurrence of irregular pitch periods, which is a continuation of glottal vibration. A symbol for syllabic /l/ exists, but the use of this symbol mostly depends on the lexical stress pattern of the word rather than on the actual realization of the phones in the spoken signal. Therefore, in most cases, it is not possible to distinguish non-abrupt /l/ from abrupt /l/ from the labels alone. The complete set of allophones used in labeling the TIMIT database and their descriptions are listed in Table 4.1.

4.2.2 Predicting Landmarks from Phonetic Transcription

It is estimated that overall the TIMIT database contains about 100,000 landmarks. Therefore, handlabeling all the landmarks of the TIMIT database would be a timeconsuming piece of work. However, since the consonant landmarks mark the places where there are abrupt changes in the spectrogram, the landmarks corresponds to a boundary between a pair of phones. As a consequence of this property, the expected landmarks can be automatically determined from the phonetic transcriptions.

The complete table of phone-to-landmark mapping is shown in Table 4.2. Because the landmarks generally do not depend on the articulator-bound features, the mapping table is represented by broad classes of the phones for the sake of compactness. The broad class assignment of each TIMIT symbol is listed in Table 4.2(a). For each pair of adjacent segments, the landmark type that is typically expected to be present at the boundary is tabulated in Table 4.2(b). The blank cells correspond to the contexts where acoustic discontinuities are not likely to be found.

Due to ambiguity in the phonetic transcription, this mapping algorithm does not predict all the landmarks that are actually realized. For example, /l/'s at syllable-final position are sometimes realized without abrupt change at the vowel-sonorant boundary, but the landmark mapping algorithm predicts a −s landmark at the boundary because TIMIT's phonetic transcription does not distinguish between abrupt and non-abrupt /l/'s.

The symbol [q] can represent both glottal stops and irregular pitch periods, but in the mapping algorithm, it is assumed to be irregular pitch periods. Most of the glottal stops are at postvocalic positions followed by an obstruent consonant, and in those contexts, the expected sequences of landmarks are the same even if [q] is considered as a vocalic region. This assumption may change the positions of the landmarks around the glottal stops, but considering that the glottal stops are relatively short in duration, the difference will be small.

Although the flaps are variants of the stop consonants /t/ and /d/, they also have similarities to a sonorant consonant. The flaps are produced with a rapid tap of the tongue against the roof of the mouth instead of a complete closure and release. As a result, intraoral pressure does not build up, and air continues to flow through the vocal tract. Thus, the voice source is not turned off completely during the flap sound, and no burst noise can be observed in the spectrogram. Because these acoustic cues

Category	Symbols used in TIMIT	Note
VOC	ae, aa, ah, ey, eh, ow, ao, uw, uh, ih, iy, ix, axr, ax-h, aw, ay, oy, ux, er, ax w, y, r	Vowels Glides
SON	m, n, ng, em, en, eng, nx l, el	Nasals Liquids
FLP	dx	Flaps
IPP	q	Glottal stops
HVO	hv	Voiced /h/
FRI	s, sh, f, th, z, zh, v, dh jh, ch b, d, g, p, t, k hh	Fricatives Affricates Stops Unvoiced /h/
SIL	bcl, dcl, gcl, pcl, tcl, kcl epi, pau, h#	Closures Silences

(a) Classification of TIMIT segments. The classification is mainly based on the articulator-free features of the segments. These classes are used in Table (b).

		Following Symbol						
		VOC	SON	FLP	IPP	HVO	FRI	SIL
Previous Symbol	VOC		−s	−s			−g	−g
	SON	+s		−s	+s		−g	−g
	FLP	+s	+s		+s		−g	−g
	IPP		−s	−s			−g	−g
	HVO						−g	−g
	FRI	+g	+g	+g	+g	+g		−b
	SIL	+g	+g	+g	+g	+g	+b	

(b) Mapping from a pair of segment categories to a consonant landmark

Table 4.2: Mapping tables to convert TIMIT phonetic transcriptions into expected consonant landmarks

are close to those of sonorant consonants, flaps are mapped to s-landmarks.

An abrupt change can be observed between a nasal and a flap, because there is a substantial dip in energy for the flap sound and the formants above F1 mostly fade out. These changes are similar to those seen in a sonorant consonant. Therefore, the transition between a nasal and a flap is mapped to an s-landmark as well.

Chapter 5

Detection of Individual Landmark Candidates

5.1 Detection of Spectral Change

The first part of the landmark detection task consists of finding acoustic discontinuities in the speech signal. Figure 5-1 shows a rough block diagram of this process. First, the energies of six different frequency bands are calculated from the broadband spectrogram of the signal, and the abrupt changes in the amplitude of each band-energy are located with a two-pass algorithm. Each block of the diagram will be explained further in the following sections.

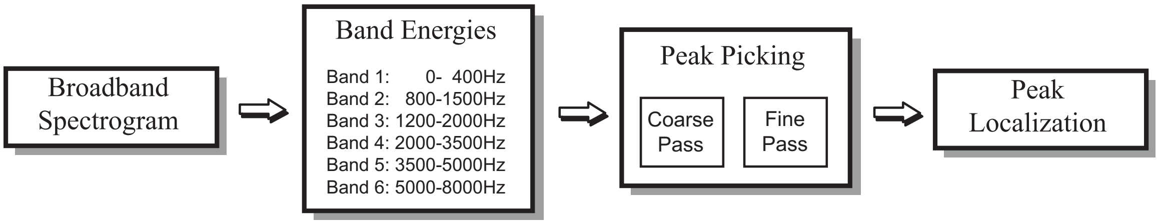


Figure 5-1: A rough block diagram of the peak-finding process

Parameter	Value
Sampling rate	16kHz
Hanning window size	6ms
Window shift	1ms
FFT frame size	512 pt

Table 5.1: The parameters used to calculate the broadband spectrogram

No.	Range
Band 1	0– 400 Hz
Band 2	800–1500 Hz
Band 3	1200–2000 Hz
Band 4	2000–3500 Hz
Band 5	3500–5000 Hz
Band 6	5000–8000 Hz

Table 5.2: Six bands used in the landmark detection algorithm

5.1.1 Broadband Spectrogram

A consonant landmark corresponds to an abruptness in the speech spectrum. Because this spectral change takes place in a short time period and affects a wide frequency range, a broadband spectrogram is best suited for the purpose of finding the time-points of abrupt changes with accuracy.

The parameters of the spectrogram are listed in Table 5.1. The spectrogram is computed with a 6ms Hanning window, shifting by 1ms steps to detect temporal information with high resolution. For each time frame, a 512-point FFT is used to provide enough frequency resolution for proper calculation of energy bands and other spectral cues.

5.1.2 Energy Bands

As explained in Section 2.1, landmarks affect different frequency regions, depending on the type of the landmark. To accommodate this property, the spectrogram is divided into six different bands as shown in Table 5.2. Shannon *et al.* [45] observed that the manner features can be perceived by human correctly when the signal is degraded into a small number of frequency bands with cut-off frequency at 800, 1500, and 2500Hz. Therefore, this division into six frequency bands is expected to capture the acoustic changes due to landmarks properly.

Energy in frequency band 1 reveals the presence of glottal vibration. Bands 2 and 3 include the 0.8-2kHz region, in which a zero may be introduced in sonorant consonants [9]. The frequency regions of Bands 2 and 3 are overlapped lest the movement of additional zero from one band to another should be mistaken with the introduction of a zero. As illustrated in Figure 5-2, when the frequency bands are overlapped, the movement of the zero can be captured at least in one of the bands. Aspiration and frication noise affects the entire frequency range, but the change is most prominent in Bands 4 and above. This frequency region is divided into smaller bands for a more reliable detection.

The energy in each of these six bands is calculated by averaging the square magnitude of the spectrogram over the frequency band. The energy band is calculated in dB. Figure 5-3 shows an example of the energy levels calculated in the six energy bands for an utterance of the sentence ‘Critical equipment needs proper maintenance.’

5.1.3 Finding the Peaks

Because the broadband spectrogram uses a 6ms window, a short-time disturbance can affect the spectrum by a great amount. To avoid such effects without sacrificing time resolution, the following methods are used: Introduction of *rate of rise* and calculating peaks in *two passes.*

Rate of Rise

The first difference of a signal is generally used to estimate the rate of change. The *rate of rise* (ROR) is similar in its purpose, but differs in that instead of taking the difference between adjacent samples, it takes the difference between samples that are farther apart. The distinction is illustrated in the following equations:

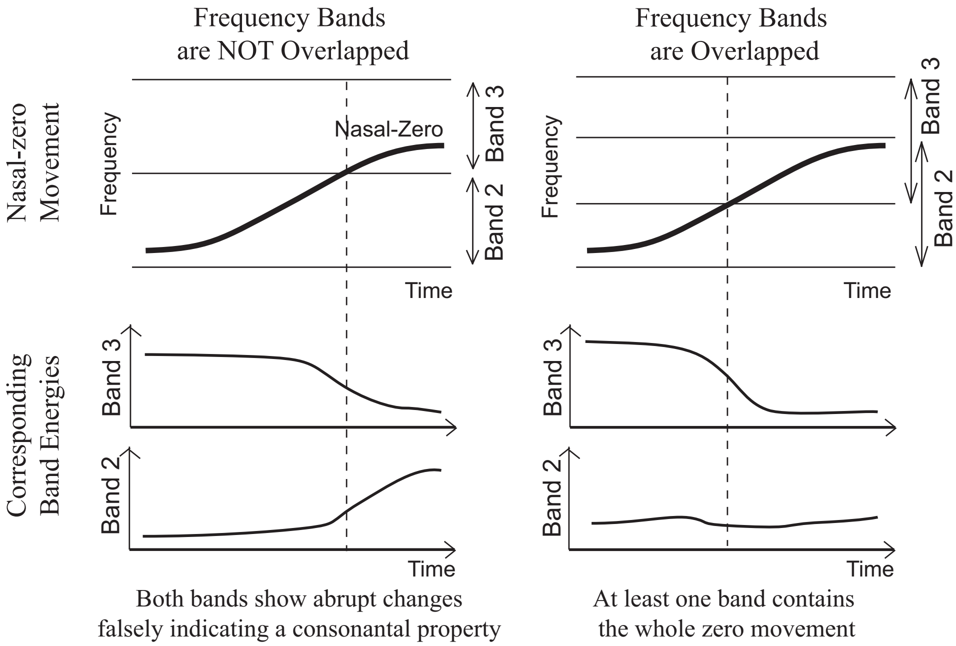


Figure 5-2: If the frequency bands are not overlapped, the movement of a nasal-zero from a band to another can be mistaken as an abrupt spectral change.

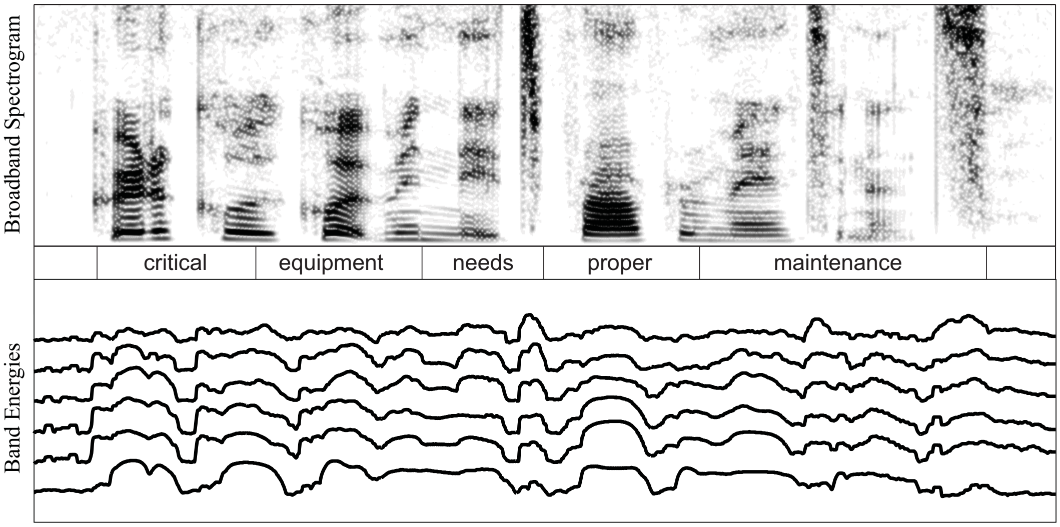


Figure 5-3: Example of a broadband spectrogram and its six energy bands

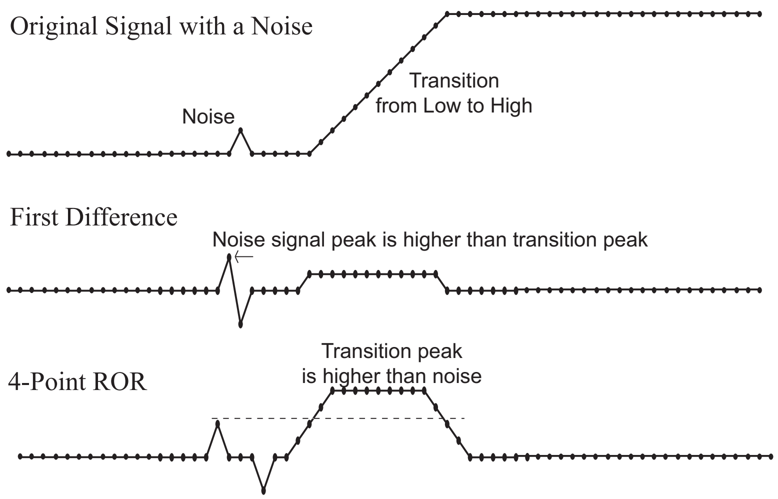
$$\begin{array}{ll} \text{First Difference} & d[n] = x[n+1] - x[n] \\ \text{6-point ROR} & r_6[n] = x[n+3] - x[n-3] \end{array}$$

Because the rate of rise calculates the change after a certain period of time, it is less affected by a change that occurs during a short period of time, and instead detects a steady change over a longer period. Figure 5-4(a) illustrates this effect. A signal at the transition from low energy to high energy is given as an input. The first difference of the signal is more sensitive to the introduction of noise than to the gradual increase during the transition, and so the peak height made due to the noise is higher than that from the transition. On the other hand, 4-point ROR, which is equivalent to the accumulated sum of four consecutive first differences, highlights a consistent increase during a certain time-period more than the distortions introduced via a sudden fluctuation by noise.

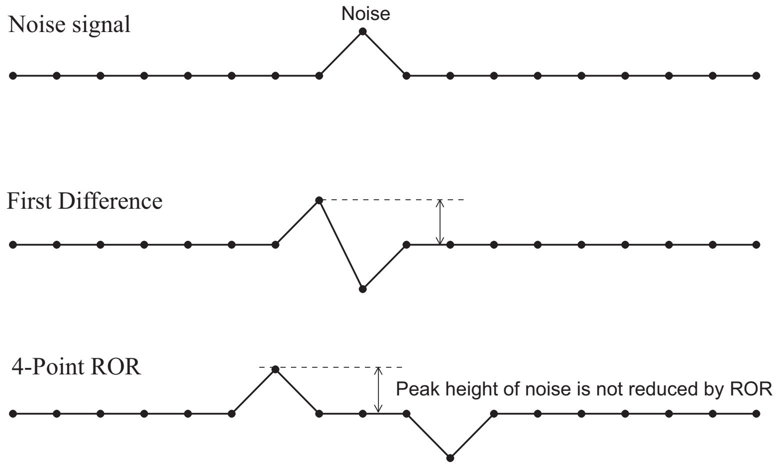
$$\begin{aligned} r_6[n] &= x[n+3] - x[n-3] \\ &= (x[n+3] - x[n+2]) + (x[n+2] - x[n+1]) + \cdots + (x[n-2] - x[n-3]) \\ &= d[n+2] + d[n+1] + d[n] + d[n-1] + d[n-2] + d[n-3] \end{aligned}$$

However, another example shown in Figure 5-4(b) illustrates the fact that even though the ROR emphasizes the steady rise in a signal, it does not reduce the magnitude of the noise signal itself. The height of the peak introduced by a sudden noise is the same in the first difference and the 4-point ROR. Therefore, a low pass filter is needed for a more robust estimation of abruptness.

An example of ROR based on a speech signal is provided in Figure 5-5. Note that the local maxima of rate of rise indicate increases in the corresponding band energy, and local minima indicate decreases. Dotted lines are inserted at some of the peaks of the rate of rise for easier comparison.



(a) ROR detects steady change of the signal more robustly than sudden noise.



(b) The peak height is the same in both first difference and ROR.

Figure 5-4: Illustrations comparing the first difference and 4-point ROR

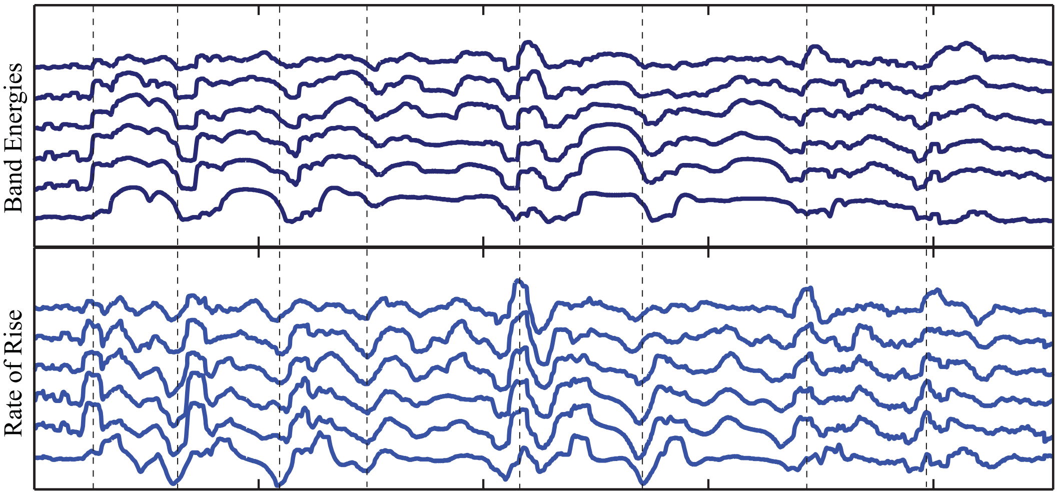


Figure 5-5: Band energies and the corresponding rate of rises

Two Pass System

Low-pass filtering with a large window reduces the noise significantly and still retains the overall energy change of the signal. However, it blurs the details, especially the abruptness, of the signal at the same time; thus, a short window must be used to pinpoint the exact location of the change. Therefore, to avoid noise and to keep the high time-resolution, two parallel processes are applied: one with a short time window (*Fine Pass*), another with a longer one (*Coarse Pass*). A shorter ROR distance is used for fine pass to detect the location with more accuracy.

Coarse pass uses a 16ms time window and 20ms ROR distance. The ROR distance is decided so that it is longer than the duration of a transition, but shorter than the duration of a schwa vowel or a frication noise. The criterion is explained in Figure 5-6. Figure (a) compares the ROR of a signal based on three ROR distances, which are shorter than, equal to, and longer than the transition time. It can be observed that if the ROR distance is shorter than the transition period, the height of the ROR peak is lower than those of the longer distances. However, the ROR distance longer than the transition period does not give a higher peak. On the other hand, Figure (b) illustrates that if the ROR distance is too long, so that the ROR distance exceeds the duration of the segment, the center of the ROR peak moves farther from the

	Coarse Pass	Fine Pass
Smoothing window	16 ms	8 ms
ROR distance	20 ms	10 ms
Threshold	7 dB	5 dB

Table 5.3: The parameters for peak picking in coarse and fine passes

actual abruptness. Therefore, the ROR distance should not exceed the duration of a segment.

On the other hand, the fine pass uses the parameters half the size of that of the coarse pass as shown in Table 5.3. This may be less robust to noise and may detect other peaks from sudden fluctuations in the signal, but the correct peak will be located at a more accurate place because less information about abruptness is filtered out.

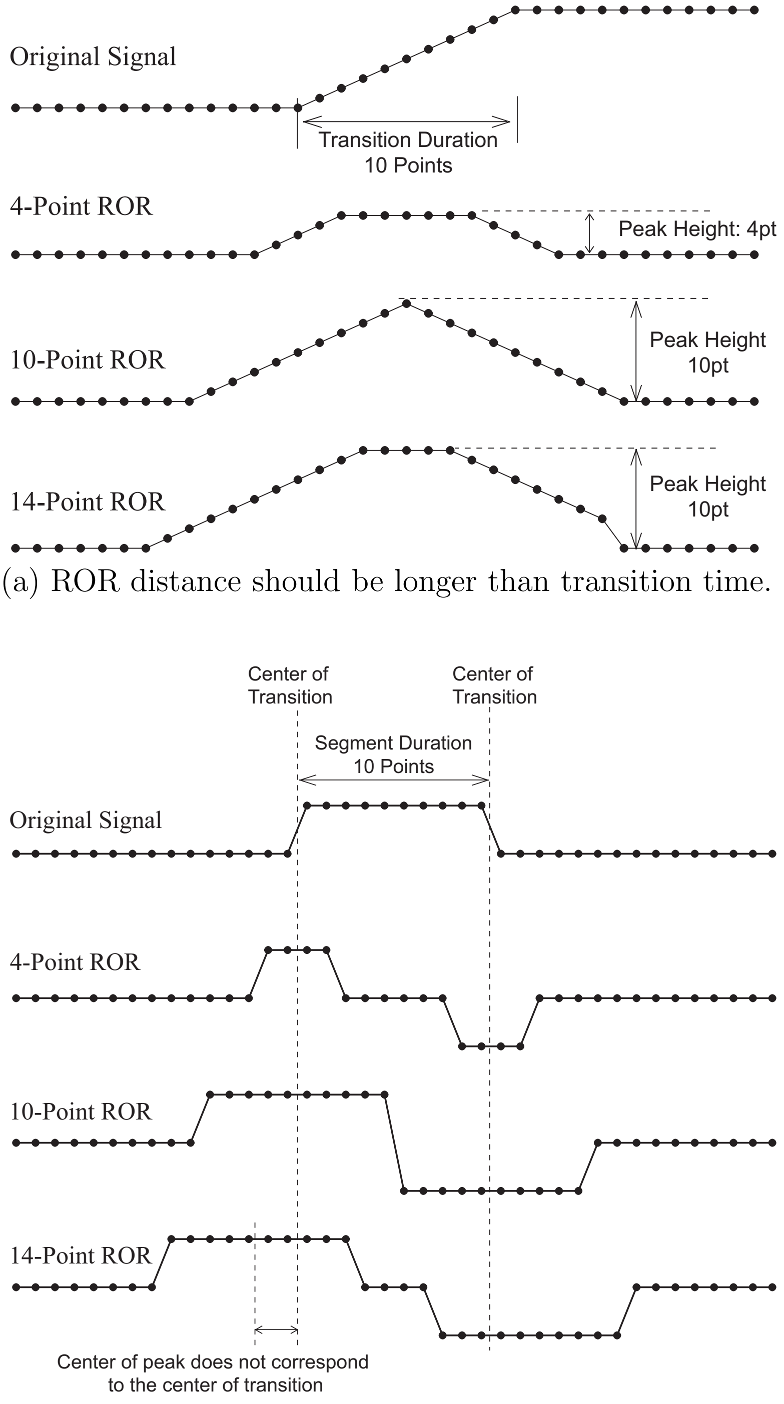
Peak Picking

After calculating ROR, the peaks of the ROR are found with Mermelstein's peak-picking algorithm [40]. The peaks of ROR correspond to the abrupt changes in the original signal. A simplified procedure of Mermelstein's peak-picking algorithm is shown below:

1. For each region over the threshold, find a maximum in the region.
2. Find a local minimum to the left and right of the maximum peak.
3. Take the difference between the minimum and the maximum peak height. If the difference exceeds the chosen threshold, split the region and apply the algorithm recursively for each region.

Some examples of peak-picking results based on this algorithm are shown in Figure 5-7. The algorithm picks at least one peaks from each region over the threshold, and if there is more than one peak in a region, picks the largest one. However, if the height of the dip between two peaks is larger than the threshold, the region is divided into two and a peak is picked in each of them.

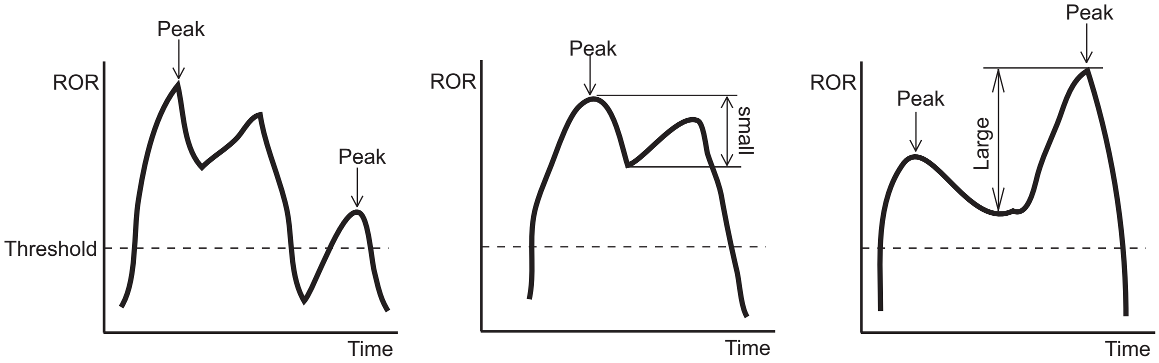
Mermelstein's algorithm is applied to pick the local maxima whose absolute values are larger than 7dB for a coarse pass or 5dB for a fine pass, and the minima with values



(a) ROR distance should be longer than transition time.

(b) ROR distance should not exceed the segment duration.

Figure 5-6: An illustration for the criterion for choosing ROR distance



(a) Peaks are picked from each region over the threshold (b) Larger peak is picked (c) Both peaks are picked

Figure 5-7: A simple illustration of the peak-picking algorithm

less then −7dB and −5dB respectively. The expected threshold for the landmark candidates is 9dB, [35] but this threshold is lowered to avoid losing landmarks. Using the lower threshold has the advantage not only in that it finds low-height peaks, but also in that it differentiates between two closely adjacent peaks. Falsely detected candidates due to the low threshold can be handled by extracting additional cues around the landmark by a method which is explained in more detail in Chapter 6.

Peak Localization

Because the coarse pass is designed to be robust to noise, and the fine pass to increase the temporal resolution, the existence of acoustic discontinuity in a frequency band is determined by a coarse-pass peak, and its exact location is determined by a fine-pass peak. Therefore, the following process is applied for each coarse-pass peak in each band, so that more exact time-points of acoustic changes can be determined. The largest positive fine-pass peak within ±15ms of each coarse-pass +peak is considered to be the localized +peak, and the largest negative one is chosen when looking for a −peak. When no such fine-pass peak is present in the vicinity of a coarse-pass peak, the coarse-pass peak is considered an error and is ignored.

5.2 Finding Landmark Candidates from the Peaks

A peak in the ROR of an energy band only means that there is a certain change in the given frequency range. This abrupt change may reflect a true landmark, when it indicates a transition between different voice sources, or an opening of the velopharyngeal port, but there are other events that can create variations in a frequency band, such as formant frequency movements. To avoid false landmarks of this type, possible landmark positions are extracted from the peaks in all six of the energy bands.

5.2.1 Criteria

A g-landmark candidate is determined by the Band 1 ROR peak location. Band 1 is directly related to the g-landmark, because the fundamental frequency, which is generated by vocal fold vibration, falls in the Band 1 frequency range.

Although onset of glottal vibration does affect the whole frequency range, the peaks in Band 2–6 are not used in g-landmark detection because they are more dependent on the context in which the landmark occurs. For example, when a fricative is followed by a vowel, the high-frequency energy in the fricative region has high amplitude, and relatively lower amplitude in the vowel, which leads to decrease of high-frequency energy at the boundary. On the other hand, if there is no fricative consonant before a vowel, there is an increase in high frequency energy at the onset of the glottal vibration. Because of this inconsistency, the ROR peaks in the higher band energies are ignored in determining the g-landmark candidates.

On the other hand, b- and s-landmark candidates are determined by ROR peaks in Bands 2–6. A sonorant consonant introduces a zero in the 800–2000Hz region, and this phenomenon results in a significant energy drop in Bands 2–4. Moreover, an obstruent consonant can be denoted by its frication noise or burst, which spans most of the high-frequency range (especially Bands 4–6). Therefore, when three out of five bands have peaks of the same sign at the same time, the position is considered to be a candidate for both an s- and a b-landmark.

5.2.2 Method

The g-landmark candidates can be determined directly from the ROR peaks, since these candidates are at the same position as the ROR peaks in Band 1. For other landmarks, however, this is not the case: A time-point where at least three peaks are present in five different bands must be found, and in some cases, these three peaks may not be detected at the exact same time-point. They may differ by several milliseconds, and it is necessary to develop a measure to determine which of the peaks can be considered to be *at the same position*.

For this purpose, a clustering algorithm is used. The following criteria are used to align the peaks:

1. The time of the ROR peaks in a cluster can be at most 50ms apart from one another.
2. No two peaks in the same frequency band can be in the same cluster.
3. No two clusters can be overlapped in time.

The 50ms criterion is used because some landmarks, especially those of sonorant consonants, have slower transition periods. The offset of a fricative also shows different transition times in different frequency bands. When there are more than two landmarks within a 50ms region, the second and third criterion can separate them.

The MaxCut algorithm [17] is applied recursively to cluster the peaks. An illustration of the algorithm is given in Figure 5-8. The procedure of clustering is as follows:

1. Create a graph on which the MaxCut algorithm will be applied.

    - **Vertices:** Each vertex represents a localized peak extracted from the previous section. The peaks from bands 2–6 are all merged in one graph, as shown in the “Merge” tier in Figure 5-8.

    - **Weighted Edges:** Each pair of vertices has an edge between them. The weight on each edge is decided by the time difference between the peaks. When the two peaks violate the second criterion (i.e., when the two peaks

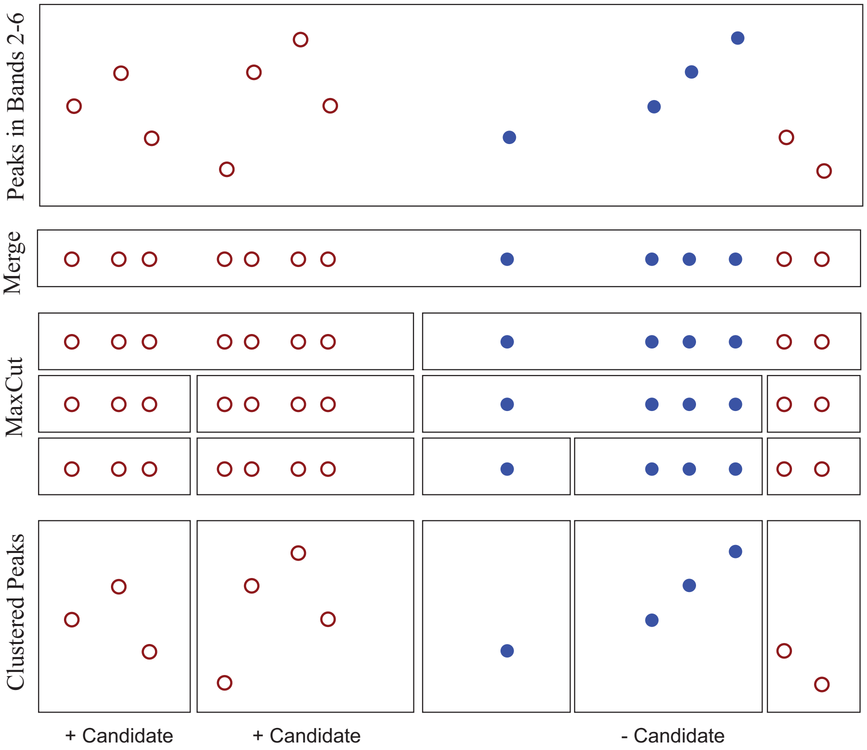


Figure 5-8: An illustration of a simple peak clustering example. Red unfilled circles represent +peaks and blue filled circles represent −peaks.

are from the same band), the weight of the edge is set to be a certain fixed value larger than 50ms.

2. Decide on a cut point which maximizes the sum of the weights on the edges that connects the divided clusters.

   Normally MaxCut is a non-deterministic polynomial-time hard (NP-hard) problem, which cannot be solved in polynomial time computational complexity. In this case, however, the number of possible clusterings does not exceed the number of vertices, due to the third criterion. Therefore, the sum of weights of all possible clusterings can be calculated in $O(n^2)$ time, where n is the number of vertices, by the following procedure:

   (a) The sum of weights for the first cut point—i.e., the cut between the first and the second vertex—is the sum of the edges $(1, i)$ for all $i > 1$.

   (b) Given the sum of weights for the $k - 1$-th cut point W_{k-1}, the sum of weights for the k-th point can be calculated by

   $$W_k = W_{k-1} - \sum_{i=1}^{k-1} w(i,k) + \sum_{j=k+1}^{n} w(k,j)$$

3. After the cut point is decided, the previous step is applied to each divided cluster recursively, until the clusters satisfy all three criteria. Each row in the "MaxCut" section of Figure 5-8 shows the clustering result after each recursion step.

4. For each cluster, check whether it includes at least three peaks. If so, accept the average time of the peaks as the time-point of s- and b-landmark candidates. In Figure 5-8, only three of the five resulting clusters pass this criterion, and so three landmark candidates are determined.

Figure 5-9 shows a step-by-step illustration of the candidate detection process in a real speech signal. When a signal is given, the energies in six bands are estimated first, and using ROR and peak-picking algorithm the ROR peaks are extracted from

Landmark Type	Cues	Description
g-Landmark	Abruptness Sonorant Levels	Height of fine pass peak Low-frequency energy on both sides
b-Landmark	Abruptness Silence Non-Silence	Height of fine pass peak Minimum energy on one side Maximum energy on the other side
s-Landmark	Abruptness Lowered Energy Vocalic Energy Increased Tilt	Height of fine pass peak Minimum energy on one side Maximum energy on the other side Difference of tilt between both sides

Table 5.4: Additional cues used to determine the correctness of landmark candidates.

each band. The g-landmark candidates correspond to the peaks in Band 1, and the b- and s-landmark candidates correspond to the cluster of the peaks in Bands 2–6.

5.3 Meaningful Cues for Each Candidate Types

Being selected as a candidate means that it is possible that there might be a meaningful landmark at the position, but it does not mean that there will always be a consistent event corresponding to the landmark type. The landmark candidates are determined only from the difference in the energy bands, and so more cues are needed to identify meaningful landmarks from random disturbances in the speech signal, and to identify the correct type of the landmark (especially to differentiate between s and b). Table 5.4 summarizes the cues that are used for each landmark candidate. Each of these cues will be discussed in more depth in this section.

5.3.1 Glottis Landmark

Abruptness

The abruptness of a transition can be measured by the peak height in frequency Band 1 in *fine pass*. In the processing described in Section 5.1.3, fine peaks were detected with only a 5dB threshold because that can differentiate two or more peaks that are

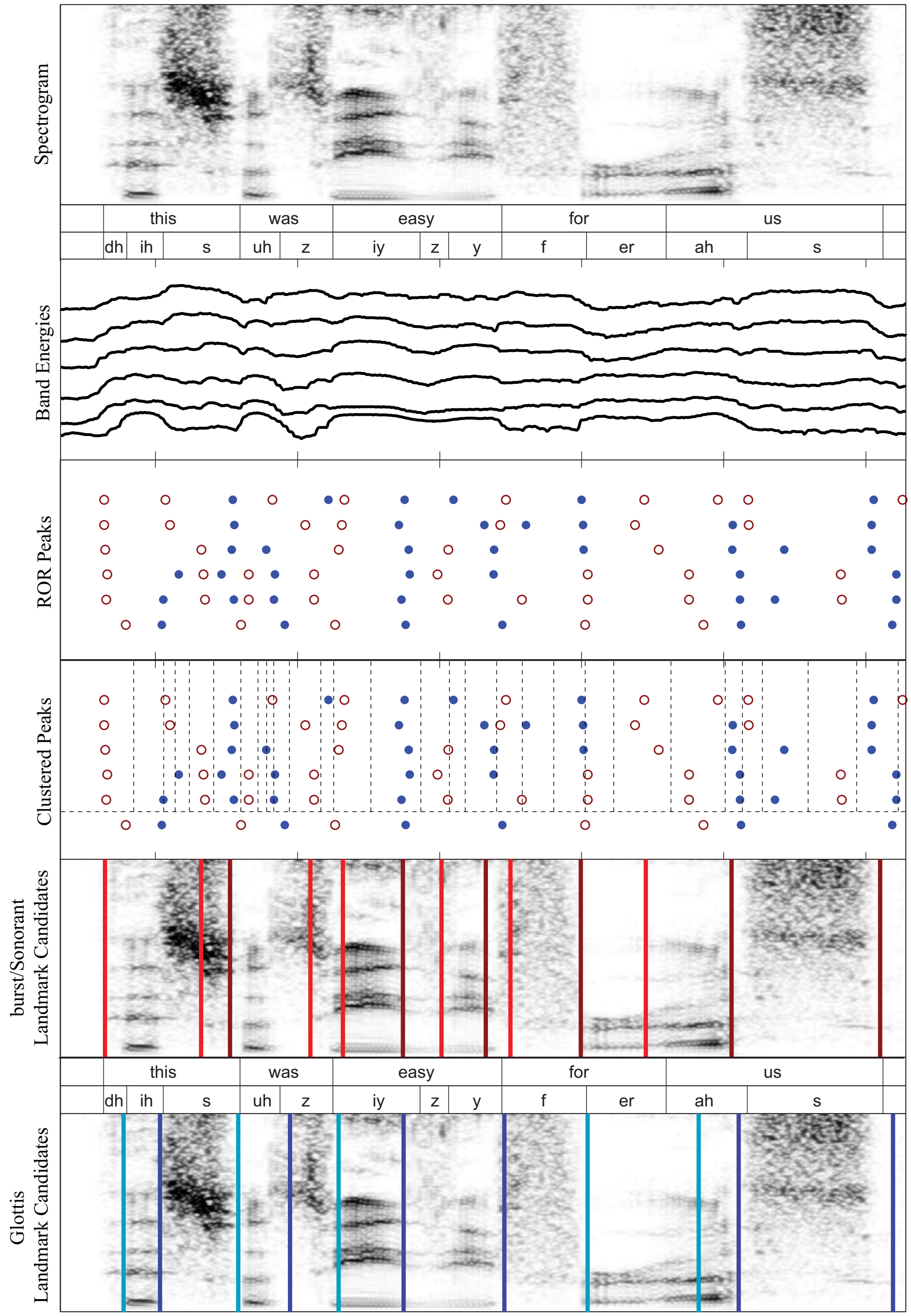
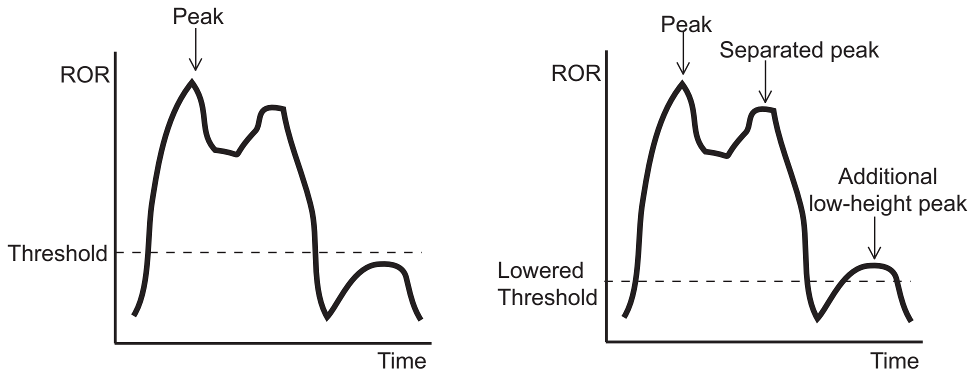


Figure 5-9: Consonant landmarks and corresponding ROR peaks. Red circles represent positive peaks, and blue dots represent negative peaks. In the landmark candidate section, lighter lines represent +candidates and darker lines represent −candidates.



(a) High Threshold (b) Low Threshold

Figure 5-10: Using a low threshold in the Mermelstein peak-picking method differentiates multiple peaks, but at the same time introduces peaks with smaller heights.

close together in time. While using such a low threshold is useful for differentiating multiple peaks, it has the disadvantage that it also generates spurious small peaks as illustrated in Figure 5-10. The parameter of peak height can be used to filter out these small insertions.

Sonorant Levels

Because abruptness is calculated by a difference of two energies, having a high abruptness does not guarantee that the energy level is high on one side and low on the other. This problem especially pertains to silent regions, because a small change in the background noise can be detected as a large jump when measured in dB. Therefore, it is necessary to measure the sonorant level on both sides of a g-landmark candidate. A g-landmark must be adjacent to a vowel or a sonorant consonant. Therefore, a g-landmark must have a high sonorant level on one or the other side of it.

The sonorant level is determined from the fine-pass Band 1 energy. It is defined to be the highest energy level such that Band 1 energy is higher than it for at least a 20ms time span, as shown in Figure 5-11. By using the criterion of *20ms time span*, we are able to exclude the effects of a sudden disturbance shorter than 20ms.

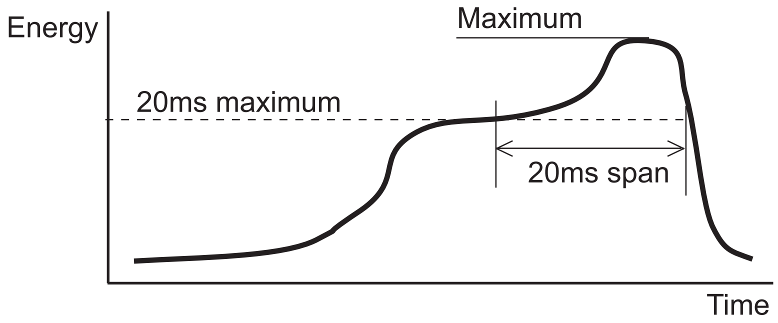


Figure 5-11: An illustration of the definition of 20ms-span maximum.

5.3.2 Burst Landmark

Abruptness

As was the case for g-landmarks, abruptness of the fine pass peak can be a crucial criterion for filtering out wrongly detected b-landmark candidates. A b-landmark candidate is postulated when there are at least three peaks within a defined time region. Therefore, instead of using five separate cues, one larger frequency band (1.2–8kHz) is used for this purpose. Low-frequency energy is not included, to avoid a voice bar during a stop closure from reducing the level of abruptness.

For the abruptness cue, the energy in the 1.2–8kHz frequency band is obtained using a 20ms Hanning window and a 13ms ROR distance. The window size and ROR distance is a little larger than the for fine-pass processing step, because the exact time of transition may be different in different frequency ranges, resulting in a longer transition period when the wider frequency band is used.

Silence

The most important criterion for the b-landmark is the existence of a silent region next to the landmark. This is because, in order to make a burst noise, a complete closure must be made to build up enough pressure so that when the pressure is released, turbulence noise will create a burst. Fricatives may be adjacent to a vowel or another fricative instead of to a silence, but those cases are not defined as b-

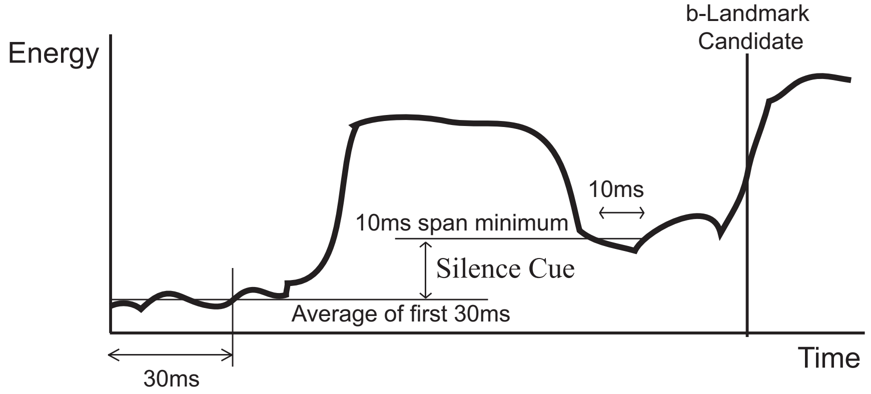


Figure 5-12: An illustration explaining calculation of silence cue

landmarks because the boundary between a vowel and a fricative is marked with a g-landmark. Moreover, the boundary between two adjacent fricatives does not have consistent abruptness; the degree of abruptness depends on the individual phonemes rather than on the general characteristics of frication noise.

The silence cue differentiates b-landmarks from s-landmarks. Since s-landmarks always occur at the boundaries between vowels and sonorant consonants, no silent region can be found next to s-landmarks. This distinction is important because b-landmark and s-landmarks share the same landmark candidates.

The same frequency band (1.2–8kHz) and smoothing window size, which were described above for detecting the abruptness cue, is used for determining the presence of silence. The silence cue is defined to be the difference between a 10ms-span minimum of the band energy and average background noise (in dB), as illustrated in Figure 5-12. The average background noise is determined as the average energy of the first 30ms period of the recording. This short time span is used for the calculation of minimum energy, due to the short closure duration for some stop consonants, and also to account for the occasional short gap between sounds.

Non-silence

As was explained in the discussion of sonorant level cues for g-landmarks, a small change in the background noise level can result in an abrupt energy difference during the silent region. Therefore, by introducing the non-silence cue, landmarks for a frication noise or burst can be separated from changes in the background noise during the silence.

For computational efficiency, the same frequency band and smoothing window size used for detecting silence cue is applied for extraction of non-silence cue as well. The non-silence cue is defined to be the difference between a *10ms-span maximum* and average background noise.

Extraction of the Cues

The acoustic cues must be estimated from different locations depending on the sign of the b-landmark. Figure 5-13 gives a graphical illustration of the cue extraction points.

A +b landmark means that there must be a transition from a silent region to a region of frication noise. Therefore, given a +b landmark candidate, abruptness cue is detected at the time point denoted by the candidate, silence cue is extracted from its left-side (between the previous and the current landmark candidates), and non-silence cues are extracted from the right-side (between the current and the next landmark candidates). On the other hand, because a −b landmark denotes the transition from a frication noise into a silence, the silence cues are extracted from the right, and the non-silence from the left.

5.3.3 Sonorant Landmark

Abruptness

To ensure the validity of an s-landmark candidate, the abruptness cue is extracted for these landmarks as well. The same frequency band (1.2–8kHz) is used, because a sonorant consonant usually introduces a zero around 1.0–1.5kHz frequency and the introduction of a zero causes energy drop in higher frequency regions. The specific range of 1.2kHz–8kHz is determined to be the same frequency band as used in the processing of b-landmarks, to reduce the computation time.

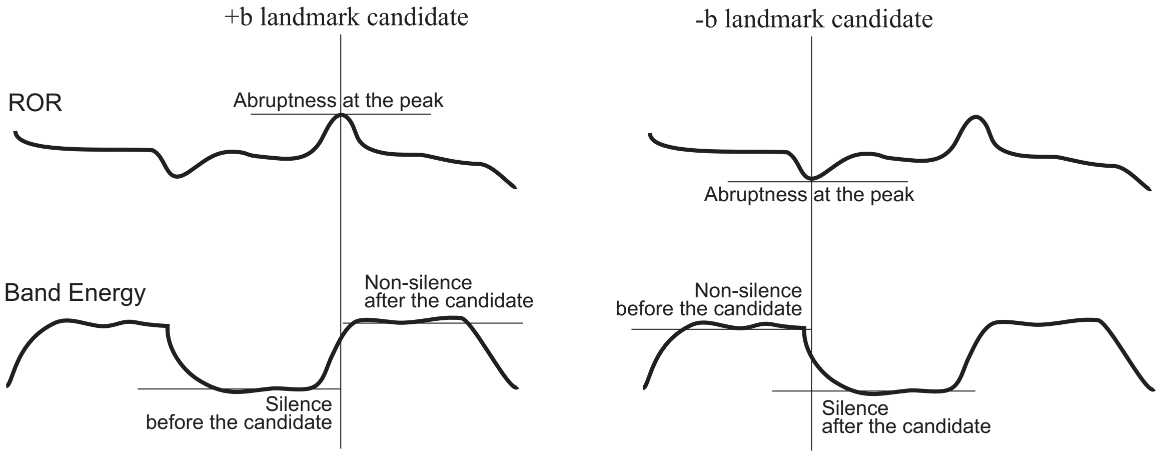


Figure 5-13: An illustration explaining where abruptness, silence and non-silence cues are detected for +b and −b landmark candidates

Lowered Energy

The energy in the high frequency band is usually lowered during a sonorant consonant. This cue is also based on the same 1.2–8kHz band, and is used with a 20ms smoothing window. The energy is defined to be the difference between 10ms-span minimum band energy and the average background noise. The parameters are set to be the same as those for the silence cues for b-landmarks, so that the cues do not need to be measured twice.

Vocalic Energy

A sonorant landmark is always located between a vowel and a sonorant consonant. A [sonorant]–[sonorant] sequence or a [sonorant]–[glide] sequence is not defined to have a sonorant landmark at the boundary between them, because the abruptness at this point is generally much lower than that for a [sonorant]–[vowel] sequence, and the energy difference varies depending on the specific phonemes or on the contexts as well. Therefore, the existence of strong vocalic energy next to a candidate can be a cue for an s-landmark.

This cue is calculated as a difference between 10ms-span maximum energy and the average background noise in the 1.2–8kHz band smoothed by a 20ms window.

The same parameters as for the non-silence cue for b-landmarks are used for this cue, to reduce redundant computation.

Increased Tilt

Tilt is defined to be the ratio of low-frequency-band energy (0–360Hz) to higher band energy (0–5000Hz). This cue is different from the previous three cues in that it utilizes the information of low-frequency energy as well as high. Therefore, this landmark can distinguish s-landmarks from b-landmarks by recognizing the existence of glottal vibration in the former. While frication noise generally tends to decrease the tilt (with possible exception of voiced weak fricatives such as /v/) due to increased energy in high-frequency region, sonorant consonants increase the tilt relative to the neighboring vowel because the nasal zeros suppress the energy in high-frequency region.

Because the duration of sonorant consonants ranges between 50–100ms depending on consonant position and speaking rate, a 30ms window is used to avoid unwanted fluctuations. The 10ms-span minimum values are measured for both the high- and low-frequency bands, and their difference is calculated because the difference in dB gives the ratio of actual values.

Extraction of the Cues

As was the case for b-landmarks, the sign of an s-landmark gives information about the segments around it. A +s landmark is a transition from a sonorant consonant into a vowel, and a −s landmark is from a vowel to a sonorant. Therefore, the acoustic cues must be extracted at different points according to the sign of landmarks. The extraction position for each cue is illustrated in Figure 5-14.

For each +s landmark candidate, the lowered energy cue needs to be extracted only from the left-side of the candidate (between the previous candidate and the current one), and vocalic energy cue should be extracted from the right-side (between the current candidate and the next one). For the same reason, for −s candidates, lowered energy cues are extracted from the right, and the vocalic energy from the left.

Because the value of tilt depends heavily on the type of vowel, it is not meaningful

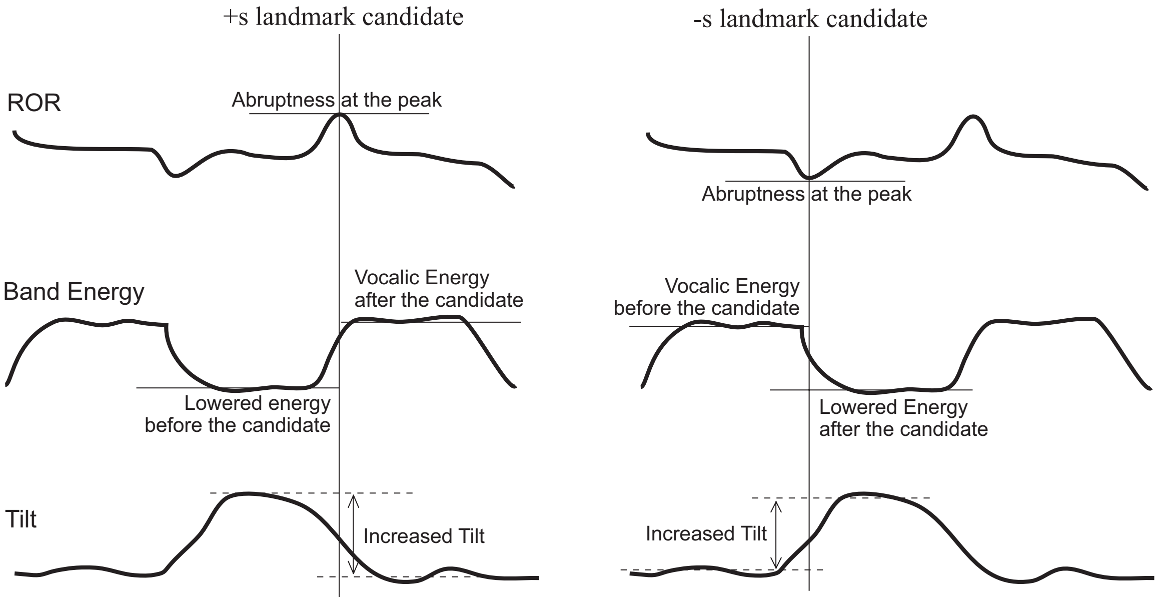


Figure 5-14: An illustration of how cues are extracted for s-landmark candidates of different sign

to examine the tilt from one side alone. Therefore, the tilts are measured from both sides and the difference is used as a cue.

5.4 Calculation of Probability

5.4.1 Introduction

Figure 5-15 shows a distribution of g-landmark cues extracted from a thousand candidate locations. The blue circled points mark the candidates that indicate the acoustically-salient abruptnesses, and the red crosses are the false alarms. The two clusters are well separated from each other, which means that the likelihood of a candidate being an important landmark can be determined effectively based on these cues. The algorithm for calculating the probability from the extracted cues is discussed in this section.

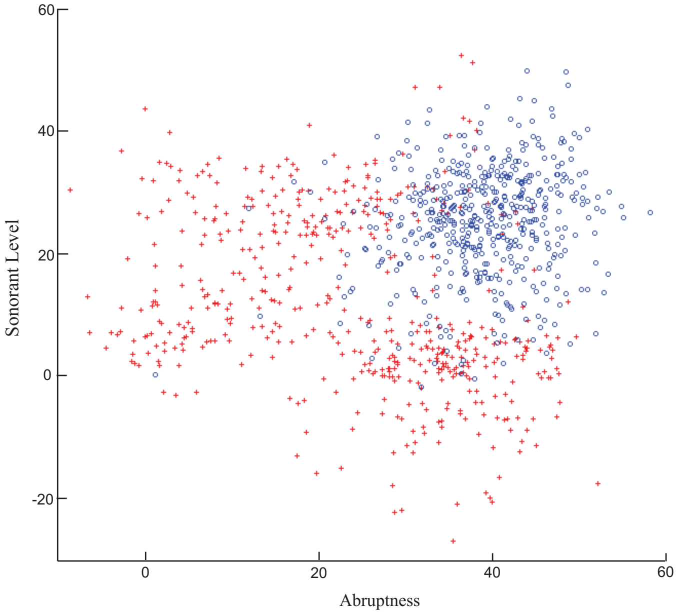


Figure 5-15: The distribution of g-landmark cues. Blue circles are the actual landmarks that are detected through landmark candidate detection process, and the red crosses are the false alarms.

5.4.2 General Formula

When a set of cues C is measured around a candidate, the probability of the candidate being a true landmark can be written as $P(\textsc{True}|C)$. Because this value cannot be evaluated directly, a simple Bayes' rule is applied to derive a more computable form.

$$\begin{aligned} P(\textsc{True}|C) &= \frac{P(C|\textsc{True})P(\textsc{True})}{P(C)} \\ &= \frac{P(C|\textsc{True})P(\textsc{True})}{P(C|\textsc{True})P(\textsc{True}) + P(C|\textsc{False})P(\textsc{False})} \end{aligned}$$

To evaluate this formula, four probability components—$P(\textsc{True})$ and $P(\textsc{False})$, $P(C|\textsc{True})$ and $P(C|\textsc{False})$—needs to be estimated. These probabilities can be trained from a reasonable data set with correct landmarks hand-labeled.

5.4.3 Estimation of A Priori Probabilities

A priori probabilities of a random candidate being a true landmark or a false alarm without any additional knowledge, $P(\textsc{True})$ and $P(\textsc{False})$, can be estimated by counting the number of true and false landmarks among the detected candidates in the training set.

$$\begin{aligned}
P(\textsc{True}) &= \frac{\text{Number of correctly detected landmarks}}{\text{Number of total detected candidates from training data}} \\
P(\textsc{False}) &= \frac{\text{Number of false alarms}}{\text{Number of total detected candidates from training data}} \\
&= 1 - P(\textsc{True})
\end{aligned}$$

5.4.4 Estimation of Cue Distributions

The probability distribution of the acoustic cues for true landmarks, $P(C|\textsc{True})$, and that of false alarms, $P(C|\textsc{False})$, can be estimated by approximating the distribution of the cues measured in the training data set. This estimation can be done by maximum likelihood parameter estimation on Gaussian mixture models. The number of Gaussian components is determined to be two, based on prior experiments.

The probability distribution of a two-component Gaussian mixture model can be written as the following, where x is a set of acoustic cues and θ is the set of all parameters (p_1, μ_1, Σ_1, p_2, μ_2, Σ_2) in the model:

$$\begin{aligned}
\Pr(x|\ \theta) &= p_1 \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right) \\
&\quad + p_2 \frac{1}{(2\pi)^{d/2}|\Sigma_2|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right) \\
&= p_1 N(x;\theta_1) + p_2 N(x;\theta_2)
\end{aligned}$$

where $p_1 + p_2 = 1$.

Commonly, an Expectation-Maximization (EM) algorithm is applied to find the maximum likelihood parameters of a mixture model for a set of training data [44]. The following procedure gives a method of finding the most suitable set of parameters for the Gaussian mixture model by an EM algorithm, when a set of data $X = \{x_1, x_2, \cdots, x_n\}$ is given.

1. Pick an initial set of parameters $\hat{\theta}$, using a simple k-means clustering method.

2. Using the calculated parameters, estimate the probability of each data point assigned to one of the two clusters: $\Pr(C_1|\ x_i, \hat{\theta})$ and $\Pr(C_2|\ x_i, \hat{\theta})$.

$$\begin{aligned}\Pr(C_k|\ x_i, \hat{\theta}) &= \frac{\Pr(x_i|\ C_k, \hat{\theta}) \Pr(C_k|\ \hat{\theta})}{\Pr(x_i|\ \hat{\theta})} \\ &= \frac{\hat{p}_k N(x_i; \hat{\theta}_k)}{\hat{p}_1 N(x_i; \hat{\theta}_1) + \hat{p}_2 N(x_i; \hat{\theta}_2)}\end{aligned}$$

3. Recompute the mixture parameters based on the previously given values.

$$\begin{aligned}\hat{\mu}_k &= \frac{\sum_{i=1}^n x_i \Pr(C_k|\ x_i, \hat{\theta})}{\sum_{i=1}^n \Pr(C_k|\ x_i, \hat{\theta})} \\ \hat{\Sigma}_k &= \frac{\sum_{i=1}^n (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \Pr(C_k|\ x_i, \hat{\theta})}{\sum_{i=1}^n \Pr(C_k|\ x_i, \hat{\theta})} \\ \hat{p}_k &= \frac{1}{n} \sum_{i=1}^n \Pr(C_k|\ x_i, \hat{\theta})\end{aligned}$$

4. Repeat steps 2 and 3 until it converges or a maximum number of steps are reached.

Figure 5-16 shows an example of the contour-plot of the probability distribution function estimated by this method, overlaid on the scatter-plot of the original cues.

5.5 Performance of Individual Landmark Detection

5.5.1 Detection Rate

The first aim of this algorithm is to make the detection rate as high as possible, even if this means that the number of insertion errors may rise as well. Elimination of the increased insertions will be handled in the processes that are explained in Chapter 6.

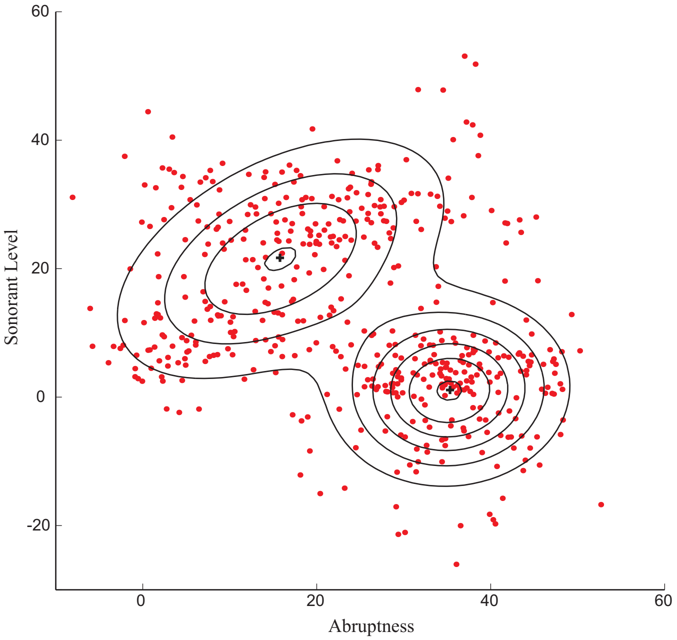


Figure 5-16: An example of cue distribution in two dimensions and the contour plot of its estimated probability distribution. The plus sign marks the mean of each Gaussian component.

Detection Criteria

In this section, the term *detected landmark* or *matched landmark* will be used to mean a labeled landmark (i.e., the landmarks predicted from the phonetic transcription of the TIMIT database using the mapping method described in Section 4.2.2) near which a corresponding landmark candidate has been detected. It can be easily understood that the landmark type and its corresponding candidate type should be the same, and at most one landmark should be matched with a landmark candidate.

A landmark represents a time-point at which a lexically-significant acoustic event occurs. Because accuracy in time is an important property of a landmark, another criterion is added: the landmark and its corresponding candidate must be within a certain distance in time. Figure 5-17 compares the results between the matching algorithms with and without the temporal restriction. The matching result on the left did not use distance criterion, and so the +g candidate on the right is matched with the +g landmark to allow matching of the +b landmark. On the other hand,

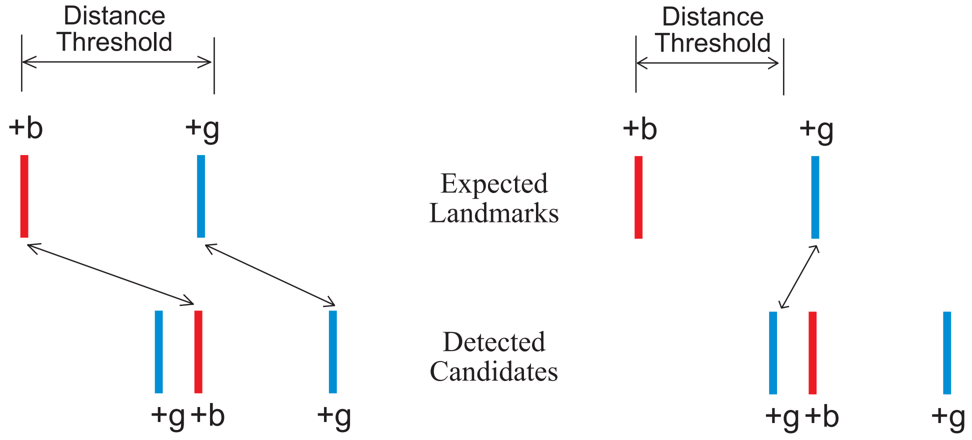


Figure 5-17: Comparison of landmark matching without distance criterion (*left*) and with distance criterion (*right*)

the figure on the right hand side uses a more strict restriction, and only the g-landmark is matched with a candidate. Short-time threshold may increase the number of unmatched candidates, but the accuracy of the matching increases.

Detection Rate vs. Maximum Distance

Figure 5-18 shows the distribution of the distance between labeled landmarks and their corresponding candidates when a distance criterion is not applied, based on the utterances in dialect region 1 of the TIMIT test set. More than 95% of the g-landmarks are detected correctly within 30ms distance, and 70% are detected correctly within 10ms.

The b- and s-landmarks show similar distributions of distance between labeled landmarks and detected candidates, except that b-landmarks are detected with more accuracy and the distances of s-landmarks are more widely dispersed. This is due to the fact that the segments represented by b-landmarks are generally burst noises and are usually very short while the segments represented by s-landmarks are sonorant consonants which have longer duration.

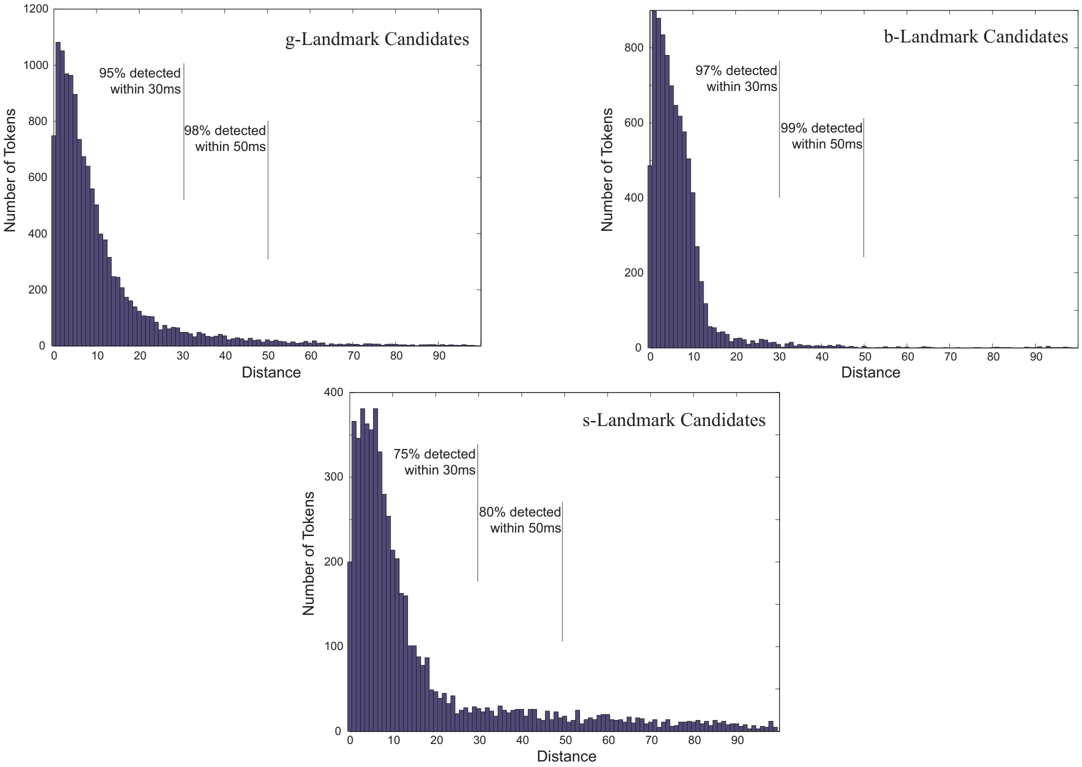


Figure 5-18: The distribution of distance between landmarks and the corresponding candidates for each landmark type estimated from the dialect region 1 of the TIMIT test set

Threshold	g-Landmark	b-Landmark	s-Landmark	Total
10ms	68.9%	82.4%	49.3%	68.5%
20ms	89.4%	95.5%	67.2%	86.7%
30ms	94.1%	97.3%	73.4%	91.0%
40ms	97.3%	98.3%	77.0%	93.7%
50ms	98.0%	99.0%	80.1%	94.9%

Table 5.5: The detection rate based on different thresholds estimated from the TIMIT test set

Detection Rate

Table 5.5 gives the detection rate of each landmark type for different distance thresholds, based on the entire TIMIT test set data. The detection rate is calculated to be the number of detected landmarks divided by the number of total labeled landmarks.

$$\text{Detection Rate (\%)} = \frac{\text{Number of Detected Landmarks}}{\text{Number of Total Labeled Landmarks}}$$

Result shows that almost all the g- and b-landmarks are detected within 50ms distance from the labeled landmarks, and 96% of g-landmarks and 96% of b-landmarks are detected correctly within a 30ms margin, but for s-landmarks only 75% are correctly detected. This is mostly due to the less abrupt nature of sonorant transitions, especially those of non-abrupt /l/ sounds which could not be automatically distinguished from abrupt /l/'s based on the phonetic transcription, and due to syllabic nasals which do not show clear distinction between vowel and nasal portions. Chen [9, 8] has developed a nasal detection module which utilizes additional acoustic cues besides band energies, and it is expected that this module will be able to compensate for the lack of spectral abruptness of sonorant consonants.

The result is comparable to the acoustic segmentation result of Glass [16], which reports 70% boundaries within 10ms of the transcription, and 90% within 20ms. Because Glass's method locates changes in the energy of the spectrogram with varying degree of sensitivity, and finds the best alignment of the acoustic discontinuities with the transcription, the result can be considered as a theoretical upper bound of the segmentation based on spectral energy change. Our result is slightly lower than that of Glass's but the landmark detection not only locates the general acoustic discontinuities, but classifies the abrupt changes according to the acoustic characteristics.

5.5.2 Insertion Rate

The insertion rate is the ratio of the number of landmark candidates that are not matched to the number of labeled landmarks. Table 5.6 shows the insertion rate of each landmark type based on a 30ms threshold.

Landmark Type	g-Landmark	b-Landmark	s-Landmark
Insertion Rate	75.9%	321%	263%

Table 5.6: The insertion rate based on 30ms thresholds estimated from the TIMIT test set

The insertion rate is high due to the lowered threshold in the peak detection process. This low threshold allows the uncertain landmarks to be detected. The b- and s-landmarks show especially high insertion rates, and this is due to a structural reason; the b- and s-landmarks share the same candidates at this point in the processing. That is, even if only correct landmarks were detected as candidates, the insertion rate of b- and s-landmarks would have to be at least 100%. This highlights the need for additional processing steps to winnow out the inappropriate candidates. The subsequent processes will be discussed in Chapters 6 and 7.

In addition, g-landmarks not only affect the low-frequency range, but also the higher frequency energies (although not reliably enough to serve as a cue.) Therefore, the b- and s-landmark candidates, which are decided solely from abrupt changes in the high-frequency bands, are apt to be detected at the point where g-landmarks are detected.

This phenomenon also increases the insertion rate significantly for b- and s-landmarks, but since these insertions correctly reflect abruptnesses, this effect does not pose a serious problem as far as the existence of a landmark is concerned. On the contrary, when the lower frequency band energy cannot be estimated clearly due to background noise, these b- and s-insertions can be used to help locate the g-landmarks. The insertions do present a problem for recognizing the *type* of landmark, which is dealt within following chapters.

5.5.3 Calculation of Probability

As was shown in the previous section, the initial candidate detection process locates most of the landmarks but it also detects a large number of false time-points. These

Type	g-Landmark	b-Landmark	s-Landmark
Type I Error	10.2%	14.1%	36.3%
Type II Error	8.6%	11.1%	18.5%

Table 5.7: Two types of errors for each landmark type estimated from the TIMIT test set

insertion errors can be dealt with partly by the probabilities calculated in this chapter.

A coarse measure of performance in probability estimation can be obtained by setting a threshold at 0.5 probability. Type I error is calculated as the ratio of candidates with probability less than 0.5 among the correctly-detected landmark candidates, and Type II error is calculated by the ratio of candidates with probability less than 0.5 among the falsely detected candidates. The correctly-detected landmark candidates are defined to be the landmark candidates which have corresponding labeled landmarks of the same type within 30ms distance.

$$\begin{aligned} \text{Type I Error} &= \frac{\text{Number of correct candidates with Prob} < 0.5}{\text{Total number of correct candidates}} \\ \text{Type II Error} &= \frac{\text{Number of false alarms with Prob} > 0.5}{\text{Total number of false alarms}} \end{aligned}$$

The two types of errors calculated from all utterances from TIMIT test set are tabulated in Table 5.7. From the measure given in the table, it can be noted that almost 90% of the falsely detected candidates can be identified solely from this coarse-grained probability measure, except for s-landmarks.

A more detailed analysis of the probability of g-landmark candidates is shown in Figure 5-19. This shows that not only are 90% of the candidates on the *correct* side of the 0.5 threshold, but most of the landmark candidates have extreme probabilities as well. That is, more than 80% of correctly-identified g-landmark candidates have more than 0.9 probability and more than 85% of false alarms have less than 0.2 probability. Therefore, by this probability measure, we will be able to distinguish false alarms from the correctly detected candidates with considerable confidence.

The next chapter will deal with the issue of distinguishing the correctly detected

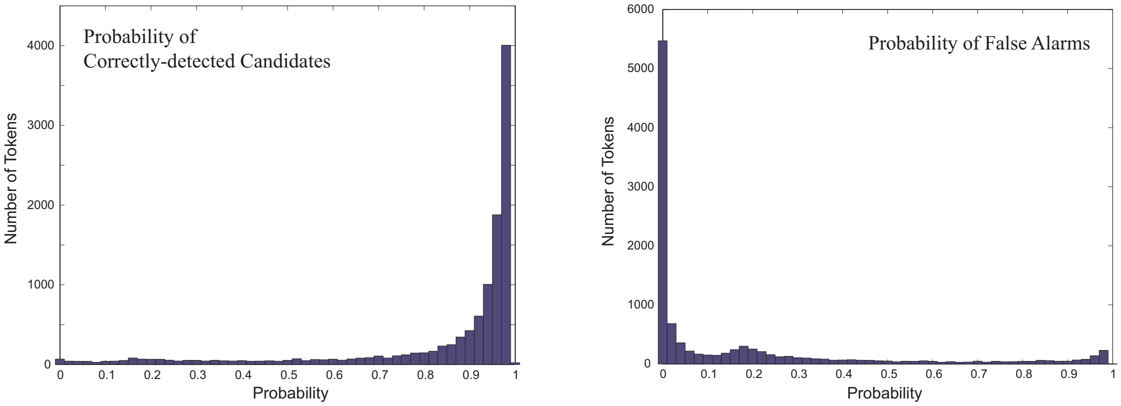


Figure 5-19: The probability distribution of correctly-detected g-landmark candidates and false alarms estimated from the TIMIT test set

landmarks from the false alarms, based on the probability cues measured in this chapter and additionally defined parameters.

Chapter 6

Landmark Sequence Determination

6.1 Motivation

This chapter deals with the problem of landmark sequence determination. It is a process that identifies and excludes the false alarms of the previous stage, and determines a sequence of landmark candidates which is likely to be the most accurate estimate of the actual landmarks. This process helps to improve the performance of the subsequent processes.

6.1.1 Landmark Selection

The initial landmark candidate detection process finds most of the places where important feature-related acoustic events occur, but in order to ensure that few landmarks are missed, the process also introduced a large quantity of false alarms. To be able to distinguish these false alarms from correctly detected landmarks, additional acoustic cues are extracted in the vicinity of the candidates, and the probability of each candidate being a true landmark was calculated from the cues.

Therefore, as a reasonable next step, it is suggested that the false alarms be identified based on these cues, and the most likely set of landmarks be selected from the previously detected landmark candidates. By removing the false alarms, unnecessary computations in the subsequent processes, such as feature extraction near the falsely detected landmarks, can be reduced.

While it is possible to filter out some of the false alarms by setting a threshold on the probability as discussed in the previous chapter, this may not be the most reasonable method of finding the true landmarks. This is because the probabilities assigned to the detected landmark candidates may depend on the context, such as adjacent vowel quality or low subglottal pressure at the end of an utterance. Therefore, a method of determining the most likely landmarks needs to be developed, incorporating some degree of knowledge about the surrounding speech signal.

6.1.2 Landmark Grouping

One of the most important uses of landmarks is to pin-point the exact place to extract additional cues for the features of the underlying segments and words of an utterance. For example, to find a spectral property of the sonorant consonant /n/ in the word 'money', such as the frequencies of the spectral peaks, the property can be found between the −s landmark located at the consonant closure and the following +s landmark at the release, as shown in Figure 6-1. Likewise, to find the VOT (Voice Onset Time) of the stop consonant /t/ in the utterance of 'her teeth', one can calculate the time difference between the +b landmark at the burst and the following +g landmark at the onset of the vowel /i/.

The cues for a single distinctive feature can be also found at several sequential landmarks. For example, the cues for a stop consonant place of articulation can be found at three different landmarks: the −g landmark at the closure, the +b landmark at the burst, and the +g landmark at the release [31, 54]. Therefore, by grouping the landmarks in a manner suitable for these subsequent processes, the performance of the processes can be improved. Figure 6-2(a) illustrates a lexical access system without an intermediate landmark grouping process. In this system, the distinctive features must be estimated from each individual landmark, and so the same features are found multiple times at different locations, and some other process then needs to decide if these separate feature bundles represent the same underlying sound or not. On the other hand, the system with a landmark grouping process in Figure 6-2(b)

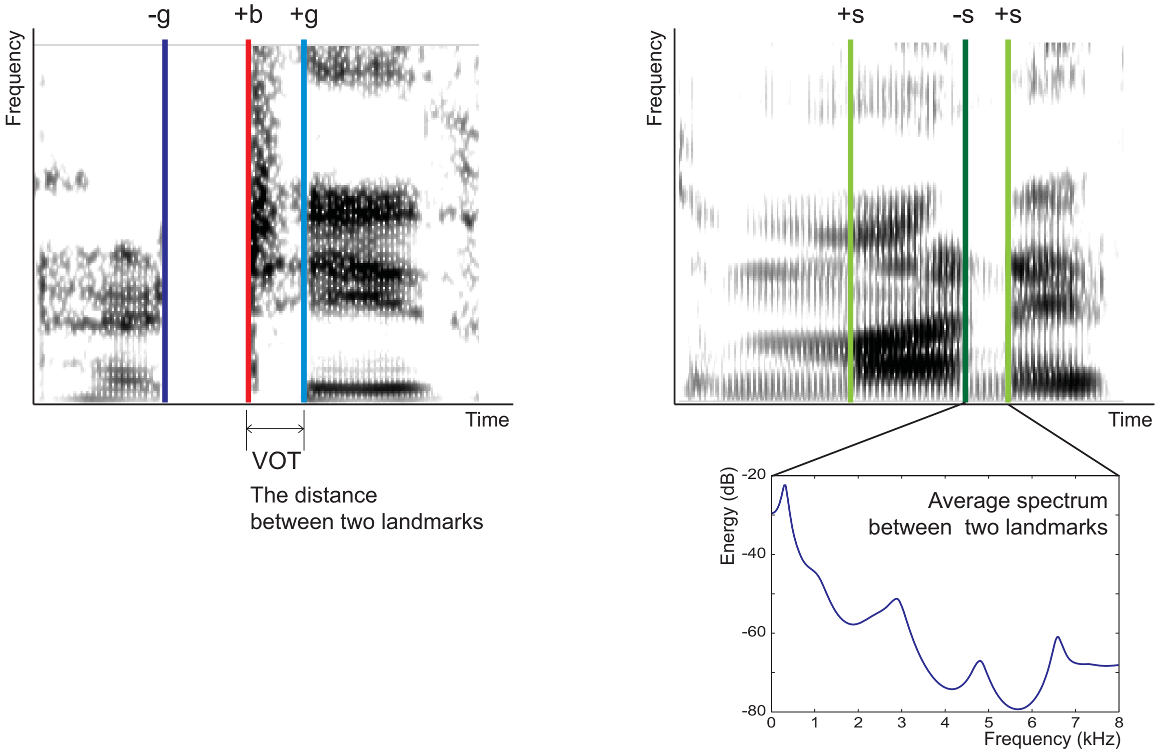


Figure 6-1: Examples of finding acoustic cues based on landmarks

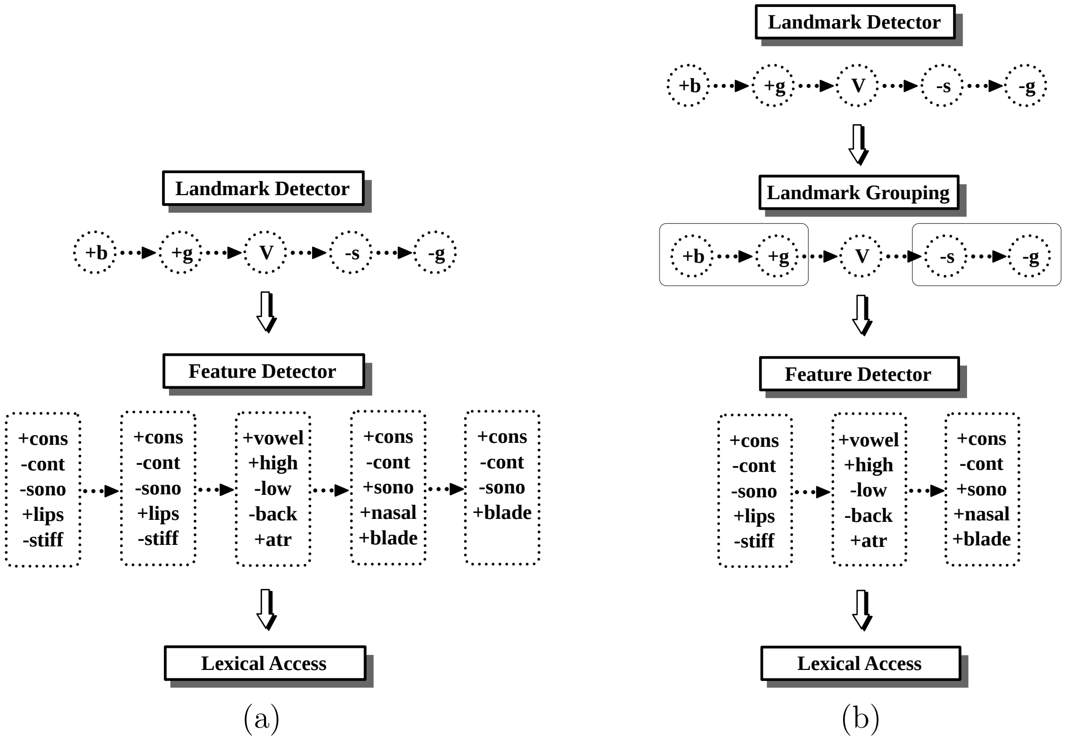

Landmark Detector
+b
+g
V
-s
-g
Feature Detector
+cons -cont -sono +lips -stiff
+cons -cont -sono +lips -stiff
+vowel +high -low -back +atr
+cons -cont +sono +nasal +blade
+cons -cont -sono +blade
Lexical Access
(a)
Landmark Detector
+b
+g
V
-s
-g
Landmark Grouping
+b
+g
V
-s
-g
Feature Detector
+cons -cont -sono +lips -stiff
+vowel +high -low -back +atr
+cons -cont +sono +nasal +blade
Lexical Access
(b)

Figure 6-2: Illustrations of lexical access systems with and without landmark grouping process

can estimate the distinctive features from a set of related landmark positions, and so the estimation are more robust, and the resulting feature sets are not redundant.

6.1.3 Relation Between Landmark Selection and Grouping

The landmark selection and grouping processes may focus on different aspects of the speech, but the two processes are related to each other. First, the landmark selection process must precede landmark grouping, because grouping cannot be reliably performed before the large quantity of false alarms is reduced. On the other hand, the landmark grouping process can provide constraints for finding a sequence of true landmarks. For example, a landmark sequence (−g, +s, +g) does not correspond to any of the possible landmark groups and such a landmark sequence should not be selected in the landmark selection process in order not to conflict with the grouping procedure.

Because of the mutual dependency of these two processes, an alternative process is developed that combines their advantages. The landmark sequence determination process explained in this chapter is designed to create an output that consists of high-probability candidates and some of the low-probability candidates which do not conflict with the constraints given by landmark groupings.

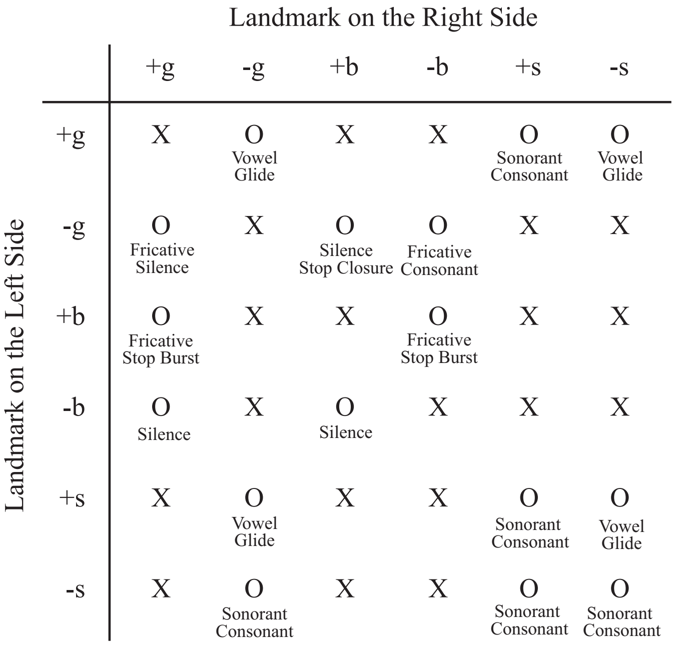
6.2 Bigram Method

6.2.1 Representation of Constraints on Possible Landmark Sequences

The constraints from the landmark grouping can be represented efficiently with a bigram model. A bigram model provides effective constraints on possible landmark sequences, because each landmark represents a movement of an articulatory organ. For example, a (+g, +g) landmark sequence cannot be produced by the voice source, because a +g landmark denotes the onset of a vocal fold vibration and two +g landmarks in a row means that the vocal folds can be turned on multiple times without any turn-off or pausing in between, which would introduce a −g landmark.

Table 6.1 shows the theoretical constraints in the bigram model. More than half of all landmark pairs cannot be produced for physiological reasons. Moreover, each *legal* (meaning physiologically possible) landmark pair can provide information about some possible characteristics of the segment between the landmarks.

A trigram model could also be used to get a stronger restriction, but because the aspects of the landmarks are mostly binary—e.g., start and end of vocal fold vibration, opening and closure of velopharyngeal port, existence of frication noise—physiological constraints can be well captured with a bigram model. The additional constraints that a trigram model can provide are mostly phonotactic constraints which are dependent on specific languages, and this is not a great advantage, considering the cost in increased complexity of computation.



		Landmark on the Right Side					
		+g	-g	+b	-b	+s	-s
Landmark on the Left Side	+g	X	O Vowel Glide	X	X	O Sonorant Consonant	O Vowel Glide
	-g	O Fricative Silence	X	O Silence Stop Closure	O Fricative Consonant	X	X
	+b	O Fricative Stop Burst	X	X	O Fricative Stop Burst	X	X
	-b	O Silence	X	O Silence	X	X	X
	+s	X	O Vowel Glide	X	X	O Sonorant Consonant	O Vowel Glide
	-s	X	O Sonorant Consonant	X	X	O Sonorant Consonant	O Sonorant Consonant

Table 6.1: A table of possible (marked with O) and impossible (marked with X) landmark pairs

6.2.2 Estimation of the Bigram Model

A bigram model of the landmark sequence is trained. Given a landmark s_1, the probability that the next landmark will be s_2 is calculated using maximum likelihood estimation.

$$B(s_1, s_2) = \frac{\bar{n}(s_1, s_2)}{\sum_{s \in \mathcal{L}} \bar{n}(s_1, s)}$$

In this formula, $\bar{n}(s_1, s_2)$ represents the empirical count of landmark sequences (s_1, s_2) in the training set. The set of all the consonant landmarks $\mathcal{L}$ is defined as $\mathcal{L} = \{$+b, $-$b, +s, $-$s, +g, $-$g, u_{st}, $u_{end}\}$. Two additional symbols are used to indicate the start and end of an utterance. The symbol u_{st} represents the start of an utterance, and u_{end} represents the end of an utterance. The results obtained from the labeled landmarks in the training set of the TIMIT database are shown in Table 6.2. The blank cells represent the impossible landmark pairs. This result verifies the theoretically expected landmark pairs listed in Table 6.1.

		Next Landmark						
		+g	−g	+b	−b	+s	−s	u_{end}
Current Landmark	+g		55.8			9.2	35.0	
	−g	33.6		45.2	14.8			6.4
	+b	90.2			9.8			
	−b	13.2		62.3				24.5
	+s		66.3			0.4	33.3	
	−s		44.3			56.0	0.7	
	u_{st}	40.3		59.7				

Table 6.2: Bigram matrix for six types of landmarks estimated from the TIMIT training set. The number denotes the probability of the landmark on the top row following the landmark on the left. When the number is not specified, the transition is illegal.

6.3 Landmark Selection with a Viterbi Search Algorithm

6.3.1 Score

To select the most likely sequence of landmarks, there must be a measure of likelihood. The likelihood score can be calculated based on the *individual probability* of each landmark candidate calculated from the acoustic cues and the *transition probability* given by the bigram. The overall score of a landmark sequence is calculated by the product of these probabilities as shown in the example in Figure 6-3.

$$P(S) = P_I(S)P_T(S)$$

The individual probability score of a landmark sequence S can be calculated by multiplying the probability of each selected landmark candidate being a true landmark, and the probability of each non-selected landmark candidate being a false landmark. The product of the probabilities of non-selected landmark candidates being false landmarks is taken into account, because otherwise an empty sequence will get the highest score.

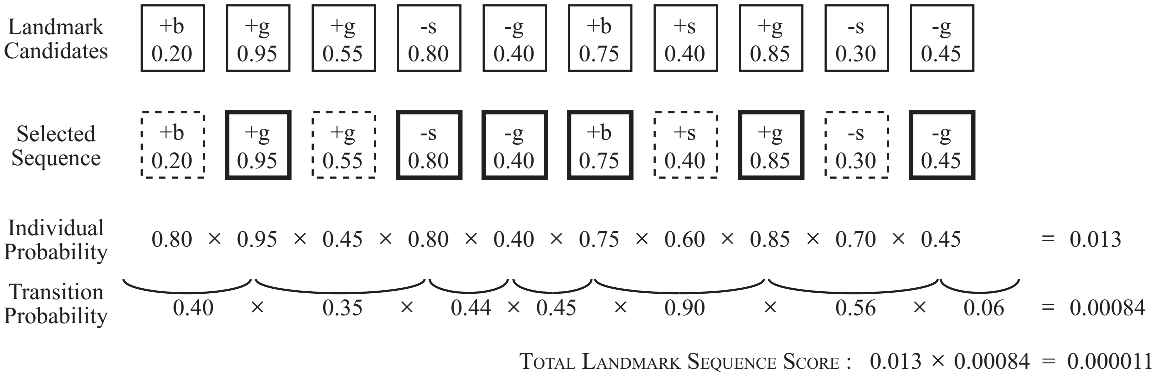


Figure 6-3: An example of scoring of a selected landmark sequence. A selected landmark sequence is represented by the items within thick-bordered boxes, and its corresponding probability score is calculated. The transition probabilities in Table 6.2 are used for this calculation.

$$\begin{aligned} P_I(S) &= \prod_{s \in S} \Pr(\textsc{True}|\ s) \prod_{s \notin S} \Pr(\textsc{False}|\ s) \\ &= \prod_{s \in S} \Pr(\textsc{True}|\ s) \prod_{s \notin S} [1 - \Pr(\textsc{True}|\ s)] \end{aligned}$$

The transition score of a sequence of selected landmarks can be calculated by multiplying all the bigram transition probabilities in the sequence, including u_{st} and u_{end}.

$$P_T(S) = \prod_{(s_i, s_{i+1}) \in S} B(s_i, s_{i+1})$$

Therefore, the total score can be computed by multiplying the individual score and the transition score.

$$P(S) = \prod_{s \in S} \Pr(\textsc{True}|\ s) \prod_{s \notin S} [1 - \Pr(\textsc{True}|\ s)] \prod_{(s_i, s_{i+1}) \in S} B(s_i, s_{i+1})$$

6.3.2 Viterbi Search

To find the landmark sequence S that maximizes the score $P(S)$, a Viterbi search algorithm is applied to a set of landmark candidates $C = (c_1, c_2, \cdots, c_n)$. The search graph used for the Viterbi algorithm is constructed as the following:

- **States**: The states used in the Viterbi algorithm start with u_{st} and end with u_{end}, and all the landmark candidates are arranged in time-order between them.

$$States = \{c_0 = u_{st},\ c_1, c_2, \cdots, c_n,\ c_{n+1} = u_{end}\}$$

- **Transition**: Transition can occur only from an earlier state to a later state. The transition score from c_i to c_j is calculated as follows. Note that impossible transition has transition probability of zero, which is equivalent to having no edge.

$$Tran(c_i, c_j) = B(c_i, c_j) \Pr(\textsc{True}|\ c_j) \prod_{i<k<j} (1 - \Pr(\textsc{True}|\ c_k))$$

  Assume that $\Pr(\textsc{True}|\ u_{end}) = 1$ because the end of utterance must be always reached.

An illustration of graph construction with a set of simple landmark candidates is shown in Figure 6-4. By finding the maximum-score transition from u_{st} to u_{end}, we can determine the most likely landmark sequence. Figure 6-5 shows the Viterbi search result that corresponds to the example in Figure 6-3. Note that the product of all the transition scores along the resulting path is the same as the probability score of the selected landmark sequence in Figure 6-3.

6.4 Results

6.4.1 Evaluation Criterion

For the evaluation, each expected (i.e., labeled) landmark is matched with a corresponding landmark candidate. The types of the landmark candidates are not considered in the matching process, and candidates more than 30ms from a labeled landmark are ignored to emphasize the importance of the accuracy in time. The formulas for

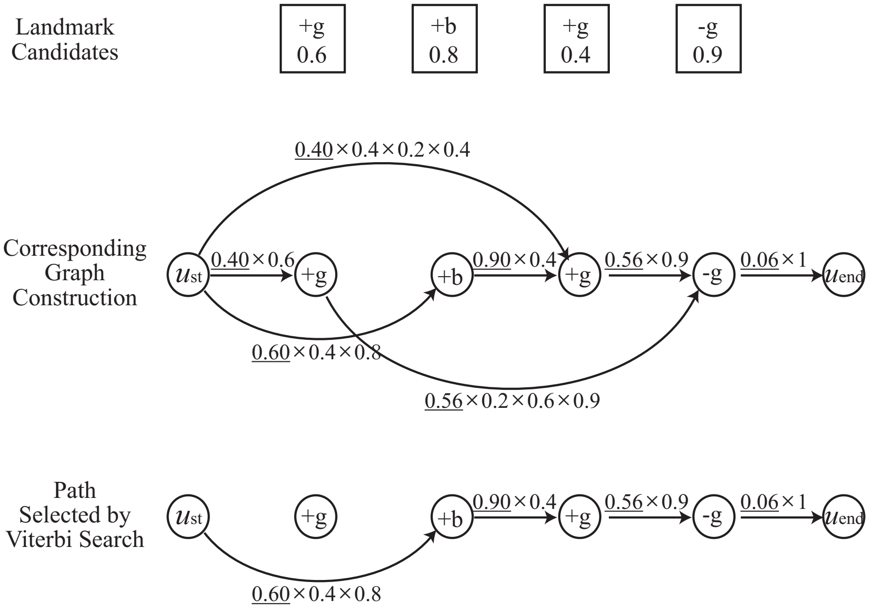


Figure 6-4: An example of graph construction and its Viterbi search result. The graph in the middle is the graph constructed from the landmark candidates above it. Every pair of vertices in the graph has a directed edge from left to right, but the edges with zero weight are not drawn here for clarity. The underlined numbers are transition probability elements, and others are individual probabilities. The result of the Viterbi search algorithm is shown in the bottom.

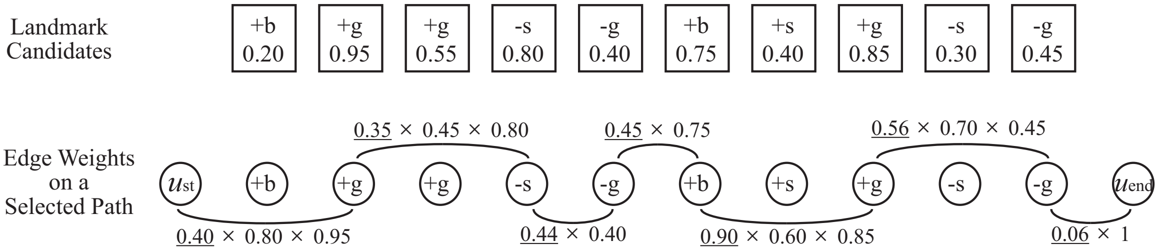


Figure 6-5: An illustration explaining the relationship between the score defined in Section 6.3.1 and the product of weights on a path of the constructed graph. The path shown above corresponds to the selected landmarks in Figure 6-3. The underlined numbers are transition probability elements, and others are individual probabilities.

Type	g-Landmark	b-Landmark	s-Landmark	Total
Detection	86.2%	74.9%	52.3%	76.8%
Deletion	4.4%	12.6%	30.7%	11.6%
Substitution	9.4%	12.5%	17.0%	11.6%
Insertion	7.6%	27.3%	18.8%	14.7%
Error	21.4%	52.4%	66.5%	37.9%

Table 6.3: The error rate of landmark sequence determination with bigram constriction estimated from the TIMIT test set

computing the rates of detection, deletion, substitution and insertion are as follows:

$$\begin{aligned}
\text{Detection Rate} &= \frac{\text{Number of landmarks matched with the same type}}{\text{Total number of expected landmarks}} \\[2ex]
\text{Deletion Rate} &= \frac{\text{Number of unmatched landmarks}}{\text{Total number of expected landmarks}} \\[2ex]
\text{Substitution Rate} &= \frac{\text{Number of landmarks matched with different type}}{\text{Total number of expected landmarks}} \\[2ex]
\text{Insertion Rate} &= \frac{\text{Number of unmatched landmark candidates}}{\text{Total number of expected landmarks}}
\end{aligned}$$

The total error rate is calculated to be the sum of the deletion, substitution and insertion rates.

$$\text{Error Rate} = \text{Deletion Rate} + \text{Substitution Rate} + \text{Insertion Rate}$$

6.4.2 Evaluation

The detection rate estimated from the whole test set of the TIMIT database is listed in Table 6.3.

Insertion errors are the detected time-points, but with no existing acoustic events nearby. Therefore, these false alarms lead to unwanted computations in the processes that follow. Applying the sequence determination process reduced the insertion rate significantly from 100–300% to 15%. There are many b-landmark insertion errors, but almost 80% of those insertions are located within 10ms of g-landmarks, and these

Type	g-Landmark	b-Landmark	s-Landmark	Total
Detection	82.4%	86.5%	54.6%	77.0%
Deletion	8.3%	9.6%	37.4%	15.2%
Substitution	9.4%	3.9%	8.0%	7.8%
Insertion	13.1%	100.4%	114.8%	56.4%
Error Rate	30.8%	114.2%	160.2%	79.4%

Table 6.4: Landmark sequence determination result without bigram constriction estimated from the TIMIT test set

insertions can be eliminated without difficulty, by introducing duration cues.

Almost 12% of the expected landmarks are substituted with landmarks of different types. The substitutions still indicate the locations of significant and informative events, and in this sense are useful in the analysis process. However, these errors may give incorrect information about the types of additional feature cues that can be found near the landmark. Among the substitution errors, 59% occur between g- and s-landmarks, mostly due to flaps and voiced fricatives which have significant energy drop in 0-400Hz range but also have persistent voicing during the closure. About 13% of the substitutions are due to the confusion of b-landmarks as g-landmarks. This type of error commonly occurs because when the burst noise of a stop consonant is too short to be recognized properly, the b-landmark label is often matched with the neighboring g-landmark.

The deletion errors are the most crucial among the three types. This type of error means that the location of an existing acoustic event is not detected at all, either correctly or as another type of landmark, and additional measure must be taken to find the ignored event. Fortunately, only 4.4% of g-landmarks are deleted completely, which enables us to estimate the general syllable structure with confidence. A large portion of the s-landmark deletions are due to non-abrupt /l/'s, which do not make abrupt changes in the spectrum.

6.4.3 The Effect of Bigram Constraints

To understand the effect of the bigram constraints on the landmark sequence determination process, we show the results of sequence determination without applying

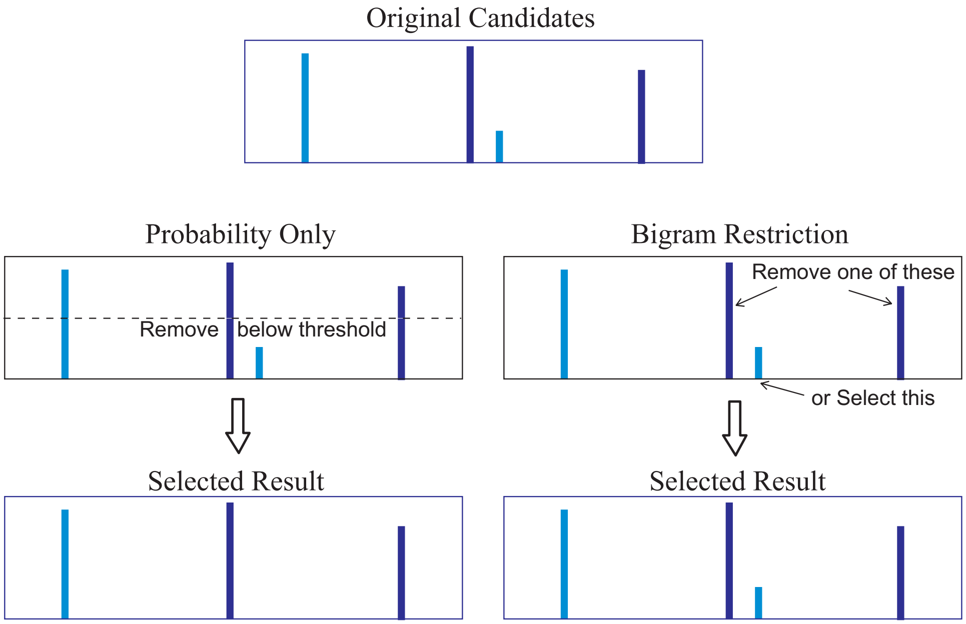


Figure 6-6: An example of how bigram constraints help increase the correct detection

the bigram. This result is shown in Table 6.4.

Generally, the overall detection rate does not differ much from the results with bigram constraints. The most significant difference is in the reduction of insertion errors from 56.4% to 14.7%. When the bigram restriction is not applied, almost three quarters of the falsely detected candidates are removed, as was expected from the distribution of probabilities. However, due to the large number of false alarms in the original set of landmark candidates, the rate of insertion errors for b- and s-landmarks still remains at about 100%. When the bigram model of possible landmark sequences is applied, these errors are reduced considerably.

Note that the deletion error rate of g-landmarks is reduced from 8.3% to 4.4% after applying bigram restriction, even though a more strict constraint is applied. This is because of the very extreme distribution of g-landmarks. Figure 6-6 illustrates this effect. When a g-landmark is detected with a low probability, it will not be detected when bigram restriction is not applied. On the other hand, when the bigram restriction is used, the low-probability landmark can be accepted in order to avoid the deletion of the neighboring high-probability candidates.

Training	Test	Detect	Delete	Subst	Insert	Error
Dialect 1	Dialect 1	76.8%	12.0%	11.2%	15.0%	38.2%
Dialect 3	Dialect 1	76.8%	11.8%	11.3%	15.0%	38.1%
Dialect 1	Dialect 3	76.8%	11.7%	11.5%	14.6%	37.8%
Dialect 3	Dialect 3	76.8%	11.7%	11.5%	14.5%	37.7%
Male	Male	75.9%	12.2%	11.8%	13.9%	38.0%
Female	Male	76.0%	12.4%	11.6%	13.9%	37.9%
Male	Female	79.2%	10.1%	10.8%	16.9%	37.7%
Female	Female	79.2%	10.1%	10.7%	17.1%	38.0%

Table 6.5: Comparison of results of landmark candidate detection followed by sequence determination across dialects and gender of the TIMIT database. Dialect 1 corresponds to New England, and Dialect 3 represents North Midland.

6.4.4 Comparison Across Dialects and Gender

The landmark candidate detection and sequence determination algorithms are trained and tested on different dialect regions and genders, to compare the effect of these factors on the detection performance. The results are listed in Table 6.5.

When different genders and dialects are investigated, the overall error rate does not change by more than 1%, indicating that the landmark detection and sequence determination processes are robust to those factors.

Individual error rates—detection, deletion, substitution, and insertion rates—sometimes change more than 2–3% depending on the situations, but they differ less than 0.1% when the same test set is used even when the models are trained in different contexts. This verifies the expectation that the distribution of acoustic cues corresponding to landmarks and the probability of bigram transitions can be trained robustly without considering the dialects and gender.

6.5 Comparison to Related Works

6.5.1 Liu's Landmark Detector

A large portion of the landmark detection algorithm has been derived from Liu's landmark detector [35]. Liu's landmark detection process was originally developed based on LAFF (Lexical Access From Features) database, which consists of a hundred

Type	g-Landmark	b-Landmark	s-Landmark	Total
Detection	91%	76%	44%	79%
Deletion	6%	19%	40%	15%
Substitution	3%	5%	16%	6%
Insertion	10%	72%	23%	25%
Error Rate	19%	96%	79%	46%

(a) Detection result based on automatically mapped landmarks

Type	g-Landmark	b-Landmark	s-Landmark	Total
Detection	96%	94%	60%	90%
Deletion	2%	3%	23%	5%
Substitution	2%	3%	17%	5%
Insertion	9%	23%	38%	15%
Error Rate	13%	29%	78%	25%

(b) Detection result based on hand-corrected landmarks

Table 6.6: Landmark detection result on TIMIT test set using Liu's landmark detection algorithm

sentences constructed from a limited number of words (200 words), but a slightly modified algorithm that is adapted to the TIMIT database has been developed as well. The result on the TIMIT test set using the TIMIT-oriented algorithm is shown in Table 6.6(a). Because the landmarks are not labeled in the TIMIT database, a set of automatically generated landmarks mapped from TIMIT transcription has been used for comparison.

Overall detection rate of Liu's algorithm is generally higher than our landmark sequence determination algorithm, but this is mostly due to low substitution rate; the number of landmarks that are completely unlocated is higher in Liu's landmark detector than in ours. Also note that the strict bigram restriction in our landmark detection algorithm could reduce the insertion rate by half.

Although overall performance of the landmark sequence is similar except for lower insertion rate in the newly developed algorithm, the new algorithm has advantage in that the landmark candidates are explicitly determined. Liu's landmark detection algorithm does not provide any cues for undetected landmarks, which comprises almost 15% of the total landmarks. Therefore, for the undetected landmarks to be located, the signal must be re-examined. On the other hand, the new algorithm first generates a large number of landmark candidates which includes almost 97% of the correct landmark positions, and a subsequent processing selects the most likely sequence. Therefore, when a landmark is not selected in the sequence determination process, it can be later added when it is needed.

Liu's algorithm has been also applied to TIMIT database with hand-corrected labels [35], and its result is shown in Table 6.6. The hand-correction includes: removing landmarks due to non-abrupt /l/, marking abrupt changes due to heavily voiced fricatives as s-landmarks, removing short burst less than 20ms apart from voice onset, and so on. When the landmark detection performance based on the hand-corrected labels were much higher than that of the automatic mapping. Therefore, it can be expected that the new algorithm would be able to achieve much higher detection rate, if correctly labeled landmarks were used instead of automatically mapped labels.

6.5.2 Other Segmentation Algorithms

The landmark detection algorithm can be compared to broad-class segmentation of speech signal as well. Segmentation task divides the consonants into intervals of uneven length, each of which represent a phonetic category.

Juneja and Espy-Wilson [30] segmented the speech signal into five broad classes—vowel, sonorant consonant, fricative, stop and silence. The segmentation was performed by extracting various acoustic parameters from each 5ms time frame, classify each frame into one of the five categories using binary support vector machines [7], and then merging the frames with same broad-class characteristics. The performance evaluation shows that the correctness of the classification is 79.8% and the accuracy is 68.1%. The correctness corresponds to our detection rate, and the accuracy corresponds to detection rate subtracted by insertion rate.

The accuracy is higher than that of our algorithm, which is 62.1%, but the computational complexity of Juneja and Espy-Wilson's algorithm is much higher because their algorithm extracts acoustic cues from every 5ms frames and then classifies each frame into one of the categories using four different SVM's. On the other hand, our algorithm extracts acoustic parameters and calculates probabilities only near acoustic discontinuities, which are mostly separated by 50–100ms from one another. This difference is because our approach is developed as a computationally-light process that provides starting points for subsequent distinctive feature extraction processes, while Juneja and Espy-Wilson's approach aimed for an algorithm that can be extended to phoneme recognition by applying complete feature hierarchy [30].

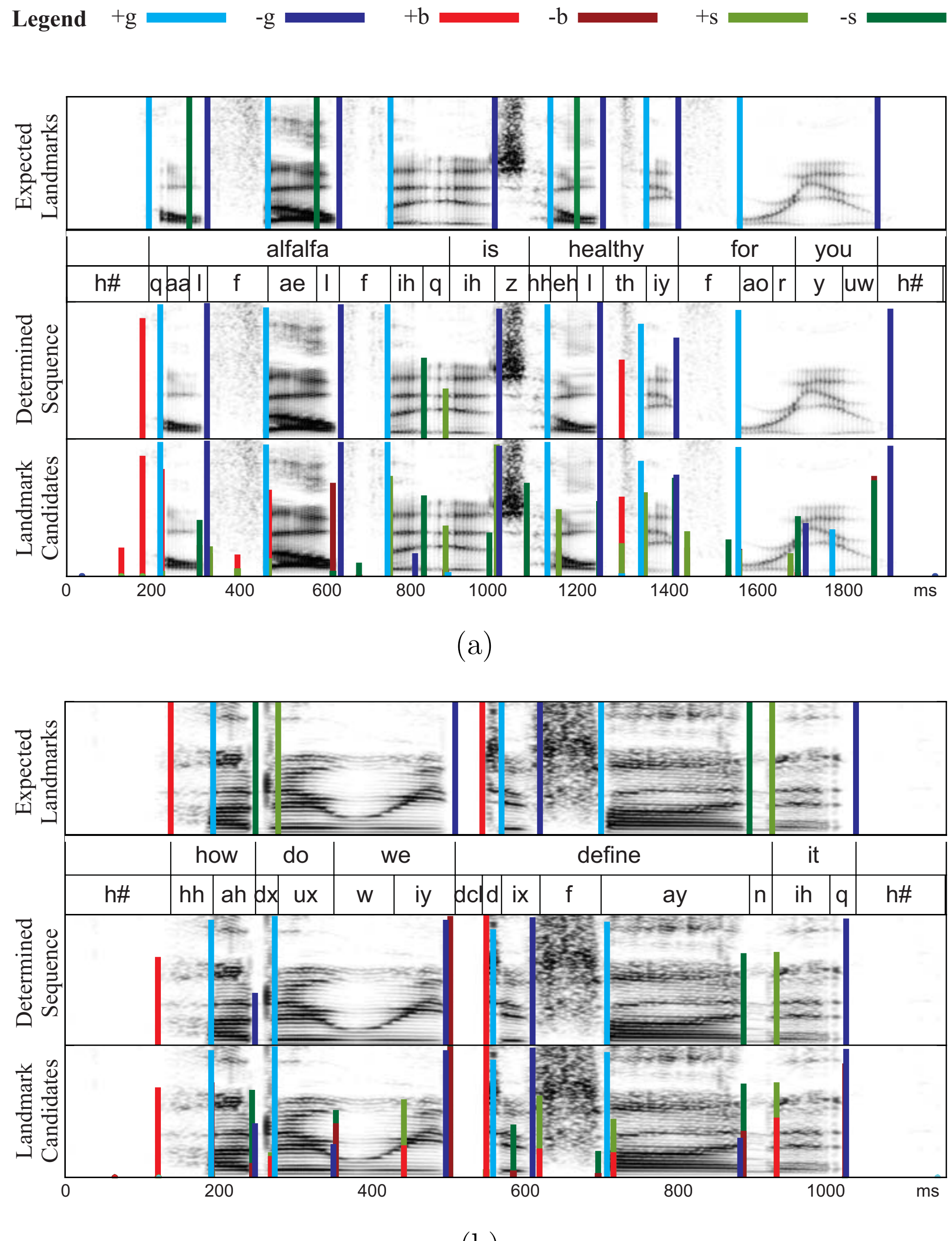
6.6 Analysis of Variable Contexts

Some examples of detected landmark sequences are shown in Figure 6-7. The examples are selected so as to include as many types of errors as possible. Although the overall error rate of the landmark sequence detection is high, the locations of most of the important acoustic events are identified either as correct landmark detections or as substitutions of a different landmark type. It should be noted that the substitutions and insertions do not occur randomly, but most of them occur in similar contexts indicating that these error patterns may reflect additional acoustic information which can be made use of.

The contexts of the common errors are analyzed and classified in this section. This analysis will provide knowledge about the acoustic properties of landmarks in more depth, and about the possible variation in the realization of landmarks in different contexts. Eventually this information can be applied in utilizing the systematic wrongly-detected landmarks as the system becomes more knowledge-based.

6.6.1 Error Type 1: Variants

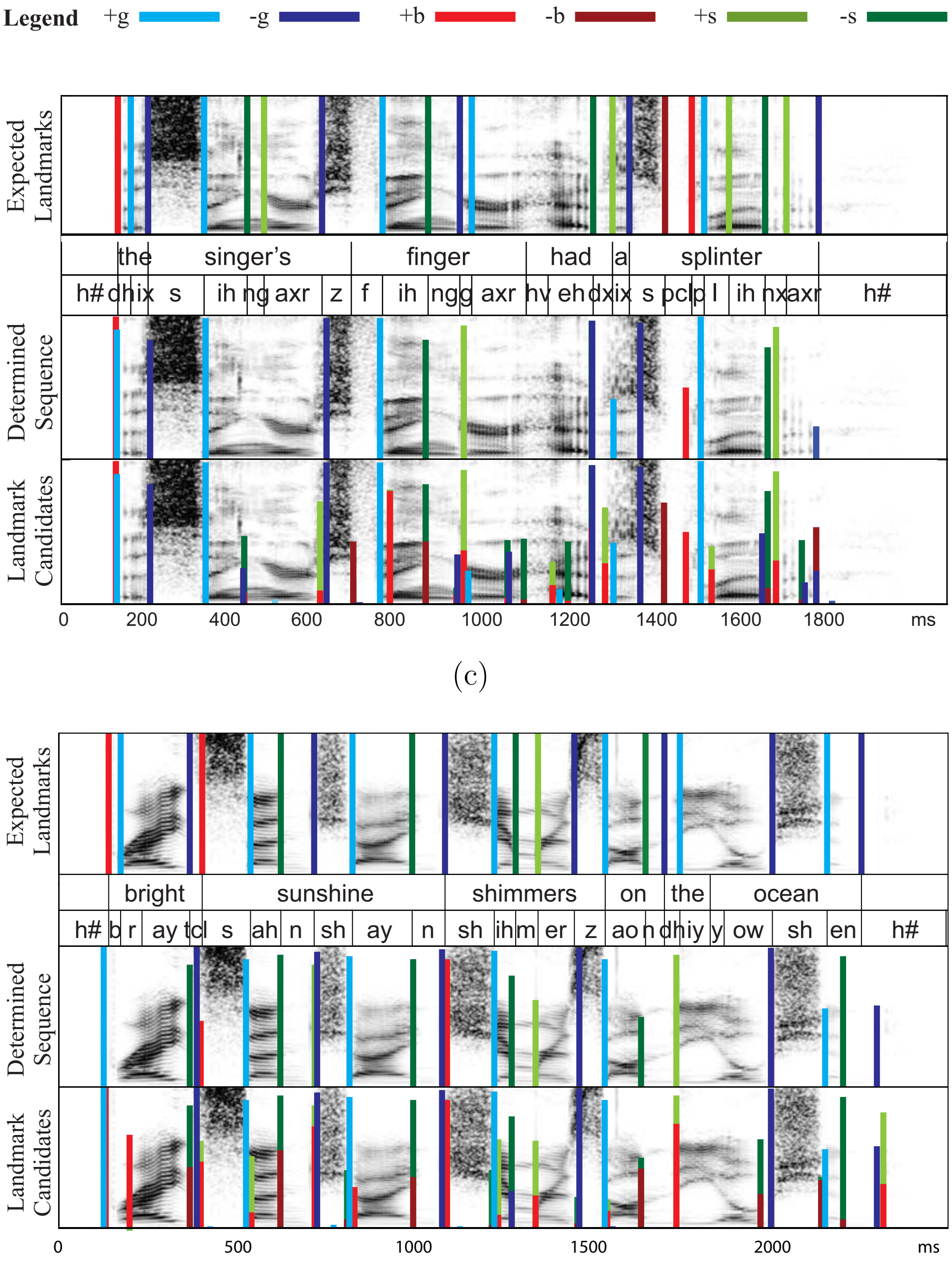
Landmarks are closely related to the manner features of pronounced words, but they are mainly defined and detected based on the acoustic properties of the signal. Therefore, when the realized acoustic property is different from the typical pronunciation of a phoneme, the detected landmark type should be different accordingly. The errors collected in this section are due to the inconsistency between the phonetic transcrip-



(a)

(b)

Figure 6-7: Examples of the results of landmark sequence determination



(c)

(d)

Figure 6-7: Examples of the results of landmark sequence determination (Continued)

tion and the realized pronunciation.

Flaps

One of the most common examples of such variants is a flap sound. In many contexts, the stop /t/ or /d/ sound is usually pronounced with a complete closure in the oral tract along with the offset of glottal vibration (−g landmark), followed by a sudden release of burst noise (+b landmark) and eventually into the onset of another vowel sound (+g landmark). In conversational speech, however, when a /t/ sound is located between a stressed and a reduced vowel, as in the word 'bottom' or 'muddy', it is usually realized by rapidly tapping the tongue against the roof of the mouth, instead of making a complete closure with pressure buildup inside the mouth. Therefore, it does not show a strong burst noise, and voicing is not suppressed during the flap sound. This means that flaps are realized with properties closer to sonorant consonants, and the expected landmarks for a flap sound are defined to be a −s landmark at the closure and +s landmark at the onset of the following vowel.

The degree of closure for an alveolar stop realized a flap can vary very widely. An extreme example is shown in Figure 6-7(b). The flapped /d/ of the word 'do' has very short closure duration, but the voicing is almost turned off during the closure and the burst noise is present. Another example is shown in Figure 6-7(c), where the flapped /d/ of the word 'had a' is flapped and the burst noise is not present, but because the voicing is suppressed during the closure, a (−g, +g) landmark sequence is detected instead of the expected (−s, +s) landmarks.

Due to such variations in the realization of flap sounds, almost half of the flaps, as hand-transcribed in the TIMIT database, are detected as g-landmarks by our algorithm, due to the large drop of energy in the lower frequency band.

Syllabic Sonorants

About 80% of syllable-final /l/ sounds do not have abruptness. Figure 6-7(a) gives an example with such /l/ sounds. When syllable-final /l/ follows a back vowel, such as in the word 'ball' or 'small', the lateral sound is pronounced with the backing of the tongue, instead of the tongue tip touching the roof of the mouth. As a result, many of these syllable-final /l/'s do not create abruptness in the spectrogram. Because the TIMIT database does not have a distinctive symbol for the non-abrupt /l/, these are viewed as deletion errors of s-landmarks, even though no s-landmark actually occurred at the time-point.

Figure 6-7(d) shows a case with −s landmark insertion in the middle of a syllabic nasal. Some of the syllabic nasals, as transcribed in TIMIT database, are not nasalized during the whole period, and the transition due to the opening of the velopharyngeal port can be identified with a spectral abruptness.

Nasalization

Nasalization is defined as the opening of the velum during voicing, creating a nasal pole-zero pair and a distinct acoustic signature. One of the commonly occurring contexts for nasalization is that of the /ð/ sound. Figure 6-7(d) shows a token of nasalization of /ð/ in the word sequence 'on the'. According to its feature specification, the /ð/ sound should be pronounced with frication noise along with the suppression of glottal vibration due to the raised oral pressure, resulting in (−g, +g) landmark pair, but occasionally, the velopharyngeal port does not get completely closed during the /ð/ sound due to the nasal sound in front of it, and the realized landmarks are −s at the closure of /n/ sound, and +s at the release of /ð/ sound.

A similar phenomenon occurs with stop consonants following a nasal, such as in the case of 'finger' or 'number'. Figure 6-7(c) shows an example of the word 'finger'. The expected landmark sequence for a nasal followed by a stop is (−s, −g, +b, +g), in which −s and −g represent the closure of nasal and stop respectively; +b denotes the onset of the burst noise (although this is often deleted after a nasal), and +g represents the onset of the vowel. In some cases, however, when a voiced stop consonant follows a nasal, the velopharyngeal port does not get closed completely during the stop closure, which leads to the deletion of both g-landmarks. In many cases +s is detected in the place of +g landmark, marking the onset of the vowel.

6.6.2 Error Type 2: Insertions

Because one of the purposes of detecting landmarks is to find the locations such that important acoustic information can be found nearby, insertion errors do not pose serious threats in the processes that follow; they merely increase the number of locations that need to be processed further. This section collects some of the common insertion errors.

Insertions Indicating Additional Information

Some of the insertions do indicate the occurrence of informative acoustic events, but they are classified as insertions because the events are not as consistent or abrupt as might be defined as typical characteristics of landmarks.

For example, glides sometimes provoke the detection of s-landmarks. The /w/ sound of the word ‘we’ in Figure 6-7(b) does show s-landmarks with more than 0.5 probability at the boundaries between the glide and adjacent vowels, although they are not selected in the landmark sequence determination process.

In addition, regions of irregular pitch periods in a vowel can result in s-landmarks, as can be seen in Figure 6-7(a) between the words ‘alfalfa’ and ‘is’. The boundaries between this region and the neighboring vowels do not show abrupt changes, except that the distance between the pitches becomes greater during the irregularity. However, due to the wide window size in the coarse pass of the landmark candidate detection process, the irregular pitch periods are considered as a period of low energy, and s-landmark candidates are apt to be found at the boundary where modal phonation begins.

The +b landmark detected at the start of the utterance in Figure 6-7(a) also indicates an additional characteristic of the speech signal. When a sentence or a phrase starts with a vowel, a glottal stop is sometimes inserted at the vowel onset. The b-landmark captures this insertion of the glottal stop, even though it does not reflect a burst after a stop consonant.

A fricative may introduce a period of pause, or epenthetic silence, as shown in the

/θ/ sound in the word 'healthy' in Figure 6-7(a). A b-landmark is *correctly* detected at the boundary between the silent region and the frication, although this was not originally expected from the phonemic transcription.

Insertions Near Other Landmarks

The onsets and offsets of vocal fold vibration are detected from the abruptness in the 0–400Hz frequency range, but these onsets and offsets also create abrupt changes across the entire frequency range, which includes the range used in the detection of b- and s-landmarks. As a result, b- and s-landmarks tend to be found right next to g-landmarks, where they are not predicted and do not correspond to separate articulatory events.

An example can be found at the end of the word 'we' in Figure 6-7(b). A $-$b landmark is detected with a high probability, which is understandable because the high-frequency energy drops down abruptly, and a silent region follows the time-point.

Since the locations of such inserted landmarks are usually at the same positions as g-landmarks, these insertions do not increase the number of locations to be further processed. In addition, because more than 70% of these false alarms occur within 7ms of g-landmark locations, most insertions of this type can be eliminated (without loss of information) simply by ignoring the b- and s-landmarks located within 7ms of g-landmarks. A simple revision of the algorithm that used a hard threshold of 7ms minimum distance between b- and g-landmarks could reduce the insertion rate of the b-landmarks by half.

6.6.3 Error Type 3: Deletions

The most serious types of errors are deletions, because this means that the acoustic event cannot be located with any of the detected landmarks. Some of the deletion errors are due to insufficient change in the signal, short duration between landmarks, or unexpected noise in the signal.

End of Utterance

The subglottal pressure tends to drop near the end of an utterance. This leads to less energy in the vowel sound, which lowers the probability of s- and g-landmarks. The examples given in Figure 6-7 are short and do not show much effect of the lowered subglottal pressure, but Figures 6-7(c) and (d) show some tapering of the probabilities toward the ends of utterances.

Therefore, the landmarks—especially s-landmarks due to the lack of abruptness—are likely to be deleted at the end of an utterance. It might be possible to compensate for the effect by analyzing the envelope of the overall energy level and using this information to compensate for the reduced abruptness.

Error Propagation

It was shown in Figure 6-6 that the application of bigram constraints enables us to detect a low-probability candidate if the adjacent candidates have high probabilities. In the opposite manner, when a deletion error occurs in the candidate detection step or when the probability of a landmark is incorrectly estimated to be very low, this error can propagate to adjacent landmarks, and the nearby landmarks can also be deleted in the landmark sequence determination stage because the sequence of landmarks has a strict grammar structure.

However, this type of error is mostly local, that is, only one adjacent landmark is affected due to the error propagation in most cases. Deletion of a g-landmark, for example, occurs mostly when the g-landmark marks a boundary of a schwa. Therefore, the g-landmark error will propagate only during the period of the schwa. When a b-landmark is deleted, it will have an affect only in that obstruent consonant cluster, but most of the obstruent consonants do not appear in clusters, and fricative-fricative sequences do not produce any landmarks in between unless there is a silent region between them. Therefore, in most cases a b-landmark deletion does not propagate at all. An s-landmark occurs next to a sonorant consonant and nasal clusters are rare, so this error will propagate only within the consonant.

In addition, even when the true landmark is completely deleted in the signal, another landmark of a different type is often detected in many cases, which leads to a substitution instead of deletion. As we have seen, substitution errors can be interpreted as information about the signal's characteristic.

Chapter 7

The Representation of Reliable and Ambiguous Regions

7.1 Introduction

7.1.1 Motivation

The method described in the previous chapter determines the most likely landmark sequence from the set of independently-detected landmark candidates by means of individual probability and transitional probability. It was established that the correct landmarks can be detected by means of their individual probabilities, and that further restriction using the bigram transitions helped in rejecting most of the wrongly detected insertions.

However, because this method determines the single most likely landmark sequence, landmarks that are realized with some level of uncertainty are sometimes deleted or substituted with different landmark types. This results in the loss of already-detected information because as was discussed in the previous chapter, most of the unclearly detected landmarks are not random errors. Instead, they occur in a limited set of contexts, such as nasalization of obstruent consonants, flaps, or sonorant-like realization of fricatives. Such contexts accompany alternative or unclear realization of distinctive feature cues due to overlapping of articulatory gestures, or incomplete closures and releases.

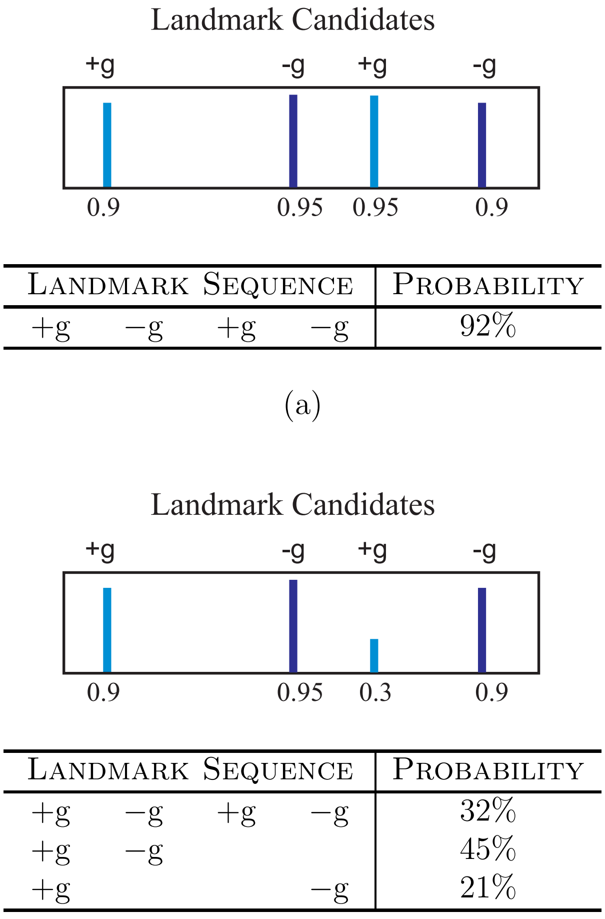
It was also noted that in those unclear instances, some of the actual landmarks had been detected during the individual landmark candidate detection process, but failed to be recognized during the landmark sequence determination process because the probability of the landmark sequence was not as high as the alternative choice.

Therefore, by locating the region where the landmarks are detected with a certain level of uncertainty, more careful inspection can be applied to the ambiguous region, so that the contexts of the region can be determined. Then, the cues that provide the most information in the given contexts can be used to disambiguate the existence of landmarks and the types of landmarks.

Another motivation for the separation of reliably detected landmarks from ambiguously detected ones is the extreme distribution of probabilities. As can be observed in the probability distribution of landmark candidates shown in Chapter 5, extremely high probabilities are assigned to most of the correctly detected landmarks; that is, more than 80% of the correctly detected g-landmark candidates have a probability greater than 0.9. When a series of landmark candidates, which do not violate the bigram transition restriction, is detected with such clarity, the landmark sequence can be decided with confidence.

Figure 7-1 gives a simple illustration of a series of reliably detected landmark candidates, as opposed to a landmark sequence that contains an uncertain candidate. If a sequence of candidates has high probability, then all the candidates in the sequence are likely to be the landmarks of the actual signal. However, when one or more of the candidates in a region are detected with low probability, the sequence determination is less clear.

Because most of the correct landmark candidates have very high probabilities and most of the false alarms are detected with low probabilities, it is likely that the individual landmark detection process will result in many sequences of candidates in which all the candidates have either extremely high or low probabilities. In such regions, the landmark sequence determination process will be able to determine the true landmark sequence with more confidence. However, some landmark candidates



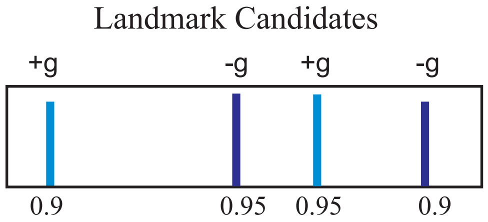
Landmark Sequence				Probability
+g	−g	+g	−g	92%

(a)

Landmark Sequence				Probability
+g	−g	+g	−g	32%
+g	−g			45%
+g			−g	21%

(b)

Figure 7-1: An example of a series of landmark candidates detected with high probabilities (top), and one with an uncertain candidate (bottom)



Landmark Sequence				Probability
+g	−g	+g	−g	92%
+g	−g			3%
+g			−g	2%
		+g	−g	3%

Figure 7-2: Unless all the landmark candidates have probability of one, there might be other alternatives, although with minuscule possibilities.

will be detected with intermediate probabilities, and in these regions, where there are candidates with less extreme probabilities, confidence will be lower.

When an utterance is pronounced clearly, the speaker produces unambiguous cues that make clear distinctions between features. This means that landmarks, which are a special class of cues, should become more prominent as well. Therefore, it can be expected that if one can identify the regions where the landmarks are detected reliably, then it would be possible to estimate the distinctive features of the sounds in those regions with more certainty.

7.1.2 Reliable and Ambiguous Regions

In this section, the terms *reliable regions* and *ambiguous regions* are defined.

A reliable region is defined to be a portion of the signal where there is only one likely landmark sequence. However, as Figure 7-2 shows, unless all the landmarks are detected with probability of exactly one, there will always be some other alternatives, however unlikely they may be. Therefore, a less stringent criterion that distinguishes reliable regions from ambiguous regions must be determined.

Once the clarification of a reliable region is in place, an ambiguous region can be simply defined to be the parts of the speech signal that do not meet the criteria for reliable regions. In other words, it means an interval where one cannot confidently say that a certain landmark sequence should be the true landmark sequence; instead, there may be two or more alternative choices. For example, a flapped /t/ sound is usually realized with either a (−g, +g) or a (−s, +s) landmark pair, and when the closure is not made clearly, both alternatives are likely to appear in the landmark sequence determination process.

In the reliable regions, the determined landmark sequences can be trusted, and can be the center of focus in later stages of processing. The uncertain regions, on the other hand, can provide some useful knowledge about the local signal by pointing out the possible alternatives. For example, when two possible alternatives for a region are (−g, +g) and (−g, +b, +g) landmark sequences, then it can be concluded that the region is obstruent. In other cases, an uncertain region can indicate which additional cues need to be detected around the region to assure correct landmark detection. For instance, when the landmark pairs (−s, +s) and (−g, +g) are two possible alternatives of a region, the correct landmark sequence can be determined by verifying the existence of voicing within the region.

This chapter will focus on finding the regions where the landmarks are detected reliably as opposed to the regions where the landmark detection is ambiguous, adding a third processing steps to the previous steps of individual landmark detection and landmark sequence determination described earlier.

7.2 Graph Representation

7.2.1 Target Representation

There are two main problems that need to be solved in the representation of reliable and ambiguous regions. One is how to tell the ambiguous regions from reliable regions, and the other is how to express alternative choices of landmark sequences in each ambiguous region.

One of the approaches that would solve such problems is to apply an N-best

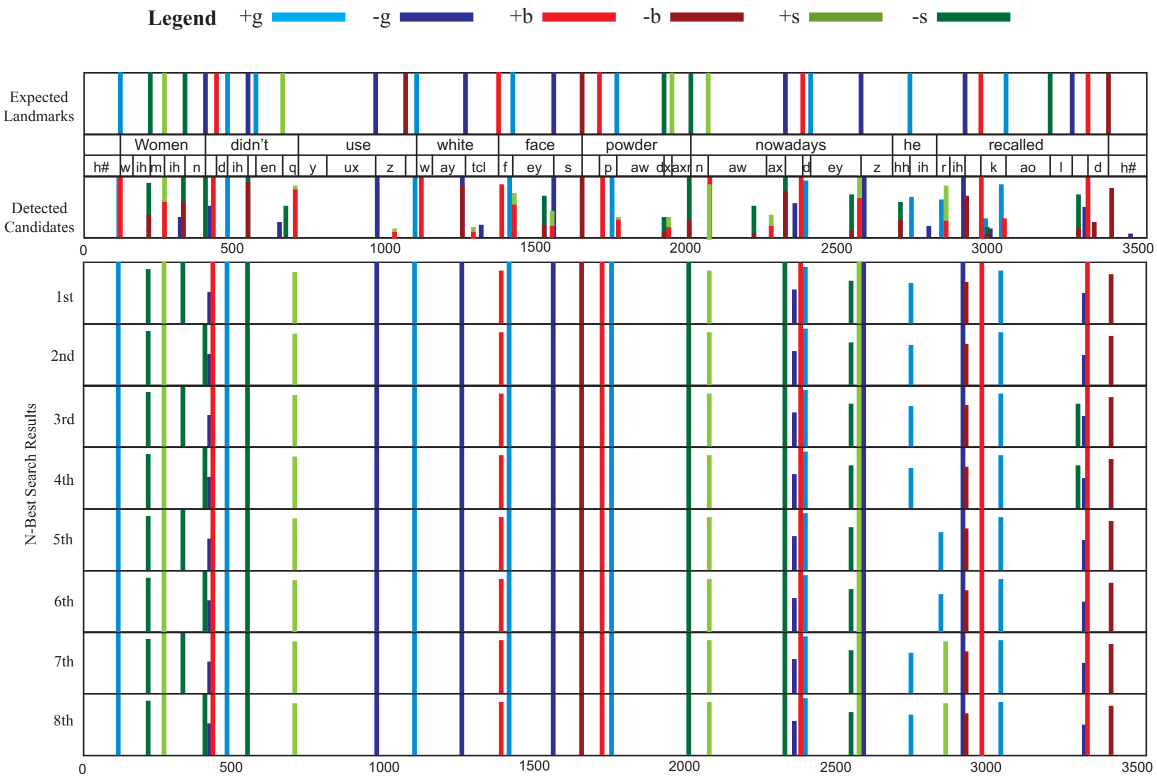


Figure 7-3: N-best result of the landmark sequence determination process

search algorithm instead of the Viterbi search [56]. By applying this approach, the true landmark sequence may be detected as one of the N most likely alternative choices. Figure 7-3 shows an example of the N-best result of a speech signal. The top tier shows the labeled landmarks, and the landmark candidates that are detected from the landmark detection process in Chapter 5 are shown below the phonetic transcription. The eight most likely sequences of landmarks determined by N-best search algorithm are shown in the last eight rows. From this example, it can be said that the region between 500–2500ms has reliably detected landmarks, whereas there might be three alternatives in the 2700–2900ms region.

However, the major drawback of this approach is that the value of N must be determined beforehand and the size of N may need to be increased exponentially with respect to the length of the signal in order to include all possible alternative choices. In addition, as the size of N increases, the number of computations and the resulting

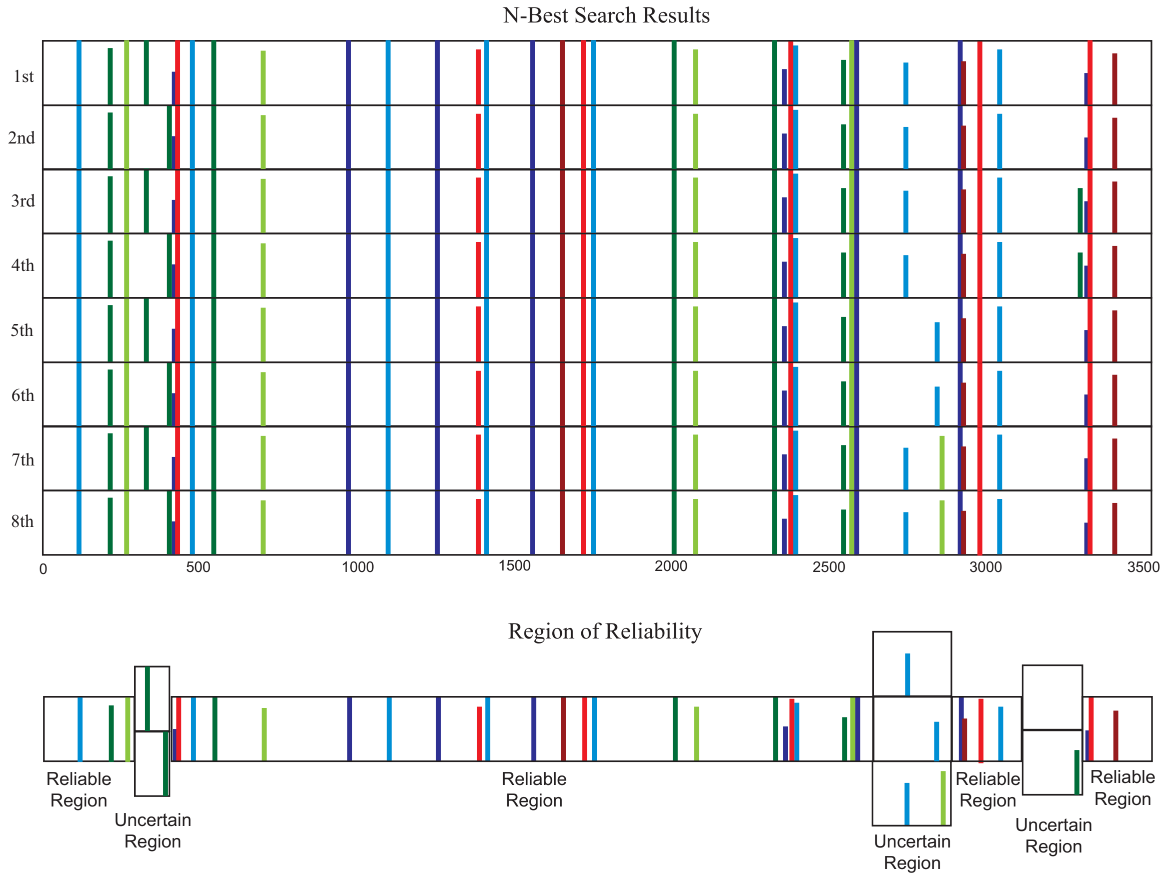


Figure 7-4: Comparison of an N-best representation and its corresponding reliability representation

set of landmark sequences would have to become as large, but the information the result represents does not increase as much. As can be seen in Figure 7-3, an N-best algorithm introduces a large amount of redundant information; the resulting landmark sequences are almost the same except for a small number of changes. To avoid this problem, it would be more ideal if the different alternatives are all collapsed so that the speech signal can be separately represented as *Reliable Regions* and *Uncertain Regions* as shown in Figure 7-4.

This representation of alternative landmark selection results has the advantage that the size of the representation is much more compact, resulting in a reduction of the redundancy in the N-best representation. The classification of reliable regions and uncertain regions is visually presented, and alternative choices for each ambiguous region are clearly identified. These are desired properties for the target representation.

Although the size of the representation could be reduced by this method, the problem of determining the size of N and the exponentially increasing complexity of N-best search algorithm cannot be resolved. In addition, the likelihood of each alternative choice in an ambiguous region cannot be reliably calculated by simply compressing the N-best search results.

Therefore, in this chapter, a different approach is proposed to generate the same representation, using a graph pruning method instead of utilizing an N-best search algorithm.

7.2.2 Graph Construction

A method that generates a weighted directed acyclic graph from a set of detected landmark candidates based on the bigram constraints and individual probabilities was proposed in Chapter 6. A simple description of the construction method is revisited below.

- **Nodes:** The nodes correspond to all the landmark candidates arranged in serial order, and two additional nodes are added which denote the start and end of an utterance—marked as u_{st} and u_{end}, respectively.

- **Edges:** Edges connect all the *legal* pairs of landmarks. The direction of the edge follows the serial order. 'Legal' means that the bigram probability between two landmarks is non-zero. The weight on each edge is calculated by the product of the bigram transition probability and the probabilities of all the candidates in between.

Each path from the start node (u_{st}) and the end node (u_{end}) represents a selection of a landmark sequence that does not violate the bigram transition rules. In the previous chapter, the Viterbi search algorithm was applied to this graph, so that the most likely landmark sequence could be determined by finding the path with the largest weight.

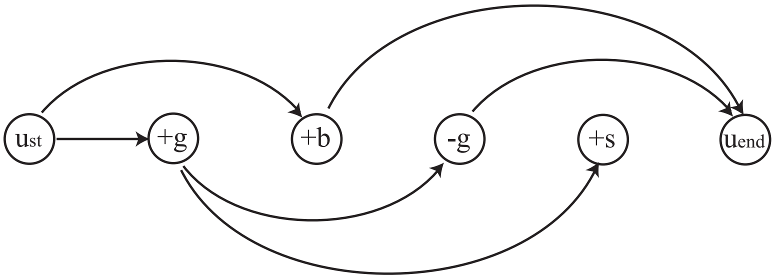


Figure 7-5: A simple example for the explanation of node pruning criterion. Omission of nodes strictly depends on the structure of the graph, and does not affect the probability of resulting paths at all.

In this section, however, the method of pruning the least likely possibilities is adopted instead of finding the most likely paths. This approach can preserve alternative possibilities of landmark sequence determination, while reducing unlikely instances at the same time.

7.2.3 Pruning Nodes

The pruning of a graph can be performed in two ways—pruning of nodes and of edges. The criterion of node pruning depends on the structure of the graph itself. Figure 7-5 shows an example of such a case. The +s node, located right before the last node, only has an edge directing to it but no edge starting from it. Therefore, no path from u_{st} to u_{end} can visit the +s node, and so deleting this node can will not affect the possible landmark sequence determination results. This can be interpreted as the following: because it is not possible to end an utterance with a +s landmark, the +s landmark can be disregarded as a candidate for the final landmark.

A generalized criterion for node pruning can be written as the following: When a node, which is not a start or an end node, has no incoming edges or no outgoing edges, the node can be deleted from the graph without affecting the search result from the start node to the end node.

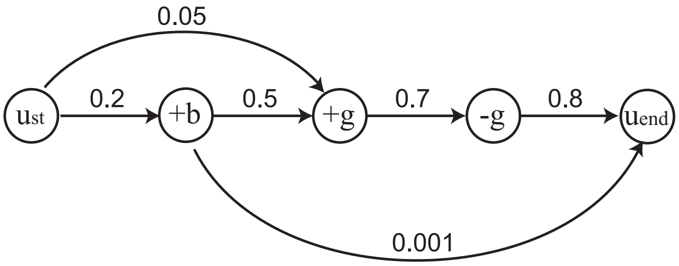


Figure 7-6: A simple illustration of edge pruning criterion. Deleting edges may change the probability of possible paths, but the product of weights on each edge remains the same.

7.2.4 Pruning Edges

Unlike the case of node pruning, some of the edge pruning may affect possible paths from the start node to the end node, and it may change the probability of each path, however small the effect may be. Figure 7-6 gives an example of when an edge may be safely pruned out. The edge connecting the +b node to u_{end} has much smaller weight than the other nodes. Therefore, it is highly unlikely that the resulting path will contain this edge, and so this edge can be pruned out without affecting the search result significantly.

However, the value of the edge weight alone cannot be a reasonable measure to prune out an edge. For example, in Figure 7-7, all three possible paths have the same product of weights, but the weights on individual edges vary from 0.001 to 0.9. Therefore, edge pruning should depend on the *probability* of each edge being in the resulting landmark sequence, rather than on individual weights. For example, even if the edge between u_{st} and the first +g node has larger weight than the edge between u_{st} and the third +g node, the probability of the former being selected should be the same as that of the latter because the weight of the following edge is much smaller. A clearer definition of this probability measure will be given in more detail in the next section.

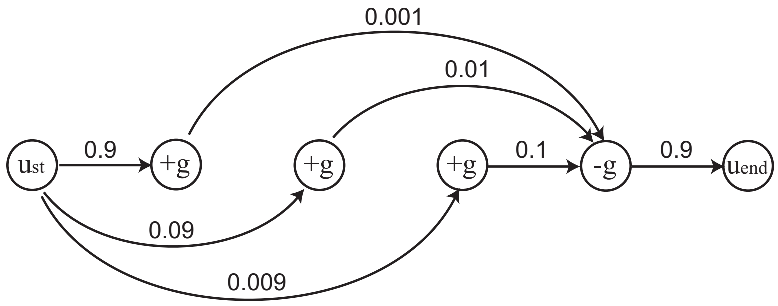


Figure 7-7: A graph with multiple paths with the same product of weights

7.3 Edge Probability

7.3.1 Definition

Some of the terms that will be used throughout this chapter are clarified in this section.

The *edge weight*, or the *weight of an edge*, is defined to be the weight calculated from the bigram transition probabilities and individual probabilities, as described in Section 6.3.2.

The *product of weight* of a path, or *score* of a path, is calculated by multiplying the weight on all the edges that the path traverses. This is the same as the overall score (see Section 6.3.1) of the landmark sequence determined by the sequence of nodes that the path visits.

The *path probability*, or the *probability of a path*, is the normalized value of the path score, such that the probabilities of all possible paths from the start node to the end node sum up to one. The normalization can be simply described with the following formula:

$$\text{Normalized Probability} = \frac{\text{Product of Weights on a Path}}{\text{Sum of Products of Weights on All Possible Paths}}$$

Figure 7-8(a) gives an example of the path scores (products of weights) and path probabilities of a graph. The numbers written on the edges represent the edge weights.

There are four possible paths going from the first node to the last. All possible paths from node 1 to node 5 and the products of weights on the paths are tabulated below the graph.

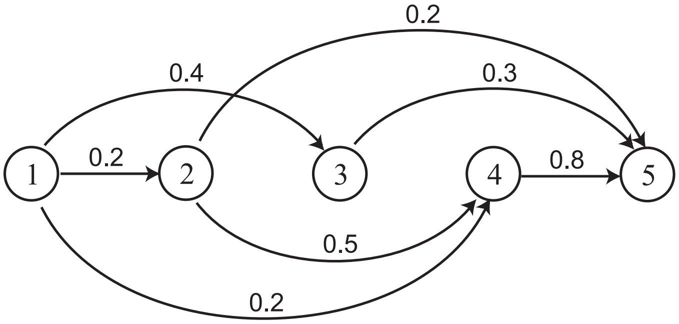
The *probability of an edge*, or the *edge probability*, is defined to be the sum of the probabilities of all the paths that traverse the edge. For example, the edge 1–2 is traversed by two different paths 1–2–4–5 and 1–2–5, so the probability of the edge 1–2 being selected as the resulting landmark sequence is calculated as the sum of the probabilities of both paths: $0.2 + 0.1 = 0.3$. Because the sum of all possible paths is defined to be one, an edge that is traversed by all possible paths will have probability of one, and an edge that can never be visited will have zero probability. The list of edges and their probabilities are shown in Figure 7-8(b).

Figure 7-9 shows the edge probabilities of the graph given in Figure 7-7. As was expected, all the possible edges that branch from the first node have the same probability, even though the weights on the edges vary widely. Note that the edge between −g and the following u_{end} node has probability of one, implying that this edge will be always traversed no matter which path is selected. This means that the −g landmark will always be selected.

7.3.2 Properties

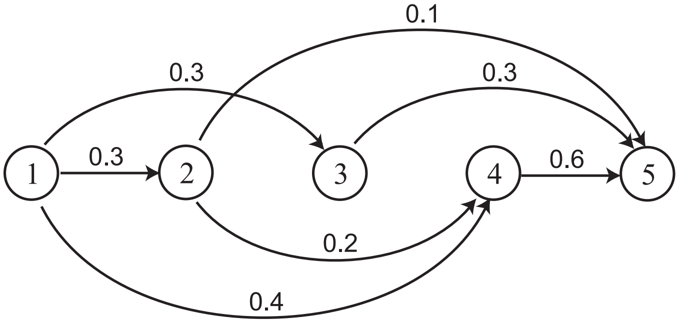
Note that the previously described method that calculates the edge probabilities is not an efficient one, since it has to find all possible paths, and calculate the product of the weights in each path. The number of all possible paths is expected to be about the order of $O(2^n)$, which grows exponentially with respect to the number of nodes. Therefore, a more efficient method that calculates the edge probabilities should be developed. As a preliminary step, some of the properties of the edge probability are examined in this section.

Property 1. *The probabilities of all the edges that start from the start node sum up to one. The probabilities of all the edges that point to the end node also sum up to one.*



Path	Product of Weights	Probability
1–2–4–5	0.08	0.2
1–2–5	0.04	0.1
1–3–5	0.12	0.3
1–4–5	0.16	0.4

(a)



Edge	Traversing Paths	Edge Probability
1–2	1–2–4–5, 1–2–5	0.3
1–3	1–3–5	0.3
1–4	1–4–5	0.4
2–4	1–2–4–5	0.2
2–5	1–2–5	0.1
3–5	1–3–5	0.3
4–5	1–2–4–5, 1–4–5	0.6

(b)

Figure 7-8: Illustration explaining the definition of edge probabilities. (a) shows a simple graph with weighted directed edges. All possible paths of the graph and the products of weights of the paths are listed below it. The probability is the normalized value of the product of weights. (b) shows the calculated edge probabilities of the graph above.

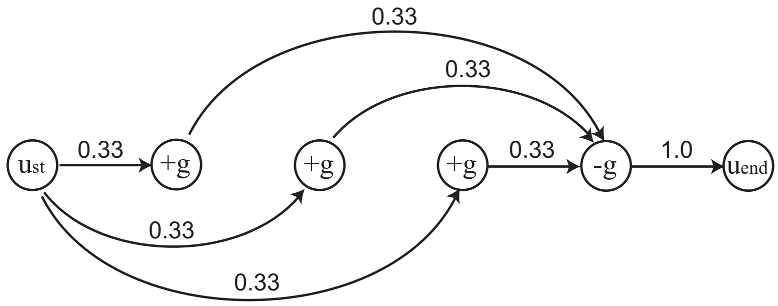


Figure 7-9: The edge probabilities of the graph in Figure 7-7

Proof. Let's say that n edges, $e_1, e_2, \cdots, e_n$, start from the start nodes, and the sets P_i for $i = 1$ to n be the sets of all paths from the start node to the end node which traverse the edge e_i. Then, it is obvious that the n sets are mutually exclusive and each of the paths from the start node to the end node must be a member of exactly one of the n sets.

By the definition of edge probability, the edge probability of e_i is determined to be the sum of all path probabilities of the paths in the set P_i. Because the n sets, $P_1, P_2, \cdots, P_n$, contain each path exactly once, the sum of the edge probabilities of $e_1, e_2, \cdots, e_n$ must be the same as the sum of the path probabilities of all the paths from the start node to the end node, which is one, according to the definition of path probability.

Therefore, the probabilities of all the edges that start from the start node sum to one. By a similar argument, it can be concluded that the probabilities of all the edges that point to the end node sums to one as well. □

Property 2. *For each intermediate node, which is not the start node or the end node, the sum of probabilities of all the incoming edges is the same as the sum of probabilities of all the outgoing edges.*

Proof. Let's say that an intermediate node v has n incoming edges, $e_1, e_2, \cdots, e_n$. Because all the paths that go through v must traverse exactly one of the n incoming edges, the sum of edge probabilities of $e_1, e_2, \cdots, e_n$ is the same as the sum of the path probabilities of all paths that go through the node v.

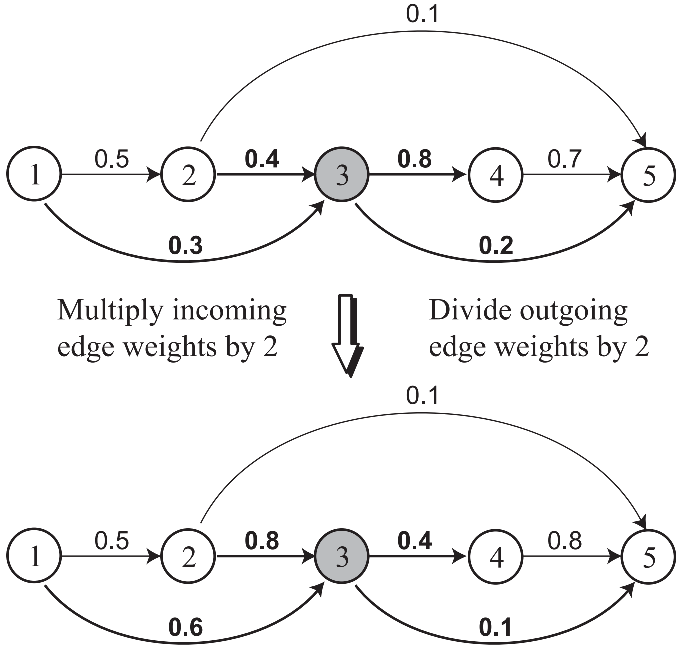

0.1
1
0.5
2
0.4
3
0.8
4
0.7
5
0.3
0.2
Multiply incoming edge weights by 2
Divide outgoing edge weights by 2
0.1
1
0.5
2
0.8
3
0.4
4
0.8
5
0.6
0.1

Figure 7-10: An example of the operation that preserves edge probabilities. The weights on the incoming edges of node 3 are multiplied by 2, and the weights in the outgoing edges of node 3 are divided by 2.

By similar reasoning, it can be shown that the sum of the path probabilities of all the outgoing edges of the node v is also the same as the sum of the path probabilities of all paths that go through the node v.

Therefore, the sum of probabilities of all the incoming edges and that of the outgoing edges are equivalent. □

The next property defines an operation on the edge weights that does not affect the path probabilities. An example of the operation is illustrated in Figure 7-10.

Property 3. [Probability-preserving Operation] *For a given node, when the weights of all the incoming edges are multiplied by a value x and the weights of all the outgoing edges are divided by the same value x, the probability of each path remains the same.*

Proof. Assume that the target node is neither the start node nor the end node. Then, the paths from the start node to the end node can be classified into one of the following two types: the ones that go through the target node v, and the ones that do not visit the node v.

When a path visits the target node v, then the path must traverse exactly one of the incoming edges and one of the outgoing edges, because no nodes can be traversed more than once. Therefore, the product of weights on this path is multiplied by x and divided by the same value x, which results in the same path score. On the other hand, when a path does not visit the target node v, then the path does not traverse any of the edges with modified weights. Therefore, the product of weights of this path is unchanged.

Therefore, if the operation is performed on an intermediate node, the path probability remains the same because the product of weights is preserved.

However, if the target node is the start node, all the paths from the start node to the end node must traverse exactly one of the outgoing edges. Therefore, the products of weights of all the paths are divided by the value x. Because all the path scores are evenly divided by the same factor, the normalized path probability stays the same. The same argument holds for the case when the target node is the end node. □

Corollary 1. *The operation described in Property 3 preserves edge probabilities.*

Proof. The edge probabilities are defined based on the path probabilities, and the path probabilities are invariant under the operation. Therefore, the edge probabilities are invariant as well. □

7.3.3 Efficient Algorithm

Based on the properties observed in the previous section, a more efficient algorithm is developed. The algorithm has two-steps; the first step changes the weights according to the probability-preserving operation explained before, so that the overall edge probabilities do not change. The second step calculates the edge probabilities using the property that the sum of incoming edge probabilities should be the same as the sum of outgoing edge probabilities.

The pseudo-code of the overall procedure is given below, and a step-by-step illustration of an example is shown in Figure 7-11.

- **Backward Step**

  Starting from u_{end} and going backward to u_{st}, do the following for each node i.

  - Let S be the sum of the weights on all the edges stemming from node i.
  - Divide the weights on all the outgoing edges of node i by S.
  - Multiply the weights on all the incoming edges of node i by S.

- **Forward Step**

  Starting from u_{st} to u_{end}, do the following for each node i.

  - Let S be the sum all the weights of the incoming edges to node j.
  - Multiply the weights of all the outgoing edges from node j by S.

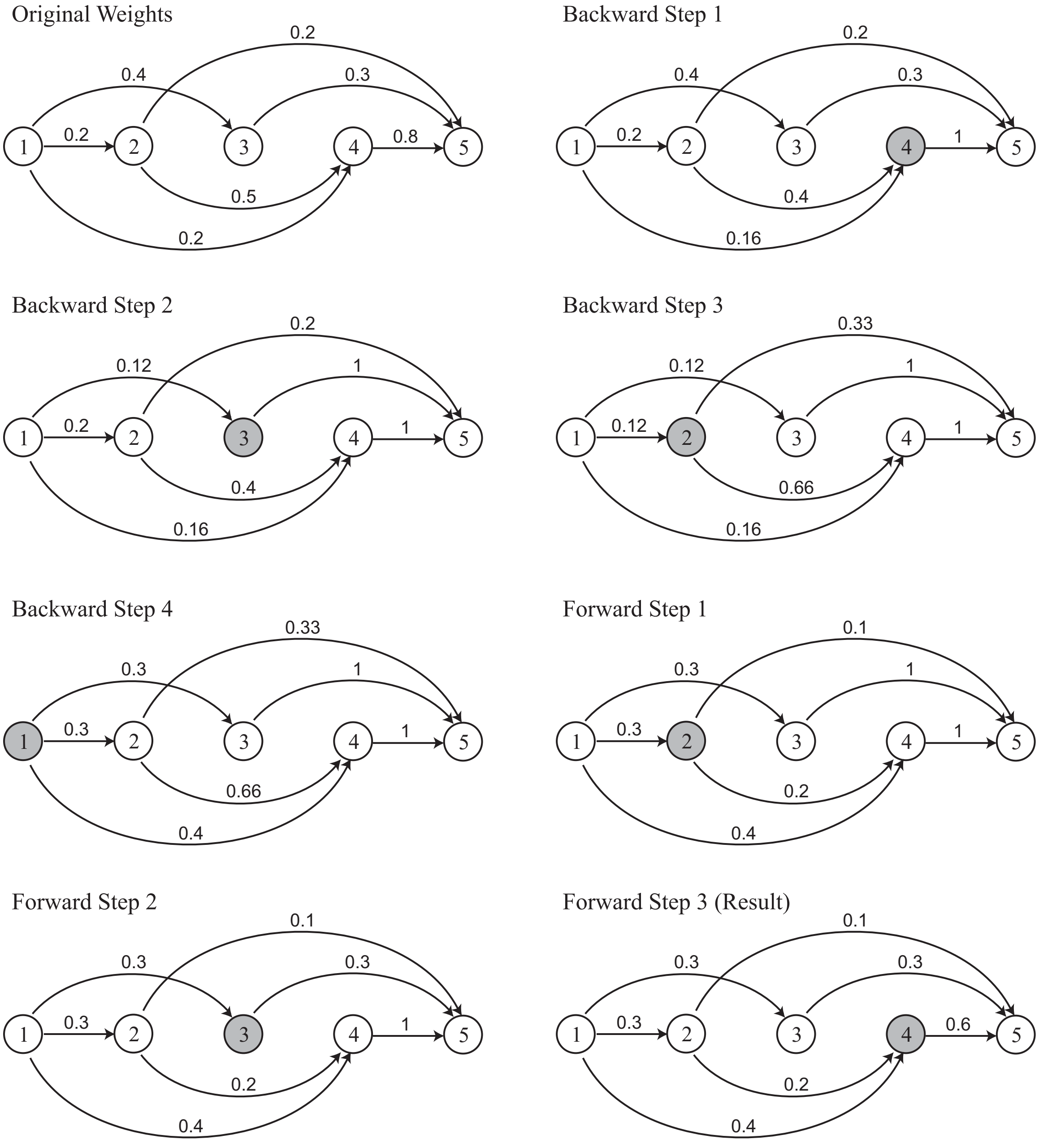
Note that the backward step normalizes the outgoing edge weights of each node, so that the sum of the weights on outgoing edges of a node equals to one. In addition, because each backward step uses the probability-preserving operation, the change in the weight does not affect the resulting edge probabilities.

Because each step is applied backward starting from the last node to prior nodes, it is also true that the products of weights on all the paths from any one node to the last node always sum up to one, assuming that there are paths connecting from the node to the last node. A detailed proof is given below:

Claim 1. *After the backward step of the algorithm is performed, the products of weights on all the paths starting from any one node v to the end node always sum up to one, assuming that there are paths connecting from the node to the end node.*

Proof. This claim can be proved by a simple mathematical induction.

Assume that there is a graph with n nodes, and that all the nodes that do not have any possible path to the end node have already been removed. Then, there must be an edge that connects the $(n-1)^{st}$ node to the end node, because there must be a path between them. Since all the edges should connect from prior nodes to later nodes, this edge is the only path that starts from $(n-1)^{st}$ node. Therefore, after the backward step, the edge weight of this edge must have become one.



Original Weights

Backward Step 1

Backward Step 2

Backward Step 3

Backward Step 4

Forward Step 1

Forward Step 2

Forward Step 3 (Result)

Figure 7-11: Step-by-step illustration of the probability calculation algorithm

Now, assume that k^{th} node to the end node all have the claimed property, that is, the product of weights on all paths from i^{th} node to the end node sum up to one for all $i \geq k$. Let's say that $(k-1)^{st}$ node v has m different outgoing edges, namely $e_1 = (v, v_1), e_2 = (v, v_2), \cdots, e_m = (v, v_m)$. Because the edges always connect to later nodes, all of the nodes $v_1, v_2, \cdots, v_m$ must have the claimed property.

Any path from node v to the end node must first go though one of these m nodes. Let P_i be the set of all paths from the node v to the end node that traverses the edge e_i. The sum of products of weights of the paths in P_i is the same as the sum of products of weights of the paths that connects node v_i to the end node, multiplied by the edge weight of the edge e_i. Because it is assumed that the product of weights on all paths from v_i to the end node sum up to one, the products of weights of the paths in P_i must sum up to the edge weight of e_i.

Because all the paths from node v to the end node must traverse exactly one of the m edges, $e_1, e_2, \cdots, e_m$, the products of the weights on all paths from the node v to the end node is the same as the sum of the edge weights of $e_1, e_2, \cdots, e_m$. Because the outgoing edge weights are normalized in the backward step so that they sum up to one, the sum of edges weights of $e_1, e_2, \cdots, e_m$ must be one. Thus, the products of the weights on all paths from the node v to the end node must sum up to one as well.

By mathematical induction, the products of the weights on all paths from any node in the graph to the end node sum up to one. □

If the target node is the start node, the following simple corollary can be derived from the previous claim.

Corollary 2. *After the backward step is done, the product of weights of any path is the same as the path probability of the path.*

Proof. By Claim 1, the sum of product of weights of all the paths from the start node to the end node must be one. Therefore, by definition of path probability, the path probability is the same as the product of weights of the path. □

Because of Claim 1, we can also prove that after the backward pass, the edge weight on any edge starting from the start node must be the same as its edge probability. A more generalized claim is shown below.

Claim 2. *After backward step, the ratio among edge weights of the edges starting from the same node is the same as the ratio among the edge probabilities.*

Proof. Think of two edges starting from the same node, say $e_1 = (v, v_1)$ and $e_2 = (v, v_2)$. We need to prove that the ratio of edge probability of e_1 vs. e_2 is the same as the ratio of edge weight of e_1 vs. e_2.

By definition, the edge probability of e_1 can be calculated as the sum of path probabilities of the paths that traverse the edge e_1. Because of Corollary 2, the edge probability of e_1 is the same as the sum of product of weights on all the paths that traverse the edge e_1.

Let P_0 be the set of all paths from the start node to v, and P_1 be the set of all paths from v_1 to the end node. Then, the edge probability of e_1 can be written mathematically as the following.

$$\begin{aligned}&\text{Probability of } e_1\\ &= \sum_{p \in P_0, q \in P_1} (\text{Score of } p) \times (\text{Weight of } e_1) \times (\text{Score of } q)\\ &= \sum_{p \in P_0} (\text{Score of } p) \times (\text{Weight of } e_1) \times \sum_{q \in P_1} (\text{Score of } q)\end{aligned}$$

Due to Claim 1, it is known that $\sum_{q \in P_1}(\text{score of } q) = 1$. Therefore, the edge probability of e_1 can be calculated as

$$\text{Probability of } e_1 = (\text{Weight of } e_1) \sum_{p \in P_0} (\text{Score of } p)$$

By the same argument, the edge probability of e_2 can be calculated as

$$\text{Probability of } e_2 = (\text{Weight of } e_2) \sum_{p \in P_0} (\text{Score of } p)$$

Therefore, it can be concluded that

$$\text{Probability of } e_1 : \text{Probability of } e_2 = \text{Weight of } e_1 : \text{Weight of } e_2$$

□

This provides the following corollary that is the first step of the forward step.

Corollary 3. *After the backward step is performed, the edge weights of the edges that start from the start node are the same as their edge probabilities.*

Proof. The last step of the backward step normalizes the outgoing edge weights so that their sum is one. Because the ratio among the edge weight is the same as the ratio among edge probabilities, and the sum of the edge weights and the total of edge probabilities are both one, the edge weights and edge probabilities are the same for the edges outgoing from the start node. □

Using the properties and the claims proved above, we can finally claim that the forward step produces the edge probabilities.

Claim 3. *After each of the forward steps, the edge weights of the edges starting from the target node become the same as the edge probabilities.*

Proof. Proof by mathematical induction. The first step is already proved by Corollary 3. Now, assume that this claim is true for all the nodes that have been processed. This means that the edge weights of all the incoming edges of the current target node v must be the same as their edge probabilities.

Due to Claim 2, it is known that the ratio of the weights on outgoing edges of v must be the same as the ratio of the edge probabilities. Also Property 2 states that the sum of edge probabilities of outgoing edges of v must be the same as the sum of probabilities of incoming edges. Since the probabilities of the incoming edges are known, the outgoing edge probabilities can be calculated based on this. The procedure given in the forward step calculates these values. □

Claim 3 proves the validity of the proposed algorithm. The complexity of this algorithm is $O(n^2)$ where n represents the number of nodes, because each of the two steps consists of n iterations, each of which deals with summing and modifying the weights of at most n edges.

7.3.4 Graph Pruning Procedure

Using the edge probability defined previously, the overall pruning procedure can be performed as the following:

1. **Node Pruning**: Prune out all the structurally meaningless nodes, that is, the nodes that do not have any incoming edges or outgoing edges are deleted from the graph and the edges that are attached to the removed nodes are also deleted. This step is performed before the edge pruning step because Claim 1 assumes that every node must have a path connecting to the end node, and also because deleting meaningless nodes will reduce the computation time for the probability computation time.

2. **Probability Calculation**: Apply the edge probability calculation algorithm to the pruned graph.

3. **Edge Pruning**: Prune edges whose edge probabilities do not pass a certain criterion. The edge pruning is not performed by setting a threshold on the value of the individual edge probability. This is because there may be dozens of edges stemming from the same node, and when that happens, none of the edge probabilities may exceed the given threshold. Instead, the following algorithm is used:

   (a) For each node, find the maximum outgoing edge probability.

   (b) For each edge stemming from the node, if the ratio of its edge probability relative to the maximum edge probability does not exceed a given threshold, delete the edge.

   (c) Do the same for the incoming edges.

   By using this algorithm, the edges with maximum probability will always be retained, however small their absolute values may be.

4. **Repeat**: The procedure should be repeated because the pruning of edges may result in isolated nodes and further changes in the edge probabilities. The

algorithm will be terminated when no more pruning occurs in both node and edge pruning steps. Because the change of probabilities due to the pruning is generally smaller than the threshold value, the pruning procedure is stabilized quickly within three repetitions in most cases.

7.4 Result

7.4.1 Evaluation Criteria

The performance evaluation was carried out based on three aspects: the compactness of the pruned graph, the reliability of reliably detected landmarks, and the ambiguity of ambiguous regions. The compactness represents how small the graph can be reduced without sacrificing the detection rate of the landmarks excessively, the reliability represents how much the landmarks in the reliable regions can be trusted, and the ambiguity represents how many alternatives there are for each region detected as ambiguous.

Compactness

The compactness of the pruned graph can be evaluated by comparing detection rate to the size of the graph, such as numbers of nodes and edges, so that it can be confirmed that the excessive elements have been pruned out without sacrificing the correctly detected landmark candidates.

The detection rate is defined to be the maximum number of correctly detected landmarks among all the possible paths in the graph, divided by the number of true landmarks. For example, assume that the true landmark sequence of the graph in Figure 7-12 is (+b, +g, −g). Although all three landmarks are present in the pruned graph, there are no paths that visit all three nodes. Because there is a path that goes through two of the true landmarks, the detection rate of this graph is determined to be 2/3.

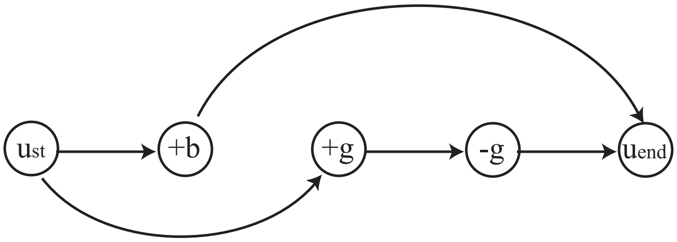


Figure 7-12: Illustration for explaining the definition of detection rate

Reliability

Reliability is estimated using three different measures: proportion of the reliably detected landmarks, and the deletion and insertion rate within reliable regions.

The proportion of reliable detected landmarks is measured by calculating the ratio between the number of reliably detected landmarks compared to the number of total labeled landmarks. This measure is calculated to make sure that some sizable portion of the true landmarks is actually detected as reliable.

Because a reliable region is represented by a linear graph with no side branches, and so has only one likely landmark sequence, the detection rate can be defined simply by the number of correctly determined landmarks out of the total number of true landmarks that are supposed to be within reliable regions. Note that this definition of deletion rate is the same as the sum of deletion and substitution rates defined in Section 6.4.1. The insertion rate in a reliable region also needs to be measured to make sure that the number of missing landmarks is low within a reliable region.

Ambiguity

Ambiguity is estimated by three measures as well: proportion of the ambiguous regions, number of alternatives in each ambiguous region, and the detection rate within ambiguous regions.

The proportion of the ambiguous regions is calculated as the complement value of the proportion of reliable landmarks. The number of alternatives is evaluated by counting the number of possible paths in an ambiguous region. Because all ambiguous regions should have at least two alternative choices of possible landmark sequences, this value should always be at least two. Detection rate within an ambiguous region is another factor for measuring ambiguity, because even if the number of alternatives is small, it would not be useful if none of the alternative sequences represent the true landmark sequence.

7.4.2 Evaluation

For performance evaluation of the pruning algorithm, a hundred utterances were randomly selected from the dialect region 1 of the TIMIT database, and the evaluation was carried out on four different edge-pruning thresholds: 0.2, 0.1, 0.05 and 0.01.

Compactness

Table 7.1 lists the size of the pruned graph and the error rate for different thresholds. When threshold 0.2 is used—i.e., when no elements with probability more than 2% are removed—the result is very similar to that of the Viterbi search result. On the other hand, when a small threshold 0.01 is used, the number of nodes is not reduced as much as that from the 0.2 threshold, but half of the nodes in the unpruned graph are removed, and the number of edges is reduced to one eighth of the unpruned graph. Also note that overall detection rate is not affected appreciably even if the size of the graph is reduced significantly, even with a small threshold.

Threshold	Viterbi	0.2	0.1	0.05	0.01	Unpruned
Nodes	25.2	26.0	29.8	34.8	43.8	87.2
Edges	24.2	27.6	36.5	53.5	86.2	640.9
Detection	76.8%	77.0%	83.3%	88.5%	93.0%	95.2%

Table 7.1: Size of graph and detection rate for different thresholds

Reliability

The deletion and insertion rate within the reliable region is shown in Table 7.2. The proportion represents the proportion of the reliable region, that is, the ratio between the number of reliably detected true landmarks and the total number of true landmarks.

Note that the detection and insertion rate within the reliable region is much smaller than that of the Viterbi search result from Chapter 6 even when a large threshold is used. Considering that the compactness measures of the graph pruned with a threshold of 0.2 and the Viterbi search result are almost similar, the large difference in the error rates is unexpected. This means that most of the errors in the Viterbi search results were detected as ambiguous when the graph pruning method was applied.

When the threshold of 0.01 is applied—that is, when no edges with more than 1% probability are pruned out—the proportion of the reliable region is still as high as 40%. In addition, the deletion and insertion error rate within the reliable regions is as small as 5%, which is almost as small as the deletion rate within the input candidates (see Section 5.5.1 for the detection rate for the individual landmark candidate detection process).

Threshold	Viterbi	0.2	0.1	0.05	0.01
Deletion	23.2%	12.8%	11.2%	7.7%	5.6%
Insertion	14.7%	8.1%	7.3%	5.6%	4.2%
Proportion	—	70.0%	63.9%	55.5%	40.9%

Table 7.2: Deletion and insertion rates within reliable regions for different thresholds

Ambiguity

The number of alternatives and detection rate within ambiguous regions are given in Table 7.3. The proportion represents the proportion of the ambiguous region, which is the complement of the proportion of the reliable region. The ambiguous region of an unpruned graph is not 100% because there were landmarks that were detected with absolute confidence and were judged to be as reliable.

Note that the detection rate within ambiguous regions increases rapidly when the threshold is lowered, and the detection rate within ambiguous regions becomes almost comparable to that of the unpruned graph when the threshold of 0.01 is used. On the other hand, the number of alternatives increases significantly if this lower threshold is used. There are mostly two alternative choices when the threshold 0.2 is used, but when the threshold 0.01 is used, the number of alternatives increases up to 13.6, although this value is still negligibly small compared to that of the unpruned graph. Assuming that each additional measurement of acoustic cue can reduce the number of alternative choices by half, we can decide on the correct landmark sequence within an ambiguous region by measuring no more than 3–4 additional cues on average.

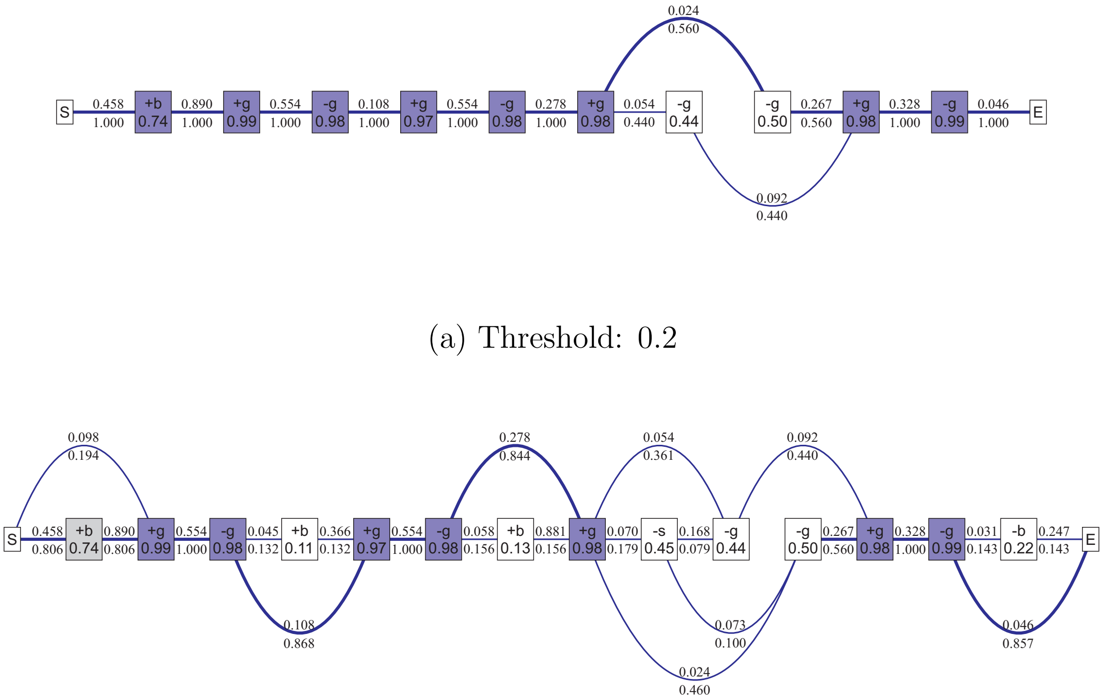
Threshold	0.2	0.1	0.05	0.01	Unpruned
Alternatives	2.2	2.9	5.0	13.6	> 1000
Detection	40.3%	70.1%	82.8%	92.0%	95.2%
Proportion	30.0%	36.1%	44.5%	59.1%	96.2%

Table 7.3: Detection rate and number of possible alternatives within ambiguous regions

7.4.3 Discussion

From the evaluation, it can be concluded that graph pruning based on edge probability can be used as an effective method to differentiate the reliable regions from ambiguous regions. The evaluation shows that substantial portions of the erroneous contexts are captured in the ambiguous regions even when a high threshold is applied.

For example, Figure 7-13 illustrates two pruned graphs generated from the same utterance by applying the graph pruning method with different pruning thresholds—0.2 and 0.05. The numbers under the landmark types represent individual probabilities of the landmark candidate, the numbers written above the edges are edge weights, and the numbers written below the edges are edge probabilities. The dark



(a) Threshold: 0.2

(b) Threshold: 0.05

Figure 7-13: Comparison of pruning examples with different thresholds. An utterance of the sentence "This was easy for us" was used for this example. The numbers under landmark types are individual probabilities, the numbers written above the edges are edge weights, and the numbers below the edges are edge probabilities. The dark shaded nodes represent the landmarks that are reliable, and lightly shaded nodes are the landmarks that are in ambiguous region, but still have a high probability of being true landmarks.

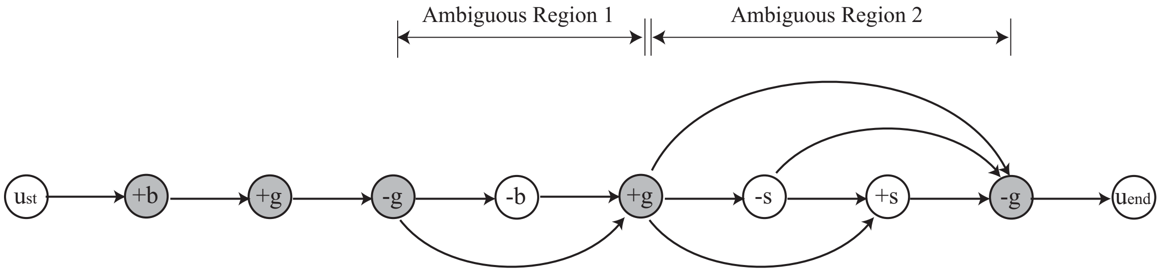


Figure 7-14: An example of a pruned graph. The shaded nodes represent the landmarks that are reliable.

shaded nodes represent the landmarks that are reliable, and lightly shaded nodes are the landmarks that are in ambiguous region, but still have high probability of being true landmarks. The edges are drawn with think lines when the edge probabilities are larger than 0.5.

The graph pruned with the threshold of 0.2 detects most of the landmarks as reliable, and only one region is detected as ambiguous. This ambiguous region corresponds to an erroneous case in which a +g landmark is missing due to a heavily voiced fricative /z/. When the graph is pruned with a lower threshold, the resulting graph may have more ambiguous regions, but the landmarks that could have been missed in the high threshold case can be detected. The example in Figure 7-13 shows that the offset of the last fricative in the word 'us', which should be detected as a −b landmark, is not detected in Figure (a) due to its low individual probability, but is detected in (b), although as one of the possibilities in an ambiguous region.

Figure 7-14 shows a graph of an actual utterance of the sentence "did you eat yet," pruned with threshold of 0.05. The weights and probabilities are not indicated for visual clarity. The graph has two ambiguous regions and five reliably detected landmarks. Note that an ambiguous region always has to be an interval between two (reliable) landmarks, while reliable regions can be sometimes represented by a single landmark.

As for reliable regions, the landmark sequence is determined with more confidence, and it can be assumed that the landmarks have been clearly produced by the speaker.

Therefore, the distinctive features within the region can be estimated with more confidence.

It has been also observed that the reliable regions may correspond to word boundaries and stressed syllables. Figure 7-15 shows a pruning example of a relatively long utterance. This is an utterance of the sentence “This, he added, brought about petty jealousies and petty personal grievances,” and the threshold of 0.05 was used for the pruning algorithm.

The regions that are judged as reliable correspond to the content words ‘brought about’, ‘jealousies’ and ‘personal grievances.’ Especially for the last two words, the underlined syllables of ‘personal’ and ‘grievances’ were detected as reliable, supporting the assumption that the landmarks are more reliably detected within the syllables with lexical stresses.

On the other hand, for ambiguous regions, certain decisions about the phonetic context can be made based on the possible alternative landmark sequences, and additional cues can be subsequently measured within the region so that the correct alternative can be chosen with more confidence. The first ambiguous region in Figure 7-14 has two alternatives, (−g, −b, +g) or (−g, +g). From these alternatives, it can be determined without additional measurement that the ambiguous region has a [−sonorant] feature, and there may or may not be a turbulence noise, which can be determined with more confidence by measuring additional cues for noise detection. The second ambiguous region has four alternatives depending on the existence of +s and −s landmarks. This means that the region is [+sonorant] no matter which alternatives may be true, and the actual landmark sequence can be confirmed by measuring additional cues for vowels and nasals.

The choice between Viterbi search versus reliable-ambiguous region notation is a trade-off between simplicity and correctness of the result, and it may be determined according to the characteristics of the application. When Viterbi search is applied, there is only one possibility, and it would become simpler to perform further processing on the result. However, there may be some undetermined landmarks and these errors may lead to problems that require additional attention. On the other

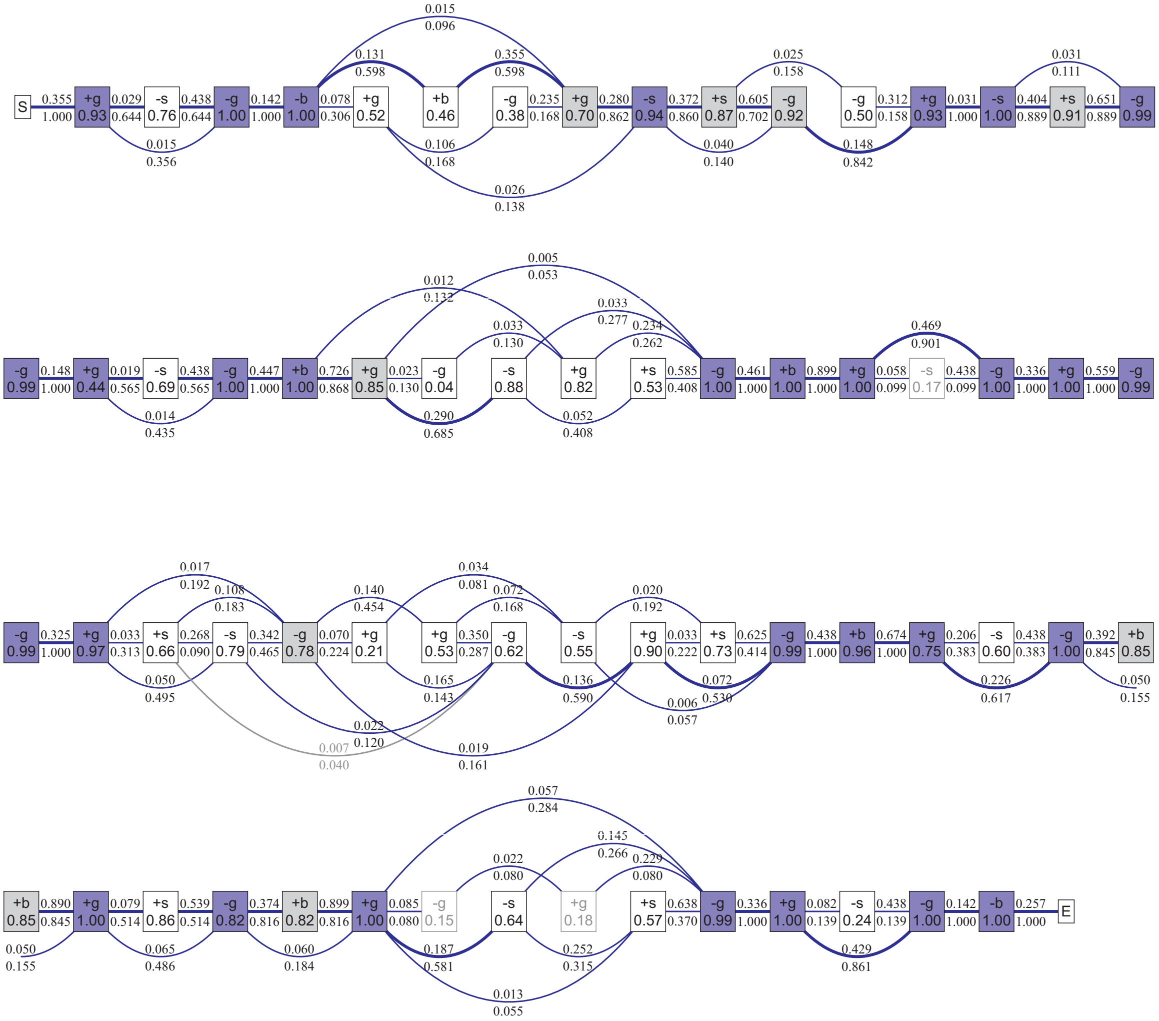


Figure 7-15: An example of a pruned graph of a relatively long utterance, "This, he added, brought about petty jealousies and petty personal grievances."

hand, when the reliability-ambiguity representation is used, the correctness of the detection within reliable regions is generally higher than the Viterbi search result. However, further processing may need to be applied to each of the reliable regions and ambiguous regions to make use of the additional information.

Chapter 8

Conclusion

8.1 Summary

8.1.1 Objective

In this thesis, a probabilistic model for finding consonant landmarks has been developed. The consonant landmark detector locates the time-points of abrupt acoustic changes, corresponding to closures and releases of consonants. Not only does this algorithm pinpoint the location of abruptness, but it also classifies the detected landmarks according to their characteristics—onsets and offsets of spontaneous glottal vibrations, existence of burst noise, and closures and releases of sonorant consonants. Therefore, this landmark detector can serve as a first step of a knowledge-based automatic speech recognition system, by finding the locations where phonetic information is highly concentrated, and also by determining the type of information that can be found in the vicinity of each detected landmark or between each group of adjacent landmarks.

8.1.2 Method

The landmark detection algorithm is developed in three stages. The first stage provides a probabilistic algorithm that detects the consonant landmarks individually, that is, without considering the relationship among different types of landmarks. The second stage makes use of the strict dependency that we observe among true landmark sequences to determine the most likely sequence among the detected landmarks. The last stage distinguishes the reliable regions, where the detected landmarks can be highly trusted, from the ambiguous regions, where more information needs to be extracted for more reliable landmark sequence detection.

Individual Landmark Detection

The individual landmark candidate detection algorithm developed in the first stage is largely a direct conversion of Liu's consonant landmark detector [35] into a probabilistic system. The new design has the following three advantages:

First, instead of implementing the landmark detector and classification module as a series of decision processes, each of which represents a piece of speech knowledge, the overall system has been redesigned to have a separate knowledge-base and a computation core system. The separation of the knowledge-base makes it possible to update the speech knowledge without requiring redesign of the whole system.

Secondly, the new algorithm allows more candidates to be detected, by increasing the sensitivity to acoustic changes. The resulting reduction of landmark deletions is considered more important than the reduction of insertions, because once a landmark sequence is determined, a deleted landmark can be retrieved only with higher level knowledge such as phonotactic constraints and lexical access, while an insertion can be discarded by measuring low-level cues alone. This change in sensitivity reduced the landmark deletion rate significantly, from 18% to less than 5%.

Finally, the algorithm was developed as a probabilistic system by assigning a probability of each landmark candidate being a true landmark. The probability measure can be used in later processing steps to sort out false alarms that have arisen due to increased sensitivity of the landmark candidate detection algorithm. The result shows that not only can 90% of the falsely detected landmark candidates be filtered out based solely on the probability measure, but also 80% of the correctly detected landmarks have probability higher than 0.9 and 85% of the false alarms have a probability less than 0.2, implying that most of the landmarks can be detected with considerable confidence.

Landmark Sequence Determination

The landmark sequence determination process developed in the second stage is meaningful not only because it determines a single landmark sequence that can be directly input to the next step of automatic speech recognition, such as distinctive feature detection or lexical access based on manner features, but because this process models the relationship among different landmark types with a simple bigram representation.

The bigram model that represents the constraints on possible landmark pairs has significance in two ways. One is that among 36 theoretically possible pairs of landmark types, only 16 pairs are articulatorily feasible. The other 20 landmark sequences are not just probabilistically unlikely: it is physiologically impossible to produce them. Therefore, the bigram representation of landmark sequence constraints can be effectively used to filter out unlikely sequences of landmarks.

Another advantage is that each of the sixteen landmark pairs describes some of the acoustic properties of the signal between the two landmarks. Because the acoustic characteristics mostly correspond to the articulator-free features of the segment, when a landmark sequence that follows the bigram constraints is observed, the articulator-free features of the signal can be directly derived from the landmark sequence.

The landmark sequence that is determined based on both individual probability and bigram transition probability shows about 12% deletion and substitution rates and a 15% insertion rate, which are generally less than those of Liu's algorithm. Most of the detected landmarks that do not correspond to the labeled landmarks occur in similar contexts, implying that these errors may provide additional information about the speech signal. Some of the common contexts include flaps, syllabic nasals and /l/'s, glides, and irregular pitch periods.

Representation of Reliable and Ambiguous Regions

The last stage of the algorithm distinguishes the regions where the landmark sequences can be determined reliably from the regions where more than one possible sequences can be hypothesized. It is expected that where one speaks more carefully, e.g., near word boundaries or lexical stresses, cues to the distinctive features will be produced more clearly and the landmarks will be detected more reliably, and so the reliably detected landmarks can be given more focus in later stages of speech recognition. On the other hand, the ambiguous regions can be given multiple possibilities of landmark sequences, and evidence for additional cues can be sought based on the possible choices, so that the landmark sequence can be determined with more confidence.

This stage uses the same measures that were used in the landmark sequence determination process—individual probabilities and bigram transition probabilities. However, instead of finding a single most likely landmark sequence using a Viterbi search algorithm, this stage uses a graph pruning method based on a specially defined edge probability measure.

Even when the pruning method is applied with a low threshold of 0.01, the size of the graph can be reduced considerably—the number of nodes is halved, and the number of edges can be reduced to one eighth of the unpruned graph on average—and still retain a detection rate as high as 93%, which is almost comparable to the theoretical maximum of 95%. The graph pruning method determines about 40% of the landmarks are reliable, and the deletion rate among the reliable landmarks is as low as 5.6%, which is significantly reduced from the 23% deletion rate of the Viterbi search algorithm. The ambiguous regions are also reduced to a manageable size, so that it is possible to find a true landmark sequence by measuring no more than 3–4 additional cues per ambiguous region.

8.2 Applications

8.2.1 Distinctive Feature Estimation

Each landmark type corresponds to a major movement of the speech organs. Therefore, when the landmark types are estimated, some of the articulator-free features can be determined directly from the landmark types. For example, a +s landmark specifies a change from a [+consonant, +sonorant, −continuant] sound to a [−consonant] sound, as in the transition between two sounds in the word ‘no’.

The bigram transition constraints applied in the landmark sequence determination and reliability representation processes help in the estimation of articulator-free features in two ways. First, because the bigram restriction asserts that the resulting set of landmarks should follow the strict rules of landmark sequencing, it reduces the possibility of contradiction between the sets of distinctive features estimated from adjacent landmarks. For example, when the bigram restriction is not applied, a sequence of (+b, −s) landmarks can be detected as a result. Because a +b landmark implies that the region following the landmark must be [+consonant, −sonorant] and a −s landmark implies that the region preceding the landmark must be [−consonant, +sonorant], the articulator-free feature between the landmark pair cannot be determined. Constraints on successive landmark pairs reduce such contradictions.

Another advantage is that, as was discussed in Section 2.2, landmark pairs are more effective than single landmarks in determining some of the articulator-free features. For example, a −g landmark alone cannot correctly determine the features of the following region—it could be either a [+consonant, −sonorant] region, or a silence—and a −b landmark cannot determine the continuant feature of the preceding region. However, when the landmark sequence (−g, −b) is detected, it can be derived that the region between the landmark pair has the features [+consonant, −sonorant, +continuant].

The bigram constraints help with the estimation of articulator-bound features as well, not only because the determination of articulator-free features restricts the number of articulator-bound features that need to be determined as stated in Section 2.2, but also because a landmark pair provides a region where the cues for a certain distinctive feature can be found throughout. For example, when a sequence of (+g, +s) landmarks is detected, defining the features of a sonorant consonant, the nasality feature of the sonorant consonant can be detected by verifying the existence of a nasal murmur in the region between the landmarks, and also by examining the nasalization of the adjacent vowel in the region between the +s landmark and the following landmark.

8.2.2 Lexical Access

A study by Shipman and Zue [46] showed that even an incomplete representation of segments can help to reduce the number of possible word candidates in an isolated word recognition system. They report that when only the CV pattern of a word is known, the average number of possible word candidates reduces to 25 words for a 20,000 word vocabulary, and in the worst case, the maximum size is reduced to 1,500 words, which is only 7.5% of the entire lexicon. On the other hand, when more detailed characteristics of the segments are known, that is, when each segment is specified to be one of the six broad classes—vowels, stops, nasals, strong fricatives, weak fricatives, and glides and semivowels—the average number of possible word candidates reduces to about 2.5 words and in the worst case, the candidate set size is reduced to 200 words.

A similar experiment can be performed on the sequence of landmarks, since the landmark types can specify some of the articulator-free features of the segments. The result was expected to be better than that of CV pattern but not as good as that of six-way classification, because the landmarks alone cannot decide the continuant and strident features clearly, and consonant landmarks cannot distinguish a series of vowels and glides from a single vowel.

The result is shown in Table 8.1. The average is calculated by averaging the number of word candidates across all the possible landmark patterns, and the weighted average is calculated using the frequency data in the Brown Corpus. The maximum is calculated by finding a landmark sequence that has the largest number of possible word candidates. The average number of possible word candidates was reduced to 5.6 words with the worst-case result equal to 722 words, which is about 3.6% of the total number of candidates. As was expected, these results fall between those for CV sequences and for a full representation of broad manner classes.

Huttenlocher and Zue [24] extended the same experiment to the case where rea-

	CV Patterns	Landmarks	Six-way Classes
Average	25	5.6	2.5
Weighted Average	600	195	40
Maximum	1500	722	200

Table 8.1: Comparison of the numbers of possible word candidates when incomplete information about the word is known. A 20,000 word vocabulary is used for the comparison, and weighted average is calculated according to the frequency in the Brown Corpus.

sonable confusion errors in the broad classes are allowed, and found out that the additional confusions did not change the number of word candidates. In addition, they performed a further experiment that examined the effect of unspecifying some parts of words completely. The result showed that when the unstressed syllables were ignored, the expected size of the set of word candidates grew from 25 to 40, but when stressed syllables were ignored, the size increased significantly to 2,000.

This experiment supports the validity of our proposed reliability-ambiguity representation, because the reliable regions are expected to correspond to lexical stresses and word boundaries. Therefore, specifying the landmarks in the reliable regions and allowing confusion within the ambiguous regions would still reduce the size of the word candidates considerably.

8.3 Future Work

8.3.1 Improving the Landmark Detector

Temporal Cues

One of the most common errors in the landmark sequence determination process was high insertion rate of b-landmarks near g-landmarks. Of all the b-landmark insertions, almost 80% occurred within 10ms from g-landmarks. This problem arises because g-landmarks not only affect the low frequency region but also create a discontinuity in the high frequency region, which is used for b-landmark detection.

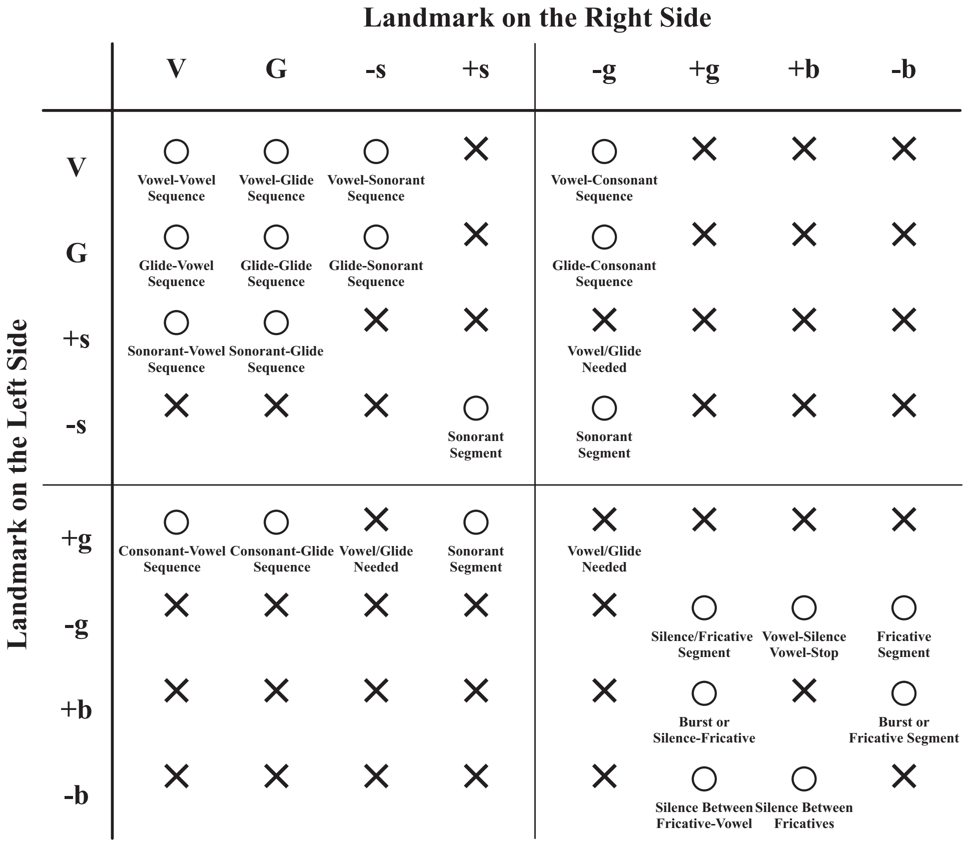
Such problems can be reduced by using the distance between landmarks in determining the landmark sequence. For example, by simply not allowing a sequence of b- and g-landmark within a 7ms interval, the insertion rate of b-landmark was reduced by half without affecting the detection rate significantly. For a more general application of temporal cues, the distribution of the distances between each possible landmark pair can be estimated and applied in the landmark sequence determination.

Three Classes of Landmarks

This thesis only dealt with three types of consonant landmarks. Because knowledge of the types of consonant landmarks makes it possible to hypothesize possible vowel and glide positions, the consonant landmarks alone can provide a large amount of information about the speech signal.

However, as was discussed in Section 2.1, the vowel and glide information can be detected reliably based on the acoustic cues extracted from a longer time period than consonant information. Therefore, if vowel and glide landmarks are incorporated together with consonant landmarks, the resulting landmark sequence will become more reliable than the result from consonant landmarks alone.

The possible transitions between all three classes of landmarks can be represented by a bigram transition model as shown in Table 8.2. Adding vowel and glide landmarks changes the allowed transition between consonant landmarks as well; for example, (+g, −g) landmark pairs are no longer allowed because there must be a vowel between these two landmarks. Such finer constraints would increase the accuracy of landmark sequence determination, especially for the instances where acoustic discontinuities are less reliable. For example, in the consonant-only case, a glide was sometimes detected with s-landmarks, but when the bigram transition including a glide landmark is used, such insertions will be deleted because neither (−s, G) landmark transitions nor (G, +s) transitions are possible. Similarly, a schwa vowel tends to have less abruptness at the segment boundary, but it will be detected with more accuracy when a vowel landmark is located in the middle of the schwa.



Landmark on the Right Side

Landmark on the Left Side

	V	G	-s	+s	-g	+g	+b	-b
V	○ Vowel-Vowel Sequence	○ Vowel-Glide Sequence	○ Vowel-Sonorant Sequence	×	○ Vowel-Consonant Sequence	×	×	×
G	○ Glide-Vowel Sequence	○ Glide-Glide Sequence	○ Glide-Sonorant Sequence	×	○ Glide-Consonant Sequence	×	×	×
+s	○ Sonorant-Vowel Sequence	○ Sonorant-Glide Sequence	×	×	× Vowel/Glide Needed	×	×	×
-s	×	×	×	○ Sonorant Segment	○ Sonorant Segment	×	×	×
+g	○ Consonant-Vowel Sequence	○ Consonant-Glide Sequence	× Vowel/Glide Needed	○ Sonorant Segment	× Vowel/Glide Needed	×	×	×
-g	×	×	×	×	×	○ Silence/Fricative Segment	○ Vowel-Silence Vowel-Stop	○ Fricative Segment
+b	×	×	×	×	×	○ Burst or Silence-Fricative	×	○ Burst or Fricative Segment
-b	×	×	×	×	×	○ Silence Between Fricative-Vowel	○ Silence Between Fricatives	×

Table 8.2: Bigram restriction of all three classes of landmarks. The capital letter V represents a vowel landmark and G represents a glide landmark. Other symbols with + and − signs stand for different types of consonant landmarks.

Processing in Ambiguous Regions

The reliability-ambiguity representation discussed in Chapter 7 can distinguish sections of the speech signal where the landmarks are produced ambiguously from sections where the landmark distinctions are clear. It has been suggested that for each ambiguous region, the list of hypothesized landmark sequences be used to extract the features that are common to all alternative choices and the features that vary depending on the alternatives, so that the common features can be confirmed and the non-common features can be evaluated by measuring additional cues. Such a process still needs to be developed.

In addition, as discussed in Section 6.6, most of the errors occur within a limited set of phonetic contexts. Therefore, it can be assumed that if the error contexts are analyzed comprehensively, by making a list of all possible phonetic contexts that trigger ambiguous landmark realization and the frequency of these contexts, this information will be useful in determining the correct landmark sequence for the ambiguous regions.

8.3.2 Other Applications

Suprasegmental Features

It has been observed that the reliably detected regions described in Chapter 7 generally correspond to lexical stresses or word boundaries. It has been generally accepted that stressed syllables have robust acoustic cues for phonetic features [11]. Gow *et al.* [19] also suggest that in continuous speech, word onsets show more robust acoustic realization of phonetic features and are less variable in terms of phonological assimilation than other parts of words. Therefore, it would be worthwhile to investigate the relationship between reliable detection of landmarks and suprasegmental features such as word onsets and lexical or prosodic stress.

The experiment by Huttenlocher and Zue [24] shows that knowledge of lexical stress and of the broad phonemic classification of the stressed syllables makes it possible to reduce the number of possible word candidates significantly. In addition, the phonetic information in word onsets plays an important role in many lexical access models [18, 39]. Thus, if it is verified that landmark reliability corresponds to perceptual islands of reliability, the landmarks within the reliable regions can be used as providing valuable information for lexical access.

Language Independence of Landmarks

It is observed in Section 6.4.4 that landmark detection is likely to be independent of dialect of American English and also of gender, although a more extensive analysis is needed to confirm the claim. It has been assumed that this property is due to the fact that the consonant landmarks do not depend on specific phonemes but depend only on the overall configuration of the vocal tract shape. If that assumption is true, it can be hypothesized that the landmarks should be independent from the characteristics of different languages as well.

Although the individual realization of landmarks may be independent from language, the transition constraints between successive pairs of landmarks are thought to be different across languages. For example, Korean speakers do not have consonant clusters and do not release syllable final obstruent consonants. Therefore, the sequences of landmark pairs (−b, +b) and (+b, −b) are less likely to be realized in Korean. Similarly, the Japanese language does not have syllable-final consonants, except for the /ŋ/ sound, which is often pronounced as a syllable instead of as a coda.

Therefore, by investigating the language-independent and language-dependent characteristics of landmarks, the landmark detection algorithm can be easily adapted to speech analysis and recognition systems for different languages, and the language-dependent characteristics may be used as one of the cues for language identification as well.

Landmarks and Speaking Styles

Similar experiments can be carried out on other characteristics of speech. For example, Boyce *et al.* [6] estimated consonant landmarks from clearly pronounced speech and conversational speech and showed that these two speaking styles can be distinguished based on the different distribution of certain landmark clusters. She also observed that the number of landmarks present in clear speech is more than that in conversational speech, due to overlapping of articulatory gestures in the latter style.

It would be also worthwhile to examine the reliability-ambiguity representations in different speaking styles. It is expected that the proportion of reliable regions in conversational speech may be less than that of clear speech, but that the landmarks near the stressed syllables or content words will still be reliably detected.

Although the characteristics of landmark sequences may vary with different speaking styles, it is likely that the acoustic characteristics of individual landmark types may not depend on speaking styles, because both speaking styles make the same or at least similar articulatory movements. However, individual landmark characteristics may be different in some atypical speech, such as the utterances of speakers with speech disorders or of young children still learning to talk. The differences may be investigated further to acquire deeper understanding of the acoustic characteristics of atypical speech, and to improve the recognition of landmarks as well.

Appendix A

Feature Bundle Representation of English Sounds

	p	t	k	b	d	g	f	v	θ	ð	s	z	ʃ	ʒ	ʧ	ʤ	m	n	ŋ
Consonant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Vowel																			
Glide																			
Sonorant	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+
Continuant	−	−	−	−	−	−	+	+	+	+	+	+	+	+	∓	∓	−	−	−
Strident							−	−	−	−	+	+	+	+	+	+			
Nasal																	+	+	+
Lateral																			
Rhotic																			
Reduced																			
Spread Glottis																			
High																			
Low																			
Back																			
Round																			
Adv. Tongue Root																			
Lips	+			+			+	+									+		
Blade		+			+				+	+	+	+	+	+	+	+		+	
Body			+			+													+
Anterior		+			+				+	+	+	+	−	−	−	−		+	
Stiff Vocal Folds	+	+	+	−	−	−	+	−	+	−	+	−	+	−	+	−			

Table A.1: Feature bundle representations for consonants in English.

	i	ɪ	e	ɛ	æ	ɑ	ɔ	o	ʌ	u	ʊ	ɚ	ə	w	j	h	r	l
Consonant																		+
Vowel	+	+	+	+	+	+	+	+	+	+	+	+	+					
Glide														+	+	+	+	
Sonorant																		+
Continuant																		−
Strident																		
Nasal																		−
Lateral																		+
Rhotic												+					+	
Reduced													+					
Spread Glottis																+		
High	+	+	−	−	−	−	−	−	−	+	+	−	−	+	+		−	−
Low	−	−	−	−	+	+	+	−	−	−	−	−	−	−	−		−	−
Back	−	−	−	−	−	+	+	+	+	+	+	+		+	−		+	+
Round						−	+	+	−	+	+			+				
Adv. Tongue Root	+	−	+	−				+	−	+	−	−	−	+	+			
Lips														+				
Blade															+		+	+
Body																		
Anterior															−		−	+
Stiff Vocal Folds																		

Table A.2: Feature bundle representations for vowels, glides and liquids in English.

Appendix B

Review of Liu's Landmark Detection Algorithm

Liu [35] has developed an algorithm that detects the acoustic landmarks by utilizing linguistic knowledge. Since the first part of the automatic landmark detection process described in this thesis—especially the general pre-processing algorithm in Section 5.1— is mostly derived from Liu's landmark detector, the algorithm is briefly reviewed here.

Liu's approach is based on a deterministic algorithm that uses a series of decision processes, each of which represents a piece of acoustic knowledge about speech signal. The detection algorithm for each type of landmarks is summarized below.

B.1 Detection of g-Landmarks

The g-landmarks are detected based on the amplitude of the energy in 0–400Hz frequency bands. The time-points that show at least a 9dB change of band-energy in 50ms period, as well as a 6dB change in 26ms period are selected to be possible locations of landmarks. The term *peaks* is used to represent the selected time-points. A +peak is the time-point where the energy increases, and −peaks is where the energy decreases.

After the peaks are selected, the sequence of peaks goes through a series of decision processes that use the following acoustic knowledge about the speech signal:

- **Pairing of g-landmarks**: An utterance always starts with a +peak, and end with a −peak. There must be an alternation of signs from one peak to the next.
- **Minimum vowel requirement**: There must be a vowel between a +g landmark and the following −g landmark. Acoustically, this means that the energy between the (+g, −g) landmark pair should be no less than a certain amount lower than the highest energy in the utterance. This amount is taken to be 20dB.
- **Duration**: The distance between a +g landmark and the next −g landmark should not exceed a certain duration—this duration threshold is decided to be 250ms. This is not an obligatory condition because a long series of vowels and nasals—such as in the phrase ‘in an animal’—can result in a longer interval between +g and −g landmarks.

A simplified flowchart of the decision process that incorporates the speech knowledge listed above is shown in Figure B-1. The flowchart consists of three sections: the initial section that starts with ‘start’ node, and the other two sections starting with nodes A and B, respectively. The initial section of the flowchart, which starts with the start node, locates the first +g landmark. After the initial part is over, the process alternates between section A and section B, determining the sequence of g-landmarks with alternating signs located at proper distances apart. After the end of signal is reached, the process terminates after final processing in section B, which filters out the g-landmark pairs that does not pass the minimum vowel requirement.

B.2 Detection of s-Landmarks

An s-landmark can exist only between a +g landmark and the following −g landmark. If peaks can be found in all of the four different frequency bands—0.8–1.5kHz, 1.2–2.0kHz, 2.0–3.5kHz and 3.5–5.0kHz —within a certain time-period, this time-point

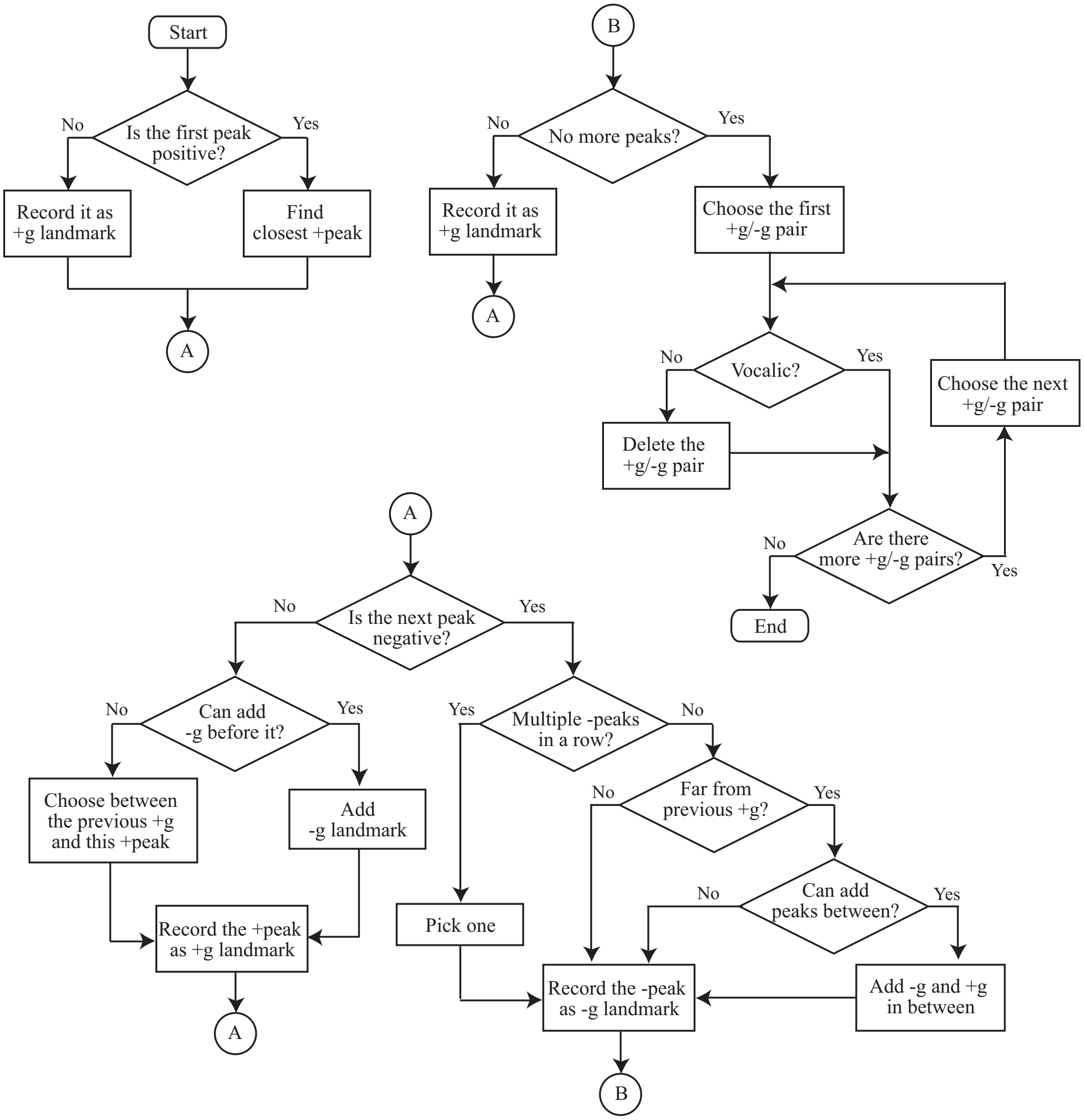


Figure B-1: A simplified flowchart that decides the g-landmark sequence that satisfies a set of restrictions.

is determined to be a *pivot*. If all four peaks are present but they are not close in time, the set of peaks is represented by the term *tagged-pivot*. The tagged-pivot is separately classified because the spectral change for a sonorant consonant is not as abrupt as that of an obstruent consonant, and different frequency band may be affected at slightly different time-points.

After the pivots are decided, the following set of knowledge-based constrants is applied to confirm the existence of landmarks.

- **Steady state**: The shape of the oral tract is stable during a sonorant consonant, and so the spectrum of the 0–600Hz frequency region stays almost the same throughout the consonant. This condition may not hold for an intervocalic position where both the closure and the release of the consonant happen sequentially.

- **Abruptness**: A sonorant consonant introduces a zero near the second formant frequency, and this causes the 1.3–8kHz frequency band energy to go through an abrupt change at the landmark position.

The thresholds for these criteria depend on the position of the sonorant consonant—intervocalic, syllable-onset, syllable-offset. For example, a syllable-offset sonorant consonant tends to be less steady than that of syllable-onset consonant and so a less strict threshold is used in processing the syllable-offset pivots.

Unlike g-landmarks, the s-landmarks do not always need to be paired with other landmarks, because the sonorant landmarks are not always abrupt. Therefore if at least one of the closure and release is abrupt enough (i.e., detected as an untagged pivot), then it is selected as a landmark.

Simplified flowcharts for the decision processes are shown in Figure B-2. The first flowchart shows the decision process of the sonorant consonant position. This process classifies each pivot position into four categories: onset, offset, intervoc and unknown. In the following processing, which is described in the second flowchart, different thresholds for abruptness and steady-state criteria are used for different categories. The second flowchart shows how the different criteria interact to decide the s-landmarks. For example, to be determined as an s-landmark, the pivot should pass abruptness and steady-state criteria and at least one of the closure and release should be untagged. However, it is also allowed that neither closure and release need to be untagged at the end of an utterance, and if the pivot is at intervocalic position, the steady-state criterion does not have to be met.

B.3 Detection of b-Landmarks

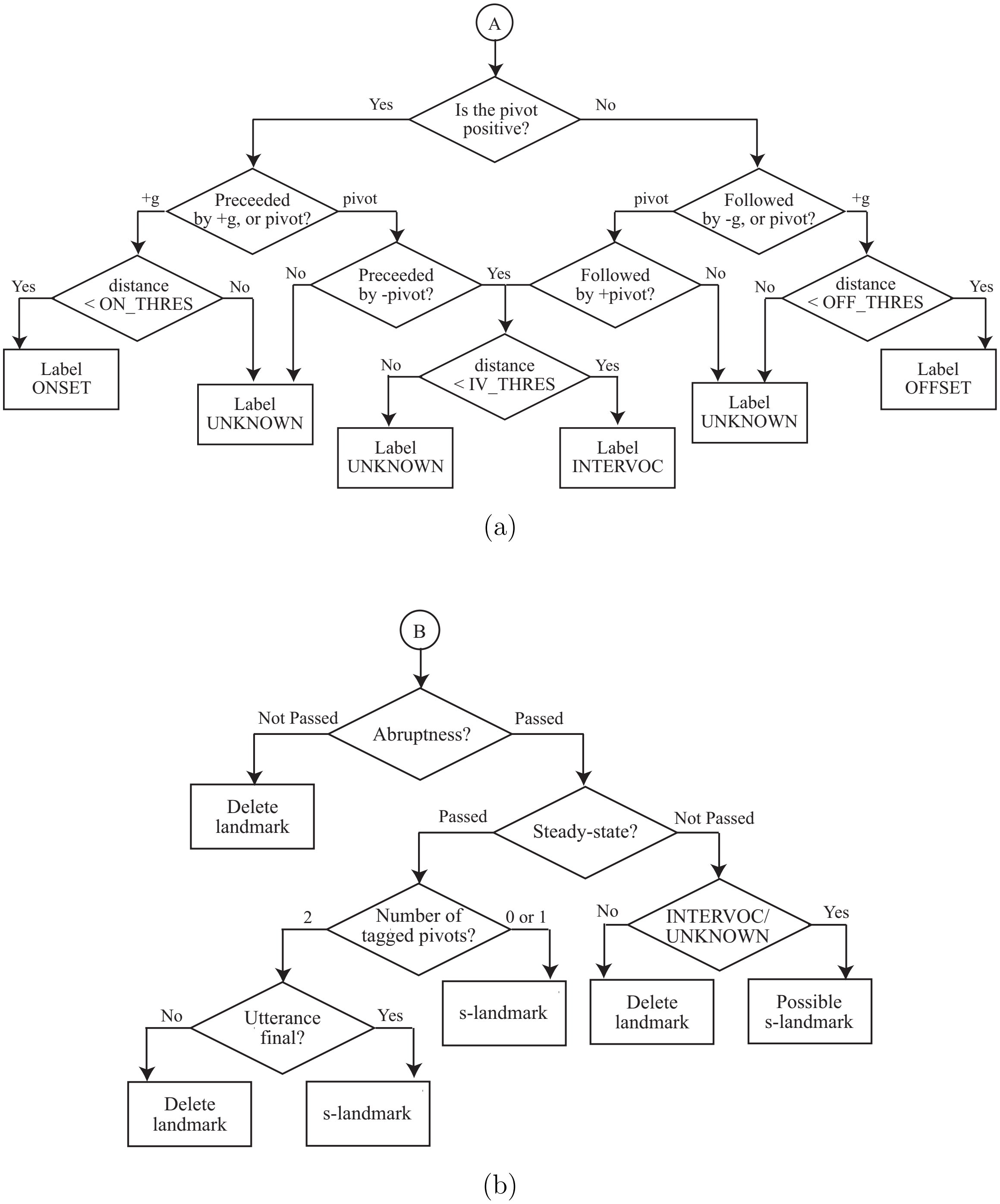
Contrary to s-landmarks, the b-landmarks can be found between a −g landmark and the succeeding +g landmark. The pivots are decided by the same criterion — peaks can be found in all four frequency bands within a certain time period.

The criteria for determining a b-landmark are the following:

- **Silence**: A +b landmark represents a stop burst, and therefore it should be preceded by a complete closure in the mouth for the intraoral pressure to be built up. Therefore, a +b landmark should be preceded by a silent region which is acoustically represented as low-amplitude signal across all of the frequency range. Similarly, a −b landmark should be followed by a silence, but to allow the existence of voice-bar, only the 1.2-8kHz frequency region is measured for this purpose. When a stop consonant is preceded by a vowel, such as in the word ‘finger’, the closure may not be complete due to the opening in the nasal tract.

- **Duration**: The silent region should be persistent over an interval of at least 5ms, and the burst onset duration—the distance between the end of the silence to the burst landmark—should be shorter than 30ms.

Figure B-3 gives a simple flowchart of the b-landmark decision process. This diagram represents the two criteria with possible exceptions for utterance-initial consonant, utterance-final consonants and stop consonants followed by nasals.



(a)

(b)

Figure B-2: Simplified flowcharts that describe the decision process of the s-landmark sequence.

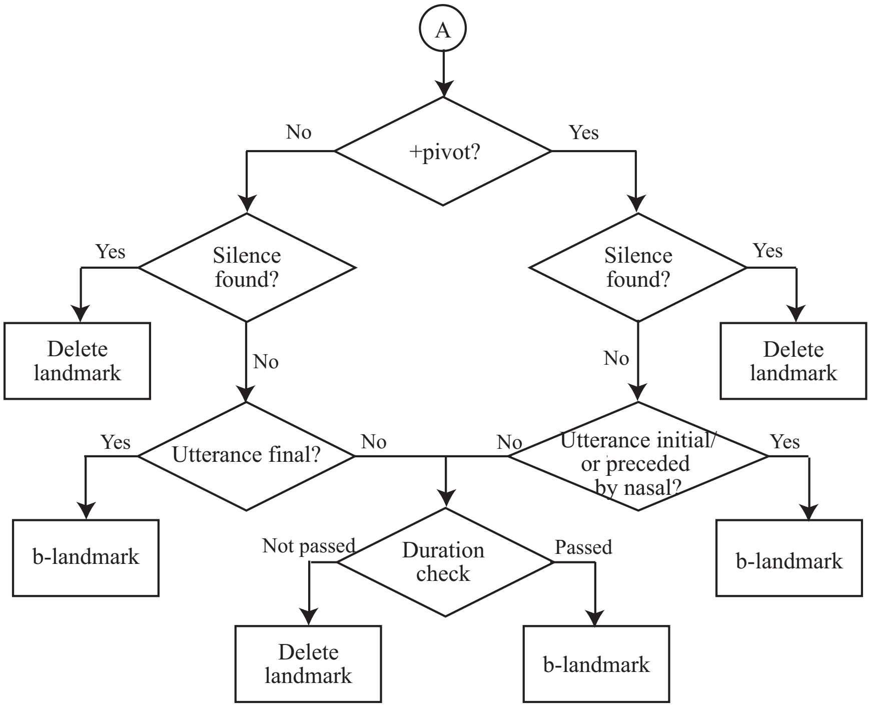


Figure B-3: A simplified flowchart that describes the decision process of the b-landmark sequence.

Bibliography

[1] Michael Ashby and John Maidment. *Introducing Phonetic Science.* Cambridge University Press, 2005.

[2] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):1001–1008, July 1989.

[3] Jeff A. Bilmes. Graphical models and automatic speech recognition. *Journal of Acoustical Society of America*, 112:2278–2279, November 2002.

[4] Jeff A. Bilmes. What HMMs can do. *Washington Tech Report*, February 2002.

[5] Nabil N. Bitar and Carol Y. Espy-Wilson. Knowledge-based parameters for HMM speech recognition. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 29–32, Atlanta, GA, May 1996.

[6] Suzanne Boyce, Ann Bradlow, and Joel MacAuslan. Landmark analysis of clear and conversational speaking styles. *Journal of Acoustic Society of America*, 118(3):1932, September 2005.

[7] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[8] Marilyn Y. Chen. Acoustic correlates of English and French nasalized vowels. *Journal of Acoustical Society of America*, 102:2360–2370, October 1997.

[9] Marilyn Y. Chen. Nasal detection module for a knowledge-based speech recognition system. *International Conference on Spoken Language Processing*, 4:636–639, October 2000.

[10] Noam Chomsky and Morris Halle. *The Sound Pattern of English.* Harper and Row, New York, 1968.

[11] Anne Cutler and Donald J. Foss. On the role of sentence stress in sentence processing. *Language and Speech*, 20:1–10, August 1977.

[12] Carol Y. Espy-Wilson. Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English. *Journal of Acoustical Society of America*, 92:736–757, 1992.

[13] Sadaoki Furui. On the role of spectral transition for speech perception. *Journal of Acoustic Society of America*, 80(4):1016–1025, October 1986.

[14] Oded Ghitza and M. Mohan Sondhi. Hidden Markov models with templates as non-stationary states: An application to speech recognition. *Computer Speech and Language*, 2:101–119, 1993.

[15] James R. Glass. *Finding Acoustic Regularities in Speech Applications to Phonetic Recognition.* Ph.D, Massachusetts Institute of Technology, 1988.

[16] James R. Glass and Victor W. Zue. Multi-level acoustic segmentation of continuous speech. In *Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 429–432, April 1988.

[17] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the Association for Computing Machinery*, 42(6):1115–1145, November 1995.

[18] David W. Gow and Peter C. Gordon. Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, 21:344–359, 1995.

[19] David W. Gow, Janis Melvold, and Sharon Manuel. How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing. *4th International Conference on Spoken Language Processing*, pages 66–69, October 1996.

[20] Mark Hasegawa-Johnson, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchhoff, Karen Livescu, Srividya Mohan, Jennifer Muller, Kemal Sonmez, and Tianyu Wang. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. In *Proceedings of IEEE ICASSP*, volume 1, pages 213–216, March 2005.

[21] Mark Hasegawa-Johnson, James Baker, Steven Greenberg, Katrin Kirchhoff, Jennifer Muller, Kemal Sonmez, Sarah Borys, Ken Chen, Amit Juneja, Karen Livescu, Srividya Mohan, Emily Coogan, and Tianyu Wang. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. Technical report, Johns Hopkins University 2004 Summer Workshop, January 2005.

[22] Shiro Hattori, Kengo Yamamoto, and Osamu Fujimura. Nasalization of vowels in relation to nasals. *Journal of Acoustical Society of America*, 30:267–274, April 1958.

[23] Andrew W. Howitt. Vowel landmark detection. *International Conference on Spoken Language Processing*, 4:628–631, October 2000.

[24] Daniel P. Huttenlocher and Victor W. Zue. A model of lexical access from partial phonetic information. *IEEE International Converence on Acoustics, Speech, and Signal Processing*, 9:391–394, March 1984.

[25] Roman C. Jakobson, Gunnar M. Fant, and M. Halle. Preliminaries to speech analysis. Technical Report 13, MIT Acoustics Laboratory, Cambridge MA, 1952.

[26] Aren Jansen and Partha Niyogi. A probabilistic speech recognition framework based on the temporal dynamics of distinctive feature landmark detectors. Technical report, University of Chicago, 2007.

[27] James J. Jenkins, Winifred Strange, and Thomas R. Edman. Identification of vowels in vowelless syllables. *Perception & Psychophysics*, 34(5):441–450, 1983.

[28] Allard Jongman. Duration of frication noise required for identification of English fricatives. *Journal of Acoustic Society of America*, 85(4):1718–1725, April 1989.

[29] Amit Juneja and Carol Y. Espy-Wilson. Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. In *Proceedings of the IEEE 9th International Conference on Neural Information Processing*, volume 2, pages 726–730, Singapore, 2002.

[30] Amit Juneja and Carol Y. Espy-Wilson. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In *Proceedings of International Joint Conference on Neural Network*, Portland, Oregon, 2003.

[31] Diane Kewley-Port. Time-varying features as correlates of places of articulation in stop consonants. *Journal of Acoustical Society of America*, 73(1):322–335, 1983.

[32] Kyung-Hoon R. Kim. Implementing the matcher in the lexical access system with uncertainty in data. M. Eng., Massachusetts Institute of Technology, 2000.

[33] Dennis H. Klatt. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of Acoustic Society of America*, 59(5):1208–1221, May 1976.

[34] Peter Ladefoged and Morris Halle. Some major features of the international phonetic alphabet. *Language*, 64(3):577–582, September 1988.

[35] Sharlene A. Liu. *Landmark Detection for Distinctive Feature-based Speech Recognition.* Ph.D, Massachusetts Institute of Technology, 1995.

[36] Sharlene A. Liu. Landmark detection for distinctive feature-based speech recognition. *Journal of Acoustic Society of America*, 100(5):3417–3430, November 1996.

[37] Steven M. Lulich, Asaf Bachrach, and Nicolas Malyska. A role for the second subglottal resonance in lexical access. *Journal of Acoustic Society of America*, 122(4):2320–2327, October 2007.

[38] Aaron Maldonado. Incorporating a feature tree geometry into a matcher for a speech recognizer. M. Eng., Massachusetts Institute of Technology, 1999.

[39] James L. McClelland and Jeffrey L. Elman. Interactive processes in speech perception: The TRACE model. In J. L. McClelland, D. E. Rumelhart, others, and eder, editors, *Parallel Distributed Processing: Volume 2: Psychological and Biological Models*, pages 58–121. MIT Press, Cambridge, MA, 1986.

[40] Paul Mermelstein. Automatic segmentation of speech into syllabic units. *Journal of Acoustical Society of America*, 58(4):880–883, October 1975.

[41] George A. Miller and Patricia E. Nicely. An analysis of perceptual confusions among some English consonants. *Journal of Acoustic Society of America*, 27(2):338–352, March 1955.

[42] Tarun Pruthi and Carol Y. Espy-Wilson. Acoustic classification of nasals and semi vowels. In *15th International Congress of Phonetic Science*, Barcelona, Spain, August 2003.

[43] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[44] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[45] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, October 1995.

[46] David W. Shipman and Victor W. Zue. Properties of large lexicons: Implications for advanced isolated word recognition systems. *IEEE International Converence on Acoustics, Speech, and Signal Processing*, 7:546–549, May 1982.

[47] Janet Slifka. Acoustic cues to vowel-schwa sequences for high front vowels. *Journal of Acoustical Society of America*, 118(3):2037, September 2005.

[48] Richard Sproat and Osamu Fujimura. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, 21:291–311, 1993.

[49] Kenneth N. Stevens. Evidence for the role of acoustic boundaries in the perception of speech sounds. In Victoria A. Fromkin, editor, *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pages 243–255. Academic Press, Inc., 1985.

[50] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, Massachusetts, 1998.

[51] Kenneth N. Stevens. Diverse acoustic cues at consonantal landmarks. *Phonetica*, 57:139–151, 2000.

[52] Kenneth N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of Acoustical Society of America*, 111(4):1872–1891, April 2002.

[53] Atiwong Suchato. *Classification of Stop Consonant Place of Articulation*. Ph.D, Massachusetts Institute of Technology, 2004.

[54] Atiwong Suchato and Proadpran Punyabukkana. Factors in classification of stop consonant place of articulation. *Interspeech*, pages 2969–2972, September 2005.

[55] Walter Sun. *Analysis and Interpretation of Glide Characteristics in Pursuit of an Algorithm for Recognition*. S.M., Massachusetts Institute of Technology, 1996.

[56] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.

[57] Yong Zhang. Toward implementation of a feature-based lexical access system. M. Eng., Massachusetts Institute of Technology, 1998.

[58] Victor Zue, James Glass, David Goodine, Michael Phillips, and Stephanie Seneff. The summit speech recognition system; phonological modeling and lexical access. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 49–52, Albuquerque, NM, April 1990.

[59] Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff. Acoustic segmentation and phonetic classification in the SUMMIT system. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 389–392, Glasgow, Scotland, May 1989.

[60] Victor Zue, Stephanie Seneff, and James Glass. Speech database development: TIMIT and beyond. *Speech Input/Output Assessment and Speech Databases*, 2:35–40, 1989.

[61] Geoffrey Zweig and Stuart J. Russell. Speech recognition with dynamic bayesian networks. In *AAAI/IAAI*, pages 173–180, 1998.