INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition*

Su-Lin Wu

TR-98-014

May 1998

Abstract

Incorporating the concept of the syllable into speech recognition may improve recognition accuracy through the integration of information over syllable-length time spans. Evidence from psychoacoustics and phonology suggests that humans use the syllable as a basic perceptual unit. Nonetheless, the explicit use of such long-time-span units is comparatively unusual in automatic speech recognition systems for English. The work described in this thesis explored the utility of information collected over syllable-related time-scales. The first approach involved integrating syllable segmentation information into the speech recognition process. The addition of acoustically-based syllable onset estimates [184] resulted in a 10% relative reduction in word-error rate. The second approach began with developing four speech recognition systems based on long-time-span features and units, including modulation spectrogram features [80]. Error analysis suggested the strategy of combining, which led to the implementation of methods that merged the outputs of syllable-based recognition systems with the phone-oriented baseline system at the frame level, the syllable level and the whole-utterance level. These combined systems exhibited relative improvements of 20-40% compared to the baseline system for clean and reverberant speech test cases.

*This report is a revised version of the author's thesis, which was submitted to the Department of Electrical Engineering and Computer Science on May 20, 1998 in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of California, Berkeley. This work was supervised by Professor Nelson Morgan. The thesis committee also included Professors Steven Greenberg, John Wawrzynek and Charles Stone.

Contents

List of Figures

List of Tables

Acknowledgements

How do I condense all my gratitude to a few short lines? In writing this, I am faced with reducing seven years of formative interactions with many wonderful people to a handful of paragraphs. What a difficult task! I am more appreciative of everyone than words can express, including those I miss naming explicitly.

I owe my deepest thanks to Professor Nelson Morgan. Without him this thesis would not exist. Besides being an unfailing source of terrible puns, he is a great research supervisor, career counselor and friend. Morgan always had a few timely words of feedback and encouragement for me even when he was at his busiest. I am sorry about all the Rolaids.

A number of the ideas in this thesis originated with Professor Steven Greenberg who shared with me his insights about how the human brain works. I very much appreciate the time Steve made for me and the discussion of everything from the syllable to scientific communication to the safety features of the Volvo.

My early graduate student career was marked with indecision. Professors Demmel, Yelick, Kahan, and Parlett gave me valuable advice and extended to me a good deal of forbearance during the early years. Later, Professors Wawrzynek, Stone, and Feldman contributed their different perspectives and served on my quals and/or thesis committees. I particularly thank Professor Wawrzynek for introducing me to the CNS group and for sharing with me his accumulated wisdom on computing and on homeownership.

The Realization Group at ICSI has a great tradition of cooperation and mutual support. Parts of the work in this thesis were undertaken in collaboration with either Michael L. Shire or Brian E. D. Kingsbury. I very much appreciated Mike's perseverance and good humor throughout our work together. I have benefited greatly from Brian's meticulous attention to detail and willingness to share his knowledge about feature analysis and about gourmet cooking. I am very glad that we could work together so well.

I heartily appreciate the help, both direct and indirect, of the inhabitants of ICSI (both past and present). In particular, I thank Jeff Bilmes, Dan Gildea, Nikki Mirghafori, Takayuki Arai, David Bailey, Geoff Zweig, Adam Janin, Warner Warren, Philipp Faerber, Alfred Hauenstein, Holger Schwenk and Florien Schiel for their help with forming my thesis. I especially thank Eric Fosler-Lussier for always having the script I need, for answering random engineering and linguistics questions, and for being willing to escape on french fry runs at a moment's notice.

The Realization Group has been fortunate to have had several "senior researchers" to help advise fledgling scientists. I have relied upon John Lazzaro for his unique understanding, Dan Jurafsky for dispensing both technical advice and Chinese food recommendations, and Dan Ellis for his technical communication acumen. Their perceptiveness on many levels has been extremely educational for me.

The SPERT boards were instrumental in the completion of this work, so I owe special thanks to the team that created them and the chief architect, Krste Asanović. David Johnson was indispensable in providing technical support for them. I also greatly appreciate Jim Beck's expertise with hardware at ICSI and his help with rebuilding my garage.

Kathryn Crabtree, Terry Lessard-Smith, Renee Reynolds and Nancy Shaw have been

enormously helpful with negotiating the various administrative hurdles of being a graduate student in the CS department and at ICSI.

I thank my parents for their encouragement and a certain amount of financial support throughout my undergraduate and graduate years.

My work has been primarily supported by a UC Berkeley Graduate Mentorship, an NSF Graduate Fellowship, JSEP grant F49620-94-C-0038, ONR grant N00014-92-J-1617, NSF grant 9712579. Additional funding came from a European Community Basic Research grant (Project Sprach) and ICSI.

Some aspects of my work reflect the early influence and instruction of the folks from Caltech, particularly Professors Charles Seitz, Mani Chandy and the late Jan van de Snepscheut. They and the graduate students that roamed the halls of Booth and Jorgensen were my initial inspiration to learn more about computer science.

I am indebted to Mr. Ohshima, Professor Daniel Chemla and all my SKA seniors and friends for helping me develop the strength of spirit to continue through and finish this project. They taught me, among many other things, the meaning of the proverb below.

Lastly, my husband, David Lipin deserves my deepest appreciation for his patience and unwavering support throughout this long endeavor. From proofreading, to listening to the tenth repetition of a talk, to doing my laundry, to giving me the best kind of moral support, Dave has always done all he could to bolster my confidence and energy. This document is dedicated to him.

A journey of a thousand miles begins with a single step.
— Chinese proverb

Chapter 1

Introduction

Shakespeare:

> Thou shalt be free
> As mountain winds: but then exactly do
> All points of my command
> To the syllable.
> Come, follow.

Automatic Speech Recognition:[1]

> Bell shall be free
> as mountain winds: but then exactly two
> all points of my command
> to this global.
> COM, follow.

This thesis is about putting the syllable back into automatic speech recognition.

For human beings, speech recognition is natural, robust, and efficient; speech is an integral part of communication between people. Every day the human speech recognition system performs feats of computation, filtering out ambient environmental noise from the speech signal, compensating and executing online adaptation for distortions due to speaker eccentricities, and rendering the result into words and sentences using complex contextual searches. Precisely how human beings perform speech recognition is not yet known. Although the physiological basis of hearing is slowly yielding to investigators, there is still much that is a mystery.

[1] An experienced dictation software user, Adam Janin (at ICSI), read this passage from Shakespeare's *The Tempest* to a commercial automatic speech recognition (ASR) package. The task is not entirely fair to the ASR system. It had never been used to transcribe anything like Shakespeare before. The system is well trained to this particular user, however, who employs it routinely in the course of his work.

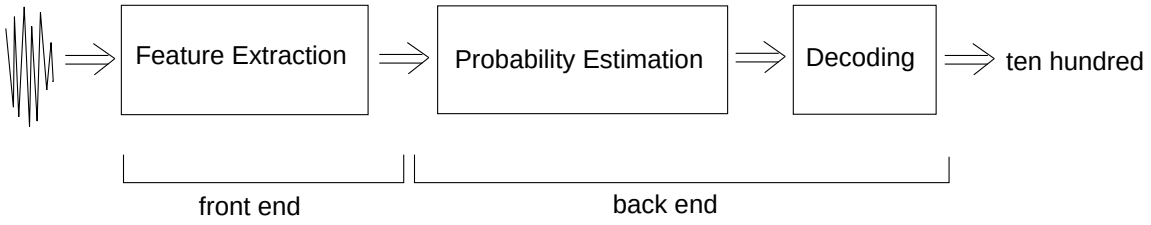

Figure 1.1: Diagram of the major parts of a typical automatic speech recognition process. Although language modeling is also a major part in ASR systems, it is not shown for simplicity.

Despite not completely understanding the neurophysiology of speech recognition, researchers have made substantial progress towards creating artificial methods of understanding speech, particularly over the last 30 years. Automatic speech recognition (ASR), however, is only just beginning to function well enough to be useful to the mass-market consumer. Designing and building artificial means of recognizing speech has proven to be difficult due to issues of complexity and robustness. Factors such as variability in speech, differences in speakers, environmental noise, confusibility of words, effects of prosodics, coarticulation, and perplexity trouble human speech recognition far less than the best automatic systems. Generally, a laboratory recognition system that performs well on artificial test data will have considerable, unforeseen difficulties with real voice input after deployment in the field– although both situations are equally intelligible to humans. Successful commercial applications often require additional tuning, data collection and analyses after field deployment to adjust systems for the differences between laboratory test sets and actual usage [188]. Ideally, this additional work should not be necessary.[2] Although speech recognition science has evolved greatly, there is still much improvement necessary before recognition by machines approaches the capabilities of human listeners.

1.1 Incorporating the Syllable into Speech Recognition

Figure 1.1 is a diagram of a typical automatic speech recognition engine.[3] A signal processing method first analyzes the spoken speech input. The process divides speech into regular time-frames and for each frame generates an array of numerical features. A probability estimator then uses the resulting acoustic features to generate posterior probabilities for each of the output categories, usually phones. The decoder, using dynamic programming, integrates the output of the probability estimation stage with additional information about the task to yield words and sentences.

Current speech recognition systems tend to employ signal processing and probability estimation techniques that focus on comparatively short sections of time. The use of small segments has the advantage of being able to encapsulate and distinguish minute changes in the speech signal. Long-time-span trends in the speech signal may be more difficult

[2] Robustness to unexpected characteristics of the speech signal is further discussed in Section 3.1.

[3] Current speech recognition technology is discussed in Chapter 3.

to identify explicitly and could be inappropriately weighted compared with other sources of speech information. Longer-time phenomena also afford fewer example patterns in a certain amount of training data for stochastic learning techniques. Since there is evidence that human speech has structure visible only at longer time intervals, approaches that consider longer time-spans present promising research avenues. To examine how to improve automatic speech recognition, the research described in this thesis looks to human speech perception for inspiration. In particular, the exploration focuses on the long-time-span unit known as the syllable, and the incorporation of certain types of syllable-based information into more standard speech recognition technology.

It is not known what sort of basic unit, or group of units, is used in human speech perception, but the syllable is one of a handful of strong contenders in a hotly-debated controversy.[4] Evidence from psycholinguistics and phonology suggests that syllable-level, long-time-span information (on the order of about 250 ms) may be crucial for speech understanding by human beings, particularly under adverse conditions [75].

For speech recognition by machines, information integrated over syllabic intervals may exhibit more robustness to unexpected characteristics of speech signals, that is, properties that are not represented well in the training information for the recognition system. While this integration can lose short-term detail, the combination of a syllable-based system with a phoneme-based system may have advantages over the individual systems alone due to their complementary strengths and weaknesses.[5] From an engineering point of view, a more natural organization of word pronunciation models, based on the syllable, may help reduce redundant computation and storage of words. The syllable may also provide a means for readily expressing long-time-span characteristics in the speech signal such as coarticulation, stress and other effects of prosody. In spite of the evident advantages, popular automatic speech recognition systems for English do not usually include the syllable as an explicit representational unit, though the concept of the syllable has played significant roles in ASR for other languages.[6]

This thesis describes two threads of research into incorporating information based on syllable-length time-spans into a recognition system for English. Illustrated in Figure 1.2, one thread explored using estimates of where syllables began (syllable onsets) as segmentation points; the other looked into improving the identification of words using information calculated over entire syllable-length intervals. By using a longer time segment, the machine learning algorithms in the recognition mechanism can potentially learn characteristics and relationships integrated over larger time spans of speech. Each experimental series culminated in combining the syllable-related information with a phoneme-based system to give an overall improvement, particularly with reverberant speech.

The concept of combining multiple sources of recognition information may very loosely approximate human perceptual processes. The experiments in this thesis that involved combining syllable-oriented and phoneme-based recognition are consistent with the idea that human speech understanding involves combining multiple representations of speech

[4] The role of the syllable in human speech perception is discussed more fully in Section 2.1.

[5] Combining systems is discussed in detail in Chapter 6.

[6] Chapter 2 discusses why typical ASR systems for English do not include an explicit representation of the syllable.

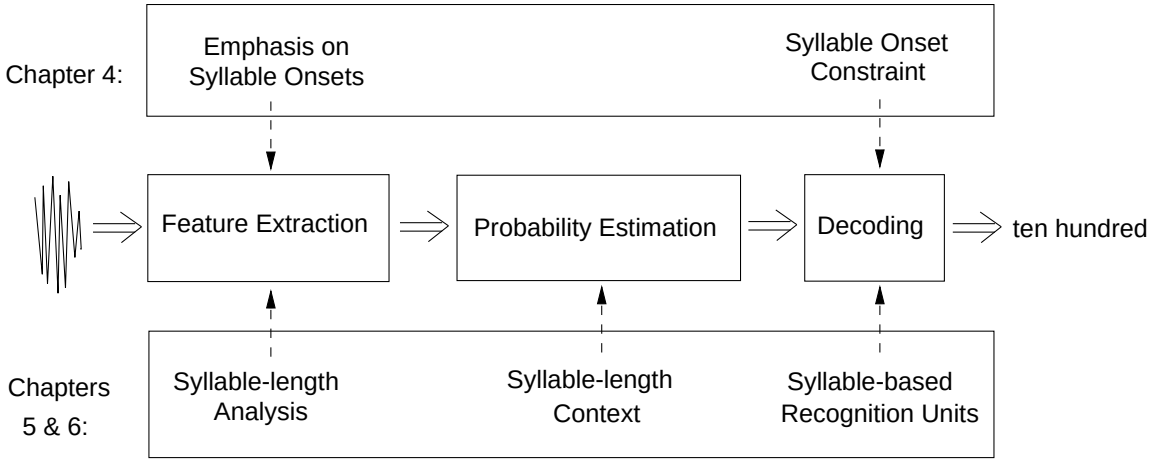

Figure 1.2: Diagram of the major parts of the automatic speech recognition process with proposed syllable elements indicated.

information, for example in the developing theories described in [74, 75, 78].

1.2 Results Overview

The goal of this thesis is to investigate the utility of syllable-level information in combination with standard phoneme-based techniques for improving recognition accuracy and robustness to unexpected properties. The course of this study involved the development of complete speech recognition systems for the Numbers corpus [30] incorporating long-time-span, syllable-oriented information.

The Numbers corpus comprises utterances from speakers saying numbers from their addresses, telephone numbers and zip codes over telephone lines in a conversational, unconstrained manner. Its relatively small size and its variety of acoustic qualities and speakers make it ideal for this work.[7] The Numbers task represents the problem of recognizing a small vocabulary task in a speaker-independent fashion under adverse conditions. Such a situation might arise in an information kiosk near a street, or for a speech recognition application accessed via a cellular telephone. Accuracy and particularly robustness to unexpected variations in the speech signal are important factors in the usability of such applications by the general public.

The first series of experiments focused on using the beginnings of syllables (syllable onsets) as cues for segmenting speech at the syllable level. The experimental methodology necessitated the design and implementation of a syllable-based decoder to incorporate syllable onset information. Pilot experiments (with correct onsets) showed that knowledge of the beginning points of syllables had the potential to improve performance by as much as 38%. Using estimated syllabic onsets (from actual speech [184]) in a phoneme-based system

[7] The Numbers corpus is discussed in more detail in Section 3.2.

realized a 10% relative improvement in accuracy over the baseline system.[8]

The second series of experiments focused on developing recognition systems using syllable-based recognition units and acoustic features computed over syllable-length intervals [80, 105, 107].[9] The analysis of the experimental results showed that while these systems did not achieve a significantly lower error rate than the phoneme-based system on the test data (several of the systems performed significantly worse), in many instances they could successfully recognize an utterance that the baseline, phoneme-oriented system could not. The syllable-based systems had strengths and weaknesses that were somewhat complementary to those of the phoneme-based system.

The final series of experiments involved integrating the syllable-based systems with the phoneme-based system using combination strategies that merged two recognition systems at one of three different stages of the recognition process. Each of these strategies resulted in somewhat different recognition performance, but in the best case (syllable-level combination) these experiments achieved an approximately 20% relative improvement in word error rate for clean speech and a roughly 40% relative improvement in word error rate for reverberant speech.[10]

The experiments illustrated that, for the Numbers corpus, using syllable-based information in combination with traditional phone-based information enhanced the overall accuracy and robustness to reverberation of the ASR system. Thus, syllable-based information and combination strategies emerge as viable areas for future ASR research, both individually and when used together.

1.3 Thesis History and Outline

This project originated as an effort to vectorize the speech decoding algorithm for a multiple-unit version of the SPERT vector microprocessor system [201], as suggested by Professor John Wawrzynek. Vectorization is most efficient when an algorithm accesses contiguous memory locations in succession; the introduction of pointers and conditionals impairs the achievement of maximum performance. Professor Steven Greenberg suggested that syllables may have many advantageous organizational, computational and storage properties, in addition to being a fundamental unit of human speech recognition. Thus, to make the decoding process easier to vectorize, I began investigating using syllables instead of words or phones as a basic organization unit for recognition. Professor Nelson Morgan suggested the focus on syllable-time-span units and features in speech recognition as a refinement of the basic broad approach, and he also suggested the combination strategy. The results of the subsequent investigation are described in the chapters that follow. While I myself have not returned to the question of vectorizing speech recognition algorithms, some of the work described in this thesis has natural extensions to parallel and concurrent processing. In particular, combining multiple systems is inherently concurrent and Eric Fosler-Lussier's

[8] These experiments with syllabic onsets are discussed in Chapter 4.

[9] The development of these syllable-oriented systems is discussed in Chapter 5.

[10] These results with combining systems are discussed in Chapter 6.

two-level decoder implementation[11] for combining multiple streams is highly parallel at the word level [54].

The rest of this document begins in Chapter 2 with a discussion of syllables as they pertain to speech recognition both for humans and for machines, along with a summary of past and contemporary work along similar lines by other researchers. Chapter 3 contains a short overview of the history and state of the art in speech recognition and includes technical details about the ICSI system which serves as the platform for the experiments of this thesis. Chapter 4 describes efforts to use syllable onset information to reduce word errors in a phoneme-based recognition engine. This work was previously published in [210]. The next chapter, Chapter 5, details the ideas, design and implementation of several speech recognition systems which incorporate syllable-related elements, and reports the performance of each. Chapter 6 relates the analysis of the differences and similarities in the experimental systems and how a syllable-oriented speech recognition system was combined with a phoneme-based, comparatively short-time span system. The chapter details encouraging experimental results with both clean and reverberant speech. Part of this work was previously published in [209]. Chapter 7 contains a summary of this project, discusses the advantages and disadvantages of the syllable-based system and draws some possible conclusions from this work. In particular, it contains some reflections on issues pertaining to the extension of these ideas to large vocabulary tasks.

[11] Based on Sakoe's algorithm [171] as described in [159].

Chapter 2

The Role of Syllable-based Information

> Syllables have been described as thrusts of the chest muscles of respiration, peaks of sonority, pulses of sound energy, necessary units in the mental organization and production of speech, a group of speech movements, and a basic unit of speech perception. [147]

This chapter discusses the syllable as a possible basic unit of speech recognition, for which there is some empirical psychoacoustic support in the case of humans and some engineering justification in the case of machines striving to imitate human abilities. For the purposes of the research described in this thesis, a "basic unit" of speech recognition is the intermediate form of speech information around which much of the recognition processing is organized for human beings or for machines. The general opinion of phoneticians and psycholinguists is that there is indeed such a unit with relatively few distinct types.[1] For this research a basic unit is ideally an output of acoustic-phonetic processing and an input to the lexical processing stages. A significant portion of the processing operates on this unit. A basic unit must be small enough to express variety in the manifestation of speech without an explosion in the number of representations, yet be large enough to be computationally efficient and possess properties that allow it to function as an organizational unit for lexical access.

Over the last 40 to 50 years, researchers have proposed many different types of intermediate units. Some of the possibilities include sub-phoneme units, phones, phones with right or left context, biphones, diphones [178] and variations [36], dyads or transemes [44], avents [134, 136, 208], triphones [7, 175], demisyllables [60], syllables [59], whole words, and phrases.

Current research in psychoacoustics and psycholinguistics suggests that the syllable might be a basic unit of human speech perception. Researchers have hypothesized that the syllable, or a related long-time-span unit, may be the key to how humans process and integrate information from the speech signal. From an engineering standpoint, the syllable

[1] Frauenfelder reviews some of the current thinking about the interface between the acoustic-phonetic level and the lexical level in [55].

may be an efficient, useful intermediate speech unit that can potentially reduce redundant computation and storage in automatic speech recognition. Higher-level knowledge of spoken language can be expressed fairly naturally and compactly in terms of syllables; yet syllables are relatively short and have constrained characteristics. In spite of the potential benefits, the syllable is not often an explicitly represented concept in modern automatic speech recognition (ASR) systems for English.

There are many questions still unanswered about the role of syllables in human language, and many practical difficulties in using syllables as units of automatic speech recognition. This chapter discusses the role of the syllable in speech recognition, for both humans and machines. Section 2.1 first reviews research literature about the syllable's role in human speech perception. Section 2.2 details the properties of syllables in conversational American English. The third section of this chapter discusses the history of the idea of using the syllable as an intermediate speech unit in recognition by machines and its advantages and disadvantages. This chapter concludes with a discussion of the relationship of this material to the experiments described later in this thesis.

2.1 Syllables in Human Speech Recognition

Much research into automatic speech recognition by machines takes guidance from the human speech communication mechanism. Even though researchers do not necessarily aim to completely mimic the human process, understanding some measure of speech recognition by humans is relevant for any study. How are syllables used by the human speech recognition system? The answers to this question are quite controversial and strongly debated in the linguistic and psychoacoustic communities.

Continuing Kahn's analogy [101], studies aimed at deciphering the role of the syllable in human perception can be thought of as akin to measuring the movement of an airplane in turbulence and attempting to infer Newton's laws of motion and gravitation. Kahn blames much of the controversy among linguists on the nature of the syllable on disagreement about which facts are most fundamental and require explanation first.

The literature on the nature of the syllable in human speech is overwhelming in size. The exposition in this section is limited to a concise summary of current thinking about the perception of syllables in human speech recognition as pertains to the area of automatic speech recognition and the research described in this thesis. The focus is on perceptual studies, rather than the study of speech generation or production, since the eventual aim is automatic speech recognition by machines. The more abstract arguments in the linguistic research community are not covered.

2.1.1 Syllable as Basic Unit

Over the past few decades it has become accepted wisdom that the process of mapping between acoustic signals and sequences of perceived sounds in humans is complex; listeners do not process speech in a linear fashion, one acoustic segment at a time like "beads on a string." Instead, acoustic information at one instant of time is relevant to several phonetic

segments, and a single phonetic segment affects a broad region of acoustic information. Speech perception is therefore highly context-dependent. The research community also has come to realize that recognizing a word is more involved than a simple mapping between acoustic-phonetic properties and an entry in the mental lexicon. It is more likely to be a rather complex, non-linear process with many heterogeneous subprocesses [81]. That the perceptual sequence is not strictly bottom-to-top is supported by the observation that listeners use high-level context to resolve confusions. For example, Savin and Bever presented words such as "cat" and "hat," mixed with background noise sufficient to cause listeners, hearing these words without context, to mistake one for the other. When given the context "it is time to feed the," listeners reported hearing "cat," even if this word is incorrect [173]. In a developing theory of speech perception, Greenberg proposes that the speech recognition mechanism is a many-layered process with dozens of coarse representations that combine in a non-linear manner in order to effect the robust and efficient speech recognition ability of human beings [74].

The community at large feels that no single perceptual unit (based on phonemes, syllables, or words) has proven to be the ideal basic unit for all auditory situations [144]. Nevertheless, researchers generally hypothesize that a few representations dominate the organizational units in the human speech perception system. The two major contenders for principal sublexical perceptual unit are the syllable and the phoneme. Although there is probably no single unit that is the sole representation of sound in speech processes, there is considerable evidence that the syllable is a major representative form, arguably more so than the phoneme. It has been suggested that many prosodic properties such as pitch, accent and stress are most naturally expressed in terms of syllables. Some researchers hypothesize the syllable to be the primary unit of segmentation in speech and the basic unit of lexical access in the human brain.

One point of evidence in favor of the syllable comes from the observation that syllables are identified more easily than phonemes by untrained, naive listeners. Rozin, Poritsky and Sotsky found that children with reading disabilities could be taught to read using Chinese characters [168]. They attributed their success to the mapping of the characters to a higher level than the phoneme, and suggested that the syllable be used to facilitate reading instruction. Anecdotal evidence suggests that normal children are able to identify syllabic segments at a younger age than phone segments. Mehler, Dommergues and Frauenfelder speculate that phones are identifiable only after subjects learn to read and write with graphemes (i.e., letters of the alphabet) which can then be related to phonemes [127]. A variety of speech pathologies can be characterized more easily in terms of syllables than with phoneme-based expressions. Syllable-based explanations have been used in explaining the misarticulations of children, hearing deficiencies, and other anomalies [147].

Many studies have tried to distinguish the roles of units such as syllables and phonemes in human speech recognition. Studies aimed at determining the basic unit of speech perception have often produced inconclusive results; it is very difficult to formulate experimental setups that isolate different human perceptual factors. Because of the complexity and inseparability of the elements in the speech recognition process in humans, experimental data can be subject to differing interpretations. Also, conclusions are further impeded by the difficulty of extrapolating from laboratory conditions to ordinary everyday speech. The experimental

results and conclusions obtained by researchers are often flatly contradictory. Rather than taking any single experiment as definitive, this thesis sides with the view that the body of research regarding the syllable must be considered as a whole. The literature supports the syllable as a prominent component of the human speech perception system, given this holistic view. Below are summaries of some of the more popular paradigms and results relevant to the issue of automatic speech recognition and the work described in this thesis.

2.1.2 Syllable Identification

One of the more popular methods for studying the basic identification units of speech in humans is the "reaction time" experimental paradigm. This sort of experiment assumes a correlation between how quickly a human subject can recognize and respond to speech stimuli and how fundamental the recognized unit is to the perception process. These experiments may take many forms, but they tend to follow the framework of asking subjects to react as quickly as they can to the perception of syllabic or phonetic targets in an artificially crafted carrier utterance. Experimenters hope that the correlation between the chosen experimental variable and reaction time is elemental and simple to characterize.

This methodology is difficult to formulate in an unassailable manner that is generalizable to greater context. Critics assert that a variety of effects are not addressed in the instantiations of this experimental paradigm. Some experiments use a very small set of specific syllables and phones. Others use sets of target syllables or phones that may not be equivalently perceived by humans owing to effects such as linguistic level or acoustic characteristics, which are independent of their relationship to the basic perceptual unit. For example, syllables of differing durations, or which have onsets with differing speeds of transition, may obscure the experimental data if categorized similarly. There are also extensive criticisms of the criticisms. The variance in experimental observations and the contradictions in conclusions brings to mind the old fable of blind men, each feeling a different part of an elephant and attributing radically different characteristics to the animal.

Nonetheless, there is a considerable body of literature based on this paradigm in the search for the perceptual units of the human speech system. The experimental results can be roughly divided between those that find faster reaction times for syllables than phones and those that find faster reaction times for phones than syllables. Massaro summarizes several studies that found subjects recognized syllables faster than phones [124]. Among these are the works of Savin and Bever [173], who found subjects responded faster to targets that were complete syllables than to phones from a syllable, and of Warren [197], who reported that identification times for monosyllabic words and nonsense syllables were shorter than for phone clusters which were in turn shorter than for individual phones. Warren also observed that the majority of his subjects failed to recognize the /n/ phone in the word "and" when asked to identify /n/s in the context of a sentence. This suggested that the monosyllabic word "and" had a perceptual identity separate from its constituent phones. Several reaction-time studies in multiple languages are summarized in [180, 181, 182] which indicate faster reaction times for syllables in French, Spanish and Portuguese, though Segui suggests that the syllabic boundaries of English may be too ill-defined to extend the general conclusion to English.

Frauenfelder summarizes a considerable body of literature pertaining to the "interface between acoustic-phonetic and lexical processing," concentrating on reaction time experiments [55]. In Frauenfelder's summary, the opposing viewpoint of the phoneme triggering faster reaction times than the syllable is supported through experimental evidence by several studies. In their experiments, Norris and Cutler attempted to enforce full analysis of targets by supplying stimuli that differed from the target only by one phone, for example "bat" and "bam." They claimed that the faster reaction times for syllables were merely artifacts of the experimental design of others [143].

What do these contradictory findings mean and what accounts for them? Kahn warns not to regard all facts as equally important and deserving of interpretation [101]. To generalize about the nature of the syllable through a relatively small number of results from a single experimental paradigm is a fundamentally unsafe methodology. The results of these experiments depend critically on the experimental procedure. The difference in reaction times for "faster" versus "slower" is often less than 150 ms, and removing the influence of higher order thinking is acknowledged to be very difficult. Therefore, any conclusions and interpretations from the results of these reaction time experiments must be taken in context, as a small part of a broader body of empirical results from diverse experimental paradigms.

Reaction time experiments and the theories developed around them tend to regard the question of sublexical perceptual units as a choice between phoneme-based units and syllable-based units. The view that there are multiple sublexical, basic units appears to be less popular among researchers investigating this issue. This multi-unit view, however, is the most likely to account for all the disparate experimental findings and it is the perspective underlying this thesis.

While reaction time experiments have been very popular and have generated conflicting conclusions, other experimental paradigms have been used in studies that contribute to the overall assessment of the syllable as a candidate basic identification unit. These experiments depart from reaction time experiments and instead follow more of a "masking" model. Researchers manipulate the stimulus with some sort of interference, either through direct obfuscation or through indirect means, and then assess the intelligibility of the resulting signal by asking subjects to identify the targets they perceive.

Psychoacoustic experiments where researchers replace phones or other short sections in stimulus utterances by noise or silence address the importance of the individual phone and other segments of speech. In "silent-center" experiments, the nucleus of a syllable is replaced with silence with minimal effect on recognition accuracy. This is known as a type of auditory illusion called "phonemic restoration." Subjects often do not notice a phone is missing if the interval has been replaced with some sort of filler, such as white noise or a cough, thus lending support to the idea that humans infer phonemes as parts of syllables after the syllable has been categorized.[2] In [96], the authors designed a variation of the silent-center experiment where the two halves of a syllable, separated by silence, are provided by different speakers, for example male and female. The results indicate that syllable onsets and offsets, taken as pair, are sufficient to derive the vowel in a syllable and

[2]For a summary, see [198].

that continuity of speech formants[3] is not necessary. With a related experimental paradigm, Furui's experiments with the identification of Japanese syllables also indicated that vowel nuclei are not needed for accurate vowel-of-syllable perception [64]. He found that the same initial part of the syllable contained crucial information for both vowel and consonant identification.

Massaro's summary [124] also discusses the experiments of Cherry and Wiley, in which by passing to subjects only the strongly voiced, high energy speech sounds, they were able to degrade speech perception to very low intelligibility [25]. The speech was rendered more intelligible, almost up to the level of the original, by adding a low level of white noise into the silent gaps between the voiced speech sounds. Even though phonetic information was not added, the insertion of noise made the speech more like normal speech in nature. Greenberg interprets their report as evidence that the addition of the noise restored a portion of the modulation spectrum of the original acoustic information, and that this patterning bears a similarity to the syllabic temporal patterning of speech [77].

Such studies indicate that human perceptual processes make considerable use of information spanning larger time-units than the single phone, and that a particular phone constituent is relatively unimportant. Ganapathiraju *et al.* discuss an analysis of data from hand-transcriptions of Switchboard data [76, 186] that showed that the deletion rate for syllables was below 1% while the deletion rate for phones was 12% [67]. The authors took this as supporting evidence for the relative stability of the two types of recognition unit.

Warren asserts that acoustic elements form "temporal compounds" and that the human perceptual system can identify these compounds more readily than constituent sounds [199, 200, 198]. Warren *et al.* found these temporal compounds to be longer than the typical phone length through experiments with loud, clear, repeating vowel acoustic elements concatenated together. They presented sequences of concatenated vowels to listeners and asked the listeners to identify the order of the phones. The studies showed that if the individual vowel durations were about 200 ms, listeners accomplished the ordering task easily; however if the vowel durations were below about 125 ms then the ordering task was impossible. Since the average duration of a phone in speech is about 70 ms [79], this points to a larger temporal effect than phoneme identification. Further studies showed that while subjects could not reliably give the ordering of phones at below 125-ms levels, they could nevertheless easily distinguish between differing sequences even when the durations of individual items were as short as 10 ms. Additionally, the experimenters found that subjects perceived these streams of concatenated vowel sounds in terms of syllables and words with consonants not actually present in the acoustic signal. These findings indicated that the initial stage of speech perception is not phoneme recognition, and that resolution into sequences of constituent phones is not necessary for accurate speech recognition.

Massaro [124] also notes that the speech synthesis community has found that using concatenated phones is ineffectual for producing intelligible speech. In contrast, speech synthesis achieved early success using units that were at least one-half syllable in length. For example, AcuVoice Inc., San Jose California, uses stored recordings of syllables in one

[3] "Formant" refers to resonances of the vocal tract as evidenced by speech sounds. The term also refers to formant frequency.

of the more natural sounding text-to-speech systems [32]. Although some researchers used this observation as evidence of merging phonemic influences, Massaro interprets this as additional support of the syllable as a basic perceptual unit.

Although there is considerable unresolved controversy, experimental evidence weighs in favor of the syllable playing a substantial role in the identification process in human speech perception.

2.1.3 Syllable Segmentation

Separate from the issue of how well humans can recognize speech fragments is the question of the fundamental time scale on which the recognition process operates. At one extreme, one could suppose that the auditory system processes acoustic information in real-time and that sound information streams into the brain which performs immediate and continuous analyses of the world without any segmentation. Or, one can imagine the brain with an enormous buffer that takes in the information from the auditory system and stores it until the acoustic input is "finished" and then transfers the whole pattern to the brain. For human recognition to distinguish individual speech items, some kind of buffer in the brain is assumed to hold the initial part of a pattern long enough for the token to "complete" and for analysis and recognition to occur. There are several indications that this pattern buffer is syllable-length in duration.

O'Shaughnessy summarizes a number of perceptual experiments that implicate a syllabic-duration perceptual unit, in work that is closely related to reaction-time experiments [148]. Among these are "shadowing" experiments (where subjects try to repeat what they hear as quickly as they can), in which delays by subjects are typically the length of a syllable or word. These results put the upper bound on the size of the processing buffer at about the length of a single syllable or word.

In the 1970s, Massaro used recognition-masking experiments to determine the perceptual unit of analysis in the human speech recognition system [123, 124, 125]. The general paradigm in these experiments involved the presentation of pairs of artificially crafted stimuli (tones or speech sounds) separated by a variable silent interval. The extent to which the second "masking" stimulus alters listeners' perception of the earlier target is known as the "backward-masking effect." Massaro took the correlation between the human subjects' ability to recognize the initial stimulus and the amount of silence between the first stimulus and the masking stimulus as an indication of the length of the perceptual unit. If the masking stimulus is presented too close to the initial stimulus (i.e., within the perceptual processing interval) and it can not be integrated, the masking stimulus interferes with the storage and analysis of the initial stimulus image in preperceptual form. Analogous temporal phenomenon are known to exist for human visual processes. One version of this experimental paradigm uses pure tones produced by an oscillator, which has the advantage that the speech and language centers of the human brain are less involved. This can reduce the amount of indirect supposition about perceptual processes.

In Massaro's studies, he found that subjects' recognition performance of the initial stimulus improved as the silent interval between the target and the masking stimulus was

increased to 200-250 ms, after which performance reached a plateau. He concluded that preperceptual auditory storage and processing does not exceed 250 ms. From this Massaro conjectured that the perceptual unit must be the syllable, which has an average duration of roughly 200-250 ms in conversational speech, though he notes that longer syllables probably become multiple perceptual units. It is important to note that Massaro defined the syllable in terms of *duration*. This differs from the definition that used by linguists in most of their experiments, which does not respect the temporal extent of a syllable. These findings are consistent with the conclusions of both Todd and Warren. Todd mentions that humans are sensitive to time intervals of about 300 ms, intervals that match the upper limit of the duration of syllables fairly well [189].

Massaro speculates that the effect of phonemic restoration, where noise can be imperceptibly substituted for a phone in a syllable, can be explained in terms of preperceptual auditory image storage. Since the human subjects believed they actually heard phones not present, this processing must occur well in advance of the conscious level. Massaro suggests that the inserted noise was probably grouped into the perceptual unit of the syllable during preperceptual storage. Since it did not disrupt the storage and analysis of the syllable, it was incorporated into the classification of the unit. On the basis of the remaining relevant acoustic features, subjects could infer the correct syllable, then inversely identify the phonemic constituents.

Interrupted (or "gated") and alternating (or "ear switching") speech experiments have also shown that the critical duration for intelligibility appears to be about the length of a syllable. In the interrupted speech experiments, partly summarized in [124], half of the speech signal was eliminated by replacing intervals of speech with silence, where the researchers varied the length of the intervals. These experiments showed somewhat varying lengths for the duration of the perceptual unit, but all corresponded approximately to the duration of syllables rather than phones. In alternating speech experiments, the speech shifted from ear to ear of the subject through headphones. When the alternation was near the syllabic rate, the recognition abilities of the subjects were disrupted. Much faster or much slower alternation rates had less effect on speech perception by human subjects.

Segui reports that subjects identify target sequences more easily if they are contained within the same syllable, rather than spread across syllables, supporting the idea that the perceptual mechanism segments the speech signal into syllable-like units [181]. In the work of Mehler *et al.* [127] and Cutler, Mehler, Norris and Segui [38], subjects were asked to identify a consonant-vowel (CV) or consonant-vowel-consonant (CVC) target in carrier words with either CV or CVC structures as the first syllable (for example, in French, detecting /pa/ or /pal/ in "pa-lace" or "pal-mier"). French subjects identified targets that formed a complete syllable faster than targets spread across different syllables, suggesting that the syllable was indeed a unit of speech segmentation. English and French subjects receiving both English and French stimuli revealed that French subjects showed a syllable effect in both languages and English subjects showed a syllable effect in neither. Cutler *et al.* speculated that this could be because English contains a considerable amount of ambisyllabicity.[4]

[4]Ambisyllabicity, the sharing of a single phone segment between two separate syllables, is discussed further in Section 2.3.3.

Miller and Eimas [130] showed more evidence of the effect of duration on recognition; they demonstrated experimentally that the identification of phonetic targets is dependent on the length of the carrier syllable, not just on the phone itself. The work described in [153] also showed that contextual effects such as duration affected the perception of stimuli and the identification of the initial sound, indicating that long-time span information on the order of syllable-length intervals influences perception, even for non-speech stimuli (with speech-like qualities).

These experiments, taken as a whole, suggest that the syllable-length interval, aside from the actual speech content contained within, plays a crucial role in human speech perception.

2.1.4 Syllables in Lexical Access

At higher levels of human speech processing, the formulation of sound experiments that test the role of processing elements becomes increasingly complex. It is very difficult to draw conclusions from such indirect evidence and to separate the many contributing functions. This section discusses lexical access, i.e., how smaller units are mapped to words and sentences for ASR. Since the work in this thesis does not directly address this problem, only a few representative experiments are presented.

Reaction time experiments have been used as support for the hypothesis that the syllable is the primary unit of lexical access. Segui summarizes studies investigating this idea and uses his experiments in French [180, 181], partly discussed above, as support. In these studies the subjects appeared to identify the first syllable of an isolated polysyllabic word before the lexical access occurred.

In the studies of Warren *et al.*, described earlier, subjects presented with streams of concatenated vowels recognized these in terms of syllables and words, perceiving illusory consonants as required to organize the sounds [199, 200, 198]. The syllables recognized were always legal syllables in English, the subjects' native tongue, though the syllables taken together were not necessarily legal words. Warren infers that humans have an internal "syllabary" (a set of acceptable speech syllables) and use this for lexical access.

Anecdotal evidence suggests that humans recall words as a sequence of syllable-level patterns rather than by individual phones. Ladefoged [113] mentions that in the history of writing, many languages have emerged in which there is one symbol per syllable, as in Japanese. From an intuitive standpoint, 'tip of the tongue' phenomena, where humans partially recall words by their syllable structure even though the phonemic constituents themselves are not retained, and word substitution slips, in which the number of syllables in the word is preserved, also imply a syllabic basis for lexical access. Ladefoged talks about the specific patterns that occur in slips of the tongue; a syllable initial consonant exchanges with a syllable initial consonant, or a syllable final consonant exchanges with a syllable final consonant. Such syllable-oriented observations lend further weight to the conjecture that the syllable is a basic unit of lexical access in human speech perception.

2.1.5 Summary

This section discusses evidence for the syllable as a basic unit of human speech perception. There is considerable disagreement among researchers as to whether the syllable or the phoneme is more elemental to the speech recognition process. This disagreement has been fueled by conflicting results in reaction time experiments, which have supported both positions. Other experimental results from different methodologies provide additional support for the syllable as a basic unit for the identification, segmentation and lexical access of speech, without entirely superseding the phoneme. The viewpoint of this thesis is that both units, the syllable and the phoneme, play basic, coordinated roles in the phenomenon of human speech recognition.

2.2 Syllables in American English

> **Syllable** A unit of speech for which there is no satisfactory definition. [113]
>
> There is no common boundary at which the syllables join, but each is separate and distinct from the rest. [2]

2.2.1 Definition of Syllable

Despite a lengthy discussion of the role of syllables in human speech recognition, a rigorous definition of the syllable has yet to be presented owing to the lack of an adequate specification. Engineers, however, need a functioning description in order to implement speech recognition systems. Syllables are notoriously difficult to define precisely, especially in American English, although human beings appear to have an intuitive understanding of them. It is agreed that, in loose terms, a syllable is constructed about a nucleus that is usually the most intense component, and generally the sole obligatory constituent. Most syllables begin with an onset which typically consists of a single consonant, but may contain two or three consonants. Many syllables end with a coda of a single consonant, but codas can also comprise two or three consonants.

Definitions striving for more technical accuracy are problematic. Every definition seems to have exceptions and caveats or is unsatisfying for practical implementation. For example, consider the following two popular definitions: 1) A syllable is a vowel between optional consonant clusters. This, the most popularly understood rule, has many exceptions, since a syllable does not necessarily contain a vowel. A syllable can instead have a "syllabic consonant" that functions as the nucleus of the syllable, for example, the /l/ in "noodle" or the /s/ in the onomatopoeia "psst." 2) Syllables correspond to peaks of sonority. Sonority is roughly analogous to the energy contour. Peaks of sonority are therefore analogous to regions of greater sound energy and are thought to correspond to the nuclei of syllables. This definition allows consonants to take the place of syllable nuclei [147], but the sonority-based specification is vague in some cases and can lead to confusions. For example, the unmistakably monosyllabic word "spa" is considered by some to have two peaks of sonority [113].

Mechanically segmenting speech into syllables is also difficult. The "maximum onset principle," defines the onsets of syllables (the initial consonant clusters) to be as long as possible within the context of the word. For example, the word "estate" would be pronounced as "e-state," according to this rule. The /s/, however, often sounds as if it is shared between syllables. Speakers can pronounce the word as "es-tate," if the first syllable is stressed, an exception to the maximum onset principle. Treiman and Zukowski note that for the word "estate" the maximum onset principle conflicts with the sonority definition, that the word "state" does not exhibit a rise in sonority from the onset to the nucleus [192]. For the experiments in this thesis, segmentation of phonemic transcriptions follows the complex hierarchical set of rules in [101], which may not be perfect in every instance, but can be consistently applied.

The list of exceptions to postulated rules continues indefinitely. There is a large number of almost-complete technical definitions of "syllable," and phonetic segmentation algorithms. This is largely due to the continuing debate about the exact nature of the syllable and its role in human speech recognition. Ladefoged [113] and Ohde and Sharf [147] further discuss the vagaries of human syllabification and the shortcomings of various attempts at defining syllables.

People intuitively understand the concept of the syllable and can usually identify the gross syllabic characteristics in a word, such as how many syllables there are and the approximate locations of their boundaries. But listeners cannot precisely describe how they accomplish this feat. Even for words such as "meal" where the number of syllables is uncertain, non-experts can discuss, without technical detail, the ambiguous nature of the number of syllables of this word, merely by sounding the word out and using primitive quantitative arguments. In contrast to the phoneme, the human concept of the syllable seems to be universally understood, even by untrained and naive listeners.

As a technical term, the word "syllable" is highly over-used ("overloaded," in programmers parlance). Linguists, phoneticians, engineers and other researchers use the word "syllable" and can intend very different meanings. From an abstract point of view, a syllable necessarily contains a group of phones and has some acoustic manifestation. A syllable can be discussed in terms of the properties of its constituent sounds, or in terms of its production by a speaker. From a perceptual point of view, a person can believe he hears a syllable that was actually omitted, for example in the case of a rapid speaker deleting the end of words or whole function words such as "a," "of," "to" and "the." The listener, unless allowed to listen very carefully and/or view spectrograms, perceives syllables in mirage form, where the canonical acoustic cues normally associated with the syllables are not present in the speech. Listeners can use these illusory syllables for lexical and semantic access.[5] These observations suggest that the syllable exists as a perceptual concept apart from a purely linguistic definition. As alluded to during the discussion of the importance of syllabic tem-

[5]The perception of these illusory syllables does not happen as easily or frequently as with the perception of unexpressed phonemes. In the word transcriptions of the 4-hour phonetically transcribed subset of the Switchboard Transcription Project [76], the most frequently deleted words (which also were monosyllabic) were "a," "of," "to" and "the." Specifically, "a" was deleted in 2.1% of its total occurrences, "of" was deleted 1.3% of the time, "to" was deleted 1% of the time and "the" was deleted in 0.5% of its total instances. These percentages are considerably smaller than those reported for phone deletion in [78]. Further discussion of illusory syllables can be found in [52, 53].

poral structure, some syllables that are linguistically defined as single units fall outside the norm for syllabic duration. These should, perhaps, be thought of as two or more syllables from the durational point of view, for example, the monosyllabic word "scrounged." For practical purposes, engineers define syllables differently again with their own criteria and rules, often denoting a unit that has only some vague resemblance to syllables as human beings intuitively understand them.

In this background and overview, the word "syllable" is used in many different ways. Deviations from the canonical, linguistic view are noted in the text. For the purposes of the experiments in this thesis, the syllable was defined precisely from a purely functional, engineering point of view, though inspired by the syllable from the more abstract acoustical and perceptual standpoint. In these experiments, syllable unit targets in the recognition task, Numbers,[6] were defined using the half-syllable units listed in Appendix A.2. The definition used is not perfect, but it is at least consistent. The creation of the Numbers syllabary is described in detail in Chapter 5. These experiments also used the syllable-length interval, which is defined for the experiments in this thesis as a roughly 200-ms span of speech, irrespective of actual speech content. Chapter 5 contains more detailed descriptions of the application of this interval in connection with the modulation spectrogram feature analysis method and the neural network context window.

2.2.2 Number of Syllables

The lack of a definition for the syllable renders problematic any attempt to define a set of unique syllables in the English language. Related to the issue of defining the essence of a syllable is defining the boundary for syllabification. The definition of a syllable boundary is as obscure as the definition of the syllable itself [192]. Any derived list of syllables would then be subject to the same caveats, which could account for the disparity in the statistics cited below for the number of unique syllables in English. Researchers mostly take a linguistically-oriented approach. Nevertheless, the estimates are roughly within the same order of magnitude.

Aside from the various estimates and definitions, it is clear that there are many unique syllables used in human language. One estimate implies that spoken American English requires 10,000 syllables for complete coverage [167]. From data intended for speech synthesis, O'Shaughnessy [148] derives 4,400 as the number of syllables sufficient to describe virtually all American English words and mentions that the most frequent 1,370 syllables are used 93% of the time. He also notes that complete coverage of American English would require perhaps 20,000 syllables.

The number of syllables is much smaller than the number of words, but much larger than the number of phonemes. The Oxford English Dictionary has 616,500 words, including variants, combinations and obsolete words. O'Shaughnessy mentions that modern American English has over 300,000 words, though only 50,000 can be considered to be commonly used. Clearly, syllables numerically represent a large compression over whole word units. On the other hand the number of unique syllables is considerably more than the 40-80 phonemes

[6] The Numbers corpus is defined in Section 3.2.

typically assigned to American English speech. Representationally, using phonemes requires many fewer unique symbols. Thus, in the experiments in Chapters 5 and 6, the number of syllable units is much greater than the number of context-independent phones used.

Although this is not an issue for restricted, small vocabulary tasks (such as the Numbers task used for the experiments described later in this thesis), there are concerns about the scalability of syllables to large vocabulary, conversational speech tasks. Historically, the number of unique syllables, and the possible complexity of syllable structures have been cited as arguments against using the syllable as a basic unit of automatic speech recognition. This particular argument, however, has somewhat less weight today in view of the widespread use of polyphone units in speech synthesis and speech recognition. Large vocabulary automatic speech recognition systems primarily use triphones (a phone with a unique pair of left and right adjoining phones), which are approximately as numerous as syllables. Some speech recognition systems are using quadphones and quinphones as well, with a commensurate increase in the number of units. Triphone-based systems typically have several thousand models, for example Cambridge University's HTK system for Wall Street Journal [205] and Dragon's system for Switchboard [151]. Young says a large vocabulary cross-word triphone system will typically require about 60,000 triphones [211]. Thus, the number of unique triphones is of the same order of magnitude as the number of different syllables. The exact number in a system depends on the implementation. Researchers can use techniques similar to those for reducing the number of triphones used in speech recognition systems for streamlining syllabaries, the collection of syllables used in a recognition task.

The number of different recognition units is a concern for researchers, whether the units are triphones or syllables, because of the need for an adequate quantity of training data for each unit. The next section examines the number and kind of syllables needed for recognizing large vocabulary conversational speech.

2.2.3 Syllables in Conversational Speech

Statistics gathered on words used in conversations can help characterize the usage of syllables in human speech [75] and more clearly outline the scalability issues. The study reported in this section used the Switchboard corpus [186, 70] word transcriptions, taken as representative samples of naturally spoken speech. A good deal of valuable information about conversational speech can be obtained through the careful examination of a large corpus such as this, as in the findings of the Switchboard Transcription Project [76].

The Switchboard corpus is a large database of spontaneous telephone conversations between two people, unfamiliar to each other, on a variety of topics (such as summer vacations, professional dress codes, the international political situation, credit cards, etc.). Collected at Texas Instruments specifically for the purpose of furthering speech recognition research, the corpus includes about 2,430 conversations comprising 140 hours of speech. Court reporters word-transcribed these conversations, which comprise about 2 million words of text, spoken by over 500 speakers of both sexes and from every major dialect of American English. The word transcriptions include a small number of word errors and also contain a variety of transcription notations. Since only a small portion of the Switchboard corpus has been phonetically hand-transcribed, the word transcriptions formed the basis of the syllable

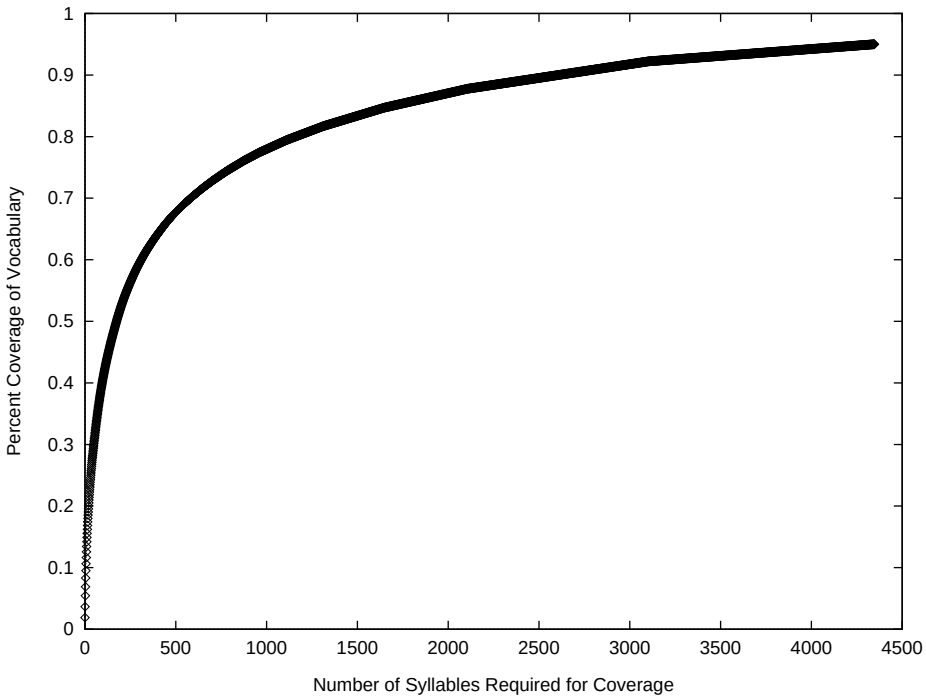

Figure 2.1: Graph illustrating the minimum number of syllables required for up to 95% coverage of the words in the Switchboard *vocabulary.*

statistics reported below.

The words in the Switchboard corpus transcriptions were syllabified by comparing words found in the transcriptions orthographically to the Celex database of English words [6], a dictionary of about 100,000 different English wordforms with pronunciations and a wealth of other information. The canonical syllables and their definitions were taken from the Celex collection, which is based on British English pronunciations. Since the analysis method used dictionary pronunciations, it did not account for deviations from the single, canonical English pronunciation for each word. The method was somewhat crude; it failed to syllabify any word not in the Celex database such as proper names, unusual words and words represented irregularly (e.g., misspellings in the transcriptions or unusually annotated words). Of the word tokens that occurred in the corpus, 4% could not be classified with these automatic methods. These represented 26% of the unique word orthographies. Although the proper nouns are an important exception, statistics gathered from the data that was successfully syllabified can still indicate general trends in human speech patterns.

This study distinguished the frequency of occurrence of syllables in the Switchboard *vocabulary* from the frequency of syllables in the full *corpus*. The vocabulary was the list of words that occur in the database, one instance per unique word item. This list was augmented by the frequency of occurrence of each of the words for the purpose of gathering counts over the full corpus. The conversational speech of the Switchboard corpus used about 26,000 different words (or baseforms composing the Switchboard *vocabulary*, also referred to as the Switchboard lexicon) and comprised a grand total of about two million

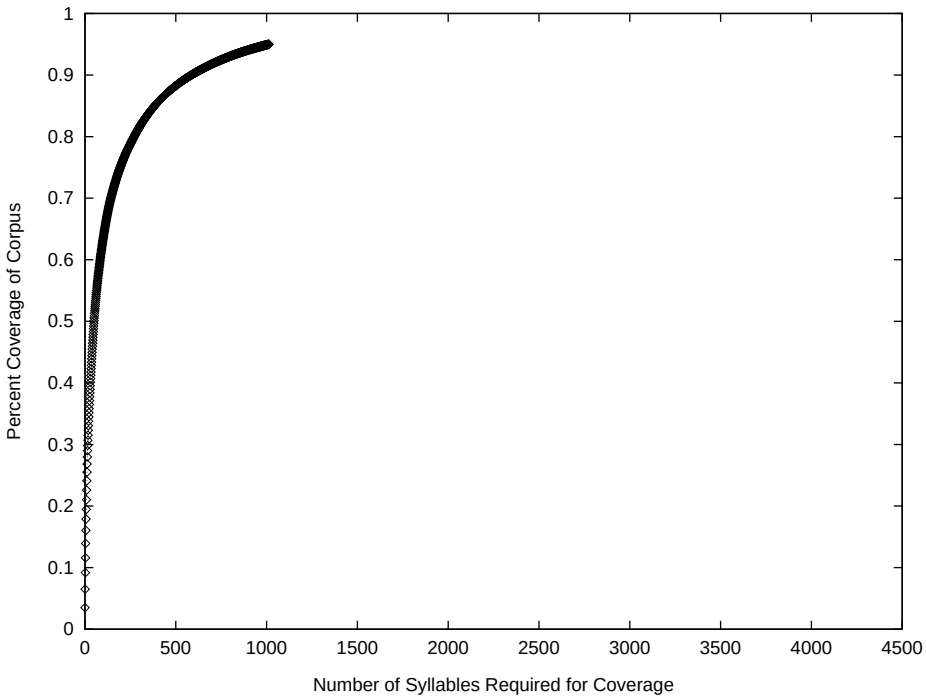

Figure 2.2: Graph illustrating the minimum number of syllables required for up to 95% coverage of the word tokens occurring in the Switchboard *corpus.*

word tokens in the word transcriptions (the Switchboard *corpus*). Complete representation of the vocabulary required about 6,000 different syllables. The study showed that 95% coverage of the words in the *vocabulary* required a minimum of about 5,000 different syllables as illustrated in Figure 2.1. For the *corpus*, however (i.e., weighting each word by its frequency of occurrence), Figure 2.2 shows that 95% coverage of the total word tokens required a minimum of about 2,000 different syllables. In addition, 75% of the corpus could be expressed with only 250 different syllables. This implies that speakers in Switchboard predominantly used a comparatively small set of syllables. In conversations, people speak in a fairly simple manner, relative to the potential complexity of English.

These methods also allowed the canonical, dictionary-equivalent versions of the syllables used to be characterized. Table 2.1 shows the frequency of n-syllable words in the Switchboard vocabulary and in the corpus word transcriptions. For every value of n, the numbers given for the complete Switchboard corpus (the right-most column in the table) are within one percent of the analogous numbers reported by French, Carter and Koenig [56] for telephone conversations, data collected over 60 years earlier. This agreement between the two is remarkable, especially across the shifts in language usage and the differences in definitions and methods.

Table 2.1 shows that while single syllable words comprised only a quarter of the vocabulary words in the conversations, these words occurred 81% of the time. French *et al.* reported the same effect. These numbers also approximately match the figures reported by Waibel [196] for a 20,000-word corpus. Words longer than two syllables made up only

N	percentage of vocabulary	percentage of corpus
1	22.39%	81.04%
2	39.76%	14.30%
3	24.26%	3.50%
4	9.91%	0.96%
5	3.21%	0.18%
6	0.40%	0.021%
7	0.057%	0.0013%
8	0.0052%	0.000037%

Table 2.1: Frequency of words with N syllables in the Switchboard vocabulary and corpus.

structure type	percentage of corpus
CVV	21.19%
CVC	19.75%
CVVC	9.99%
CV	9.51%
VC	9.14%
VV	6.98%
CVCC	3.99%
VCC	3.85%

Table 2.2: Frequency of the eight most frequent syllable (consonant-vowel) structures in the Switchboard corpus.

about 4% of the total words used in these conversations. Nouns were more likely to be polysyllabic than other grammar classes in both the Switchboard study and in the study by French *et al.* It can be concluded from these data that conversational English is in actuality more simply constructed than is commonly believed. Syllable boundaries were very often also word boundaries, therefore if syllable boundaries can be accurately detected then this gives an approach to word segmentation of the acoustic signal.

Another common belief is that English conversation requires the mastery of a great variety of syllable structure types (i.e., the pattern of distinct consonant, "C," and vowel, "V," constituents). "Scrounged," for example, is described as a single CCCVVCCC syllable in the Celex database, the longest single syllable. Of the 42 different syllable structures that occur in Switchboard, the eight relatively simple structures in Table 2.2 account for 84% of the syllables used in the corpus. Although the study by French *et al.* found somewhat different percentages, they also found that a handful of rather simple syllable structure types were used over 80% of the time.[7]

[7]It is not surprising that the French *et al.* study arrived at different percentages for this case since they used a different phone set and differing phonological conventions from those of the Celex data. For example,

This thesis is concerned only with American English; however, others are studying the use of syllables in conversational speech in other languages. Kirchhoff found similar observations for German with respect to the number of syllables in each word in conversations and the types of syllable structures used [110]. For Japanese, Arai and Greenberg found that syllables have more temporal characteristics in common with English than is popularly believed; they also reported that distributions of syllable structures in Japanese bear considerable similarity to those in English [1]. Greenberg, Hollenback and Ellis gathered additional phonological details, statistics and studies of syllable durations in conversational speech as represented by Switchboard [79].

2.2.4 Summary

Although American English has a large number of unique syllables and considerable potential for convoluted construction and complex syllabic structures, everyday speech is fairly simple. Conversational speech exhibits regularities in structure that can possibly be exploited for speech recognition. Most conversational speech can be expressed in a relatively small number of syllables, compared to the total number of syllables in American English, and these syllables tend to have clear, easily defined structures. These observations support the proposition that syllables could be used to improve accuracy in automatic speech recognition, even for large vocabulary tasks.

2.3 Syllables in Automatic Speech Recognition

Automatic speech recognition (ASR) systems typically employ phoneme- or sub-phoneme-based hidden Markov models (HMMs) concatenated into words and sentences. Although phoneme-based models are most popular, researchers in ASR have used a wide spectrum of units with varying levels of success, ranging from multi-word phrases to articulatory features and including such units as multi-syllabic groups (stressed-unstressed pairs) and syllable parts (onsets, nuclei and codas). The use of syllables and syllable-based long-time-span units in speech recognition offers many benefits, but there are also disadvantages and difficulties in implementation. The summary below describes the history of research using syllables and syllable-like units in ASR for English and includes discussion of similar work for other languages as relevant to this thesis.

2.3.1 Speech Units in ASR

A "phoneme" is an abstract conceptualization that is defined to encompass those "distinctions or contrasts that are recognized by speakers of the language as 'making different words' and acknowledged by linguists as systematically functional" [28]. For example, the distinction in linguistic meaning between the word "cat" and the word "pat" comes from

the French *et al.* study defined syllables to have at most one vowel, while the Celex database contained syllables with two vowels in the nucleus of a single syllable. The French *et al.* study probably represented several different types of Celex syllable structures as having the same syllable structure.

the difference between the /k/ and /p/ phonemes. Its function in distinguishing between different words is one of the many reasons why phonemes have been a convenient basis for the lowest level of speech decoding in recognition systems.

The manifestations of phonemes can vary considerably with context. Different sounds (phones) that are actually alternative acoustic representations of a single phoneme are grouped together and referred to as the "allophones," or conditioned variants, of a particular phoneme. Allophones are said to result from phonological conditioning, that is, language-specific rules of pronunciation. The general effect is called "allophonic variation." In continuous speech, the formation of a series of phones is achieved by the movement of the speech articulators.[8] Mechanical and neural limits on articulator motion can cause a spread of influence between neighboring phones. Known as "coarticulation," this results in variations from canonical phone expressions for the phonemic constituents of words. Allophones and coarticulation present challenges to machines attempting to classify speech sounds, which is addressed by the use of stochastic methods such as HMMs. Phone identification and segmentation, however, can be hard tasks even for experienced phoneticians. For example, transcribers can disagree when classifying the stationary segments of vowels in the manual phonetic labeling of speech data [77]. Although such experts can classify consonants fairly consistently, they will often disagree about vowel identity. Inter-labeler agreement for the Switchboard Transcription project was about 75-80% [78]. Through the inclusion of additional, relevant context, units larger than the phone may efficiently account for phonetic variations within a larger representational structure.

Early speech recognition systems took the approach of modeling whole word units and there is some evidence that this approach still performs best; every word and the acoustic-phonetic dynamics contained within can be carefully characterized. The interaction between phonemes can be extensively modeled within the context of the word. In modern large vocabulary tasks (with lexicons of 20,000 to 100,000 words), however, using whole-word units becomes difficult to implement and impractical. There is often only a small amount of training data for infrequent words, which also tend to be the longer words. Further, it is difficult to re-target such a recognizer for a new vocabulary; it requires complete retraining and re-tuning. Whole phrases (groups of words) have been modeled as well, but this strategy quickly becomes even more impractical on a large scale than words. In contrast, units smaller than words can be recombined to form new words, previously out-of-vocabulary, without retraining. While sub-word units are not a complete answer to the problem of insufficient training examples for some words, they have some advantages over whole-word models.

The syllable may be a useful compromise between the phoneme and the word. Alternatively, the syllable can be used in combination with words and phonemes to compensate for some of their disadvantages. Greenberg suggests that coarticulation and other non-linear effects of concatenating speech sounds are largely confined within a syllable [74]. Fujimura points out that coarticulation effects across syllable boundaries can be more easily defined those across phone boundaries within a syllable [59]. Moreover, as speakers omit phones and otherwise vary the pronunciation of words, the temporal characteristics of the original syllable structure are preserved in many cases, even though the phoneme sequence may

[8] Articulators include the vocal folds, soft palate or velum, tongue, teeth, lips, uvula and jaw.

have been substantially altered. Compared to whole words, sub-word units like syllables may present computational and storage advantages, particularly for large vocabulary tasks.

Another intermediate speech unit that has enjoyed some popularity is the "demisyllable," a term introduced by Fujimura in 1976 [60, 61]. A demisyllable is defined as essentially half of a syllable that has been divided after the CV transition. The exact point of division, and additional specifications, depend on a specific definition and implementation. It is estimated that there are 2,000 to 3,000 unique demisyllables in American English [97, 167]. Demisyllables have a numerical advantage over syllables, since many syllables can share the same demisyllables, but properties that belong to the syllable as a whole may not be expressed as coherently. The work described in Chapters 5 and 6 use "half-syllables," related in spirit to the demisyllable.

Researchers have encapsulated contextual information by using subcategorizations of the phoneme such as the context-dependent phone, which is a phone with some left and/or right phonetic context included. For example, instead of a context-independent phone /ae/, a context-dependent phone might be /(b)-ae-(t)/. The use of triphones, quadphones or quinphones is currently popular. It can be argued that context-dependent phone models can provide the same representational power as syllable-based units, particularly demisyllables. Syllable-level units, however, incorporate knowledge that can reduce the number of possible consonant-vowel combinations needed; as noted previously a large vocabulary cross-word triphone system might have as many as 60,000 triphones. A syllable-level unit may also be more suitable for modeling coarticulation effects that extend beyond the typical phone. Supposing human speech is organized around the syllable, the phonotactics[9] of context-dependent phone models may not reflect the underlying structure as well as syllable-based models. The "half-syllable" unit defined for experiments in this thesis encompasses a larger contiguous section of speech than a context-dependent phone and therefore can potentially incorporate properties from longer-time spans.

The next two sections address the question of the advantages and disadvantages of using syllables in ASR.

2.3.2 Syllables – Key for ASR?

For the concatenative methodology of current automatic speech recognition, the ideal unit is large enough to incorporate the majority of phonological effects, yet have relatively stable, well-defined boundaries. If the syllable is indeed a basic perceptual unit for humans, then using this information in ASR can perhaps improve recognition accuracy. Fujimura proposed the syllable (in the form of a specific definition) as a unit of automatic speech recognition in 1975 [59]. In his paper, Fujimura described the advantages of syllables for ASR, which appear to still be relevant for current ASR tasks. Fujimura has more recently proposed theories of prosodic structure interpreted as a series of syllables and boundaries with attached magnitude values (the C/D model) [62, 63] and has had success using a syllable-based unit for speech synthesis. The usefulness of the syllable in speech synthesis systems implies that syllabic units sound more like natural speech to the human ear and are

[9] Phonotactics refers to the way sounds combine with other sounds in a language. For example, the combination "nglib" can not be a syllable, according to the phonotactic rules of English [46].

easier to understand than other alternative units under the same concatenative paradigm. This suggests that the syllable possesses qualities that are conducive to perception and should be explored for automatic speech recognition. The strength of the syllable lies in the potential for greater accuracy through more accurate modeling of speech and for more engineering-oriented gains such as in execution speed and memory usage.

Recognition accuracy may be enhanced by the explicit representation of syllable effects. As discussed earlier in this chapter, the stability of the syllabic unit as a whole appears to be greater than that of the constituent phones [78]. Further, the coarticulation effects between phones within the same syllable are believed to be governed by rules that are distinct from those acting between phones of different syllables [59]. The syllable, then, may be a good encapsulation device. Simple phonological rules using syllables may easily represent long term structure.

Church makes an argument for the use of syllabic structure and stress as an intermediate level of representation between the phonetic description and the lexicon in machine parsing of phone sequences [26]. While allophonic variation has usually been thought of as problematic for recognition, in Church's proposal it is seen as a source of cues to improve recognition performance. Church points out that, theoretically, information regarding the location of syllables and stress can be derived from the distribution of allophones. Church gives the example of the aspiration of a voiceless stop in a stressed, word-initial position, an example earlier noted by Fujimura [59], Kenstowicz and Kisseberth [103] and Kahn [101]. Aspiration is an allophonic variation on the pronunciation of a phoneme pertaining to the amount of air released when the phone is produced. Because of its syllable-initial position, the /p/ in "pie" is the aspirated version of that phoneme; a larger puff of air is released than in the nonaspirated version that occurs in the word "spy" [46]. Another example is that /t/ is aspirated if it is at the beginning of a syllable, as in "ten." A preceding /s/, however, as in "stem" displaces the voiceless stop from the beginning of the syllable, so the /t/ is not aspirated. Identifying instances of those variations that occur only in syllable initial and syllable final structures can help segment sequences of phones into syllables.

Recognition systems can use this information to constrain the search for the correct lexical match. Church notes that most speech models rely on invariant features and therefore have no facility for benefiting from the information contained in allophonic variation and phonotactics. An intermediate level of representation, such as syllables, may present a way of utilizing such parsing clues. Syllable boundaries can be useful in providing hints as to where speech segments lie. Isolated word recognition has historically been more accurate than continuous speech recognition, and part of this is due to the greater ease in determining speech-silence segmentations than the more general coarticulated sound unit segmentations for continuous speech. In later case, the acoustic signal shows no clear boundaries between words. Syllable boundaries do have some acoustic manifestation, through allophonic variation, and may provide valuable, though approximate, pointers to the location of word boundaries.

Syllabic level properties of speech, such as energy contours and peaks [74], and fundamental frequency [114] are not often used in speech recognition, but may contain valuable information. Syllables can be used to incorporate prosody (the rhythmic and tonal qualities of speech) and other suprasegmental features. Suprasegmental features, which generally

span several phone segments, include stress, duration, tone and intonation [113]. Each of these can provide clues that may improve decoding accuracy.

Speech recognition systems do not commonly focus on long-term speech structure (i.e., over about 200 ms) even though considerable evidence has accumulated that indicates such a framework exists. Using the syllable and syllable-based units can facilitate the learning of long-term structure by statistical mechanisms. Long-time span analysis of speech can enhance speech recognition accuracy in addition to the role that the syllable plays as a basic perceptual unit of human speech recognition.

Using syllables may also affect the implementation of speech recognition systems. Syllables may enable systems to reduce the amount of memory used or reduce the execution time needed without sacrificing accuracy rates. The search space of syllables may have useful properties for algorithms; in comparison to words, different syllables relate to each other in a fairly well-understood and constrained manner. Thus, the syllable search space is more easily defined and possibly has reduced complexity through the reduction of redundant computation.

Many current decoding strategies use some sort of clustering, or tree-structuring of pronunciation models to reduce redundant computation. For example, the beginning portions of words often share common phone sequences. Lexicons are often represented by trees so that processing on these initial portions is not repeated. Also, word lattice generation, described in more detail in Section 3.4, has become more popular in recent years, spurring the need to produce lattices of moderate size. The syllable is a compelling size and a natural unit for the representation of lexicons that efficiently unites the common portions of words, perhaps in combination with a lexical tree. For the same reason, the syllable can also be a more efficient representation for lattices; many words can be represented with the same set of syllables.

The use of the syllable, as opposed to the word, as an organizational unit may also allow more efficient utilization of parallel and concurrent machines. These architectures offer massive computational and storage resources [3] that have been largely untapped by the speech decoding problem. The search space and representation of syllables (trees and directed graphs, for example) may be more easily structured for parallel or concurrent machines than sequential word models. Similar arguments apply to vector processing methodologies.

2.3.3 Syllables – Morass of confusion?

Along with the many advantages to using syllables in speech recognition, there are also factors that confound the incorporation of syllables into ASR. There is a tradeoff and a balance to be found between benefits from the use of the syllable and added complications and complexity. This consideration underlies the hypothesis of this thesis. Do the advantages outweigh the disadvantages for a given speech recognition methodology? Although the issues involved are complex and contentious to some degree, simplifying engineering assumptions can allow the successful use of syllable information as will be illustrated by the positive experimental results described in Chapters 4, 5 and 6.

The primary problem with using syllables in ASR lies in the lack of a definition for the

syllable and its boundaries. In particular, syllables (and therefore demisyllables and other syllable-based units) are not always clearly defined in American English.[10] The most often cited causes are syllable reduction, stress-timing, and ambisyllabicity. There is also ongoing argument as to whether a syllable's boundary is properly located within the intervocalic segment (between the consonants of a syllable) or in the consonant clusters.

Syllable reduction occurs when polysyllabic words are simplified. For instance, VCV or CVCV may be converted to CV words by the elimination of the initial syllable, or the vowel in it. Examples given by Ohde and Sharf include words like "away" and "believe," which become "way" and "blieve" [147]. The vowel in the first syllable, which also is the syllable with weaker stress, is eliminated. The reduction of unstressed syllables is a characteristic of the normal rhythm of English [28].

Stressed syllables tend to be longer and more intense. Linguists often describe American English as a "stress-timed" language [28]. Traditionally, languages characterized as "stressed-timed" have unstressed syllables that are greatly reduced in duration compared to stressed syllables; in these languages stressed syllables tend to occur with an even tempo. This contrasts with the term "syllable-timed," which has traditionally referred to languages in which every syllable has approximately equal duration. Japanese is usually given as the classic example of a syllable-timed language. Arai and Greenberg, however, have observed that for conversational Japanese, the durations of syllables varied almost as much as for English [1, 78].

Linguists believe that stress timing causes the unstressed syllables between stressed syllables to have varying, unequal durations [28]. Lately, Arai and Greenberg have found that for conversational speech, stress-timing in English may manifest as a large range of syllable durations rather than as a strictly alternating pattern [76]. Statistically, English tends to have a slightly wider range of syllable durations than Japanese. The effects of stress-timing could make syllabification in ordinary speech difficult, because the duration of unstressed syllables in American English shrinks dynamically compared to the stressed syllables. Fujimura, in [63], discusses the problem of finding "phonetically hidden" syllables and suggests heuristics for divining their locations.

Syllable boundaries are difficult to determine in many cases due to ambiguous structure, which is commonplace in American English words. Linguists disagree on how many syllables compose words such as "meal," "seal," "real," and so on [113]. One form of ambiguous structure, *ambisyllabicity*, where one segment belongs to two syllables, also makes syllabic segmentation difficult in American English. Words such as "nesting" have unclear syllable boundaries. With the pronunciation /n-eh-s-t-iy-ng/ (in ICSI56 orthography), the word can be produced as /n-eh-s-t/ /iy-ng/, /n-eh-s/ /t-iy-ng/, /n-eh/ /s-t-iy-ng/, or with an ambisyllabic /t/, as in /n-eh-s-t/ /t-iy-ng/. Moreover, the semantic meaning of a string of phones can affect the perception of the syllabification. One example of such an effect is the naturally spoken phrase /h-ih-d-n-ey-m-z/. If the listener thinks the words are "hidden aims," it translates to 3 syllables. If the listener instead thinks the phrase is "hid names," only 2 syllables are perceived [113].

[10] Syllables are not alone in suffering from a lack of definition. For instance, phone identities and boundaries can be equally or even more difficult to distinguish.

As an added complexity, a phone sequence can be syllabified differently depending on the speaker's condition, for example, as when the speech is particularly fast or slow. One problem that occurs in the experiments discussed in Chapter 4 is that connected speech syllabifies very differently from the same words spoken in isolation. Syllable boundaries can move across word boundaries. For example, the words "five eight" can be pronounced by faster speakers as /f ay/ /v ey t/ where the /v/-release resyllabifies to the "eight." In this case the syllabic onset can be associated with either the /v/-release or the /ey/. These ambiguities make it difficult to resolve the syllable boundary in an automatic fashion.

As examined previously, the definition of the syllable is amorphous and ill-defined. For implementation purposes, however, the engineer must have a concrete and clear specification of each model. The engineer must make arbitrary decisions about how to characterize a syllable and these decisions may not always be defensible from every linguistic point of view. The resulting "syllable" definitions deviate from what is acceptable to the average phonetician.

Syllabification is an open research topic. Several methods and techniques, each emphasizing a different aspect of the problem, have been developed in several different contexts. Automatic syllable parsers are available for making syllabic segmentations of specified phonetic sequences. Hammond uses Optimality Theory to explain syllable parsing in English and French [82]. A parser from Fisher [51], used for the experiments in this thesis, is based on an implementation of the hierarchical rules presented by Kahn [101]. Other efforts include [39], in which the authors use a mainstay of neural network training, error-back-propagation, to learn the syllabification of Dutch. All automatic syllabification methods have some shortcomings, yet an ASR system based on syllables is highly dependent on their results.

Despite these concerns, automatic speech recognition systems can use syllable-based information to improve recognition accuracy. ASR systems have achieved considerable success despite the difficultly of ideal *phonetic* identification and segmentation. This success hints that perfect *syllable* identification and boundaries may not be necessary. Most words are fairly straightforward to syllabify [113]. For other words, a clearly defined, consistently applied process can produce usable syllabifications that correctly assign most of the salient properties of each syllable. Although linguists might argue with the theoretical validity of such syllables, the definition is effective for engineering applications, as will be seen experimentally in Chapters 4, 5 and 6. The specification captured enough of the features of syllables to permit positive effects. Addressing the problems described in this section should further improve and expand recognition performance.

2.3.4 Other Work with Syllable-like Units in ASR

Despite these difficulties, the possible benefits of syllables have periodically motivated researchers to experiment with them over the last three decades. Although most speech recognition research has focused on the established phoneme-based paradigm, the syllable and other long-time-span units have appeared from time to time in the literature, an early example being the Hearsay system where the syllable was one of the levels of representation [117, 48]. In this section, a sampling of the areas pursued in syllable-based ASR research is

surveyed.

The conceptual ancestors of certain aspects of the approach described in Chapter 4 are the works by Hunt, Lennig and Mermelstein in the late 1970's and the SYLK project by Green, Kew and Miller in the early 1990's. Hunt *et al.* performed a pilot experiment in which they incorporated syllables into the recognition of a small vocabulary American English task by first attempting to segment the input speech signal into syllabic intervals using what the authors called the "loudness" contour of the waveform [92, 91]. This syllable-based system attempted to estimate syllable boundaries, then formed recognized syllable sequences into words and sentences. In this system, Mermelstein's automatic segmentation system assessed syllable boundaries from a loudness function computed over the entire power spectrum [129].[11] They concluded that this approach showed some promise and was worthy of further research. Waibel [196] also investigated the reliable estimation of syllable boundaries. Waibel's algorithm, which defined a syllable's onset to be the beginning of the vowel nucleus, performed comparably to similar algorithms, including the one by Mermelstein. His algorithm also ensured that boundaries identified by his process would be commensurate with abstract linguistic considerations. Waibel mentions that all the algorithms then known fall short of the ability of humans to syllabify speech, a conclusion that is still true today. In the SYLK project [73, 72], researchers chose the syllable as the "explanation unit" to address the issue of allophonic variation. Green, Kew and Miller focused their work on locating syllable onsets (defined from a phonological point of view) where their symbol methodology contained 20 distinct onsets.

Segmenting continuous speech by focusing on syllabic nuclei was discussed in the early 1980s by De Mori and Giordano [42]. For German, which bears a close relationship to English, Reichl and Ruske [162] approached the identification of syllable nuclei through neural networks. Ruske, Plannerer and Shultz [169, 154] have experimented with demisyllable-based speech recognition systems for German. In [154, 155], systems first segmented the speech signal into syllables and then used parts of syllables for the recognition process. Recently, Schiel (then at ICSI) modeled German syllables from the Verbmobil project with multi-state hidden Markov models [174]. He found encouraging success when he used phoneme-based HMMs in tandem with syllable-based HMMs for commonly occurring words and allowed the decoder to pick which of the two to use for the final hypothesis.

In 1986, Gauvain reported experiments with syllable-based recognition of isolated words in French [68]. While Gauvain found that the change increased the overall word error rate, he also found that syllable representations of words reduced the storage required by his system to one-sixth of that required for whole words. Gauvain's further analysis revealed that the syllable-based system and the whole-word based system made substantially different errors.

For American English, Rosenberg, Rabiner, Wilpon and Kahn [167] experimented with demisyllables. Although the approach used in this work is similar in spirit to the work reported in Chapter 5, it differs substantially in implementation: the Rosenberg system focused on isolated words and used dynamic time warping to match the input acoustic features to templates. The thesis work to be described later focused on continuous speech

[11] For the experiments with syllable onsets described in Chapter 4, the onsets were estimated from 9 critical-band-like regions, which supplied spectral features for a neural network.

and used neural networks to classify the units and HMMs to form words and sentences.

Recently, Hu, Schalkwyk, Barnard and Cole used syllable-like units as the basic units of their recognition system [89]. In their segment-based system, the larger units proved to be less sensitive to segmentation accuracy. Hauenstein (then at ICSI) experimented with neural networks trained for whole syllable classification [83, 84]. While his syllable-based system underperformed a more conventional phone classifier based on word error rate, Hauenstein found that the syllable classifier-based system performed better for cross-database isolated word recognition tasks. This result suggests that the syllable-based system learned some characteristics of syllables that were more transferable to a new corpus than those of phones. In essence, the syllable system may have had better capabilities for robust generalization. Jones, Downey and Mason [99] reported positive results in recognizing syllable targets compared with monophone targets, though they did not report word recognition results.

In the spirit of dealing with allophonic and lexical variation, Kirchhoff addressed the issue of acoustic variability by using phonetic features for recognition, and allowing the features to overlap within the context of a syllable [108, 109]. De Mori and Galler implemented a method of using syllable phonotactics (rules governing how syllables combine with each other) to create new word pronunciations, thus generating lexical variations for a word automatically [41]. They chose the syllable as the unit for this process because phones can often be completely deleted in a lexical variant while some form of the relevant syllable is often still detected if the word is intelligible. These authors noted that high accuracy with large vocabulary ASR can be achieved only by using many different knowledge sources. The syllable model has attributes that can assist with the combination and integration of different sources of knowledge.

Researchers have used syllables in conjunction with prosody towards improving the accuracy of ASR. Prosody, as mentioned previously, refers to the tonal and rhythmic qualities of speech. Prosodic attributes, such as duration, amplitude and F0 contour (pitch), span well past the boundaries of phones. Also known as suprasegmentals, these properties are not commonly used in automatic speech recognition, though there is considerable evidence that such information helps humans recognize speech. Stress is a primary function of prosody.

Lea, Medress and Skinner [114] used prosodic features to break up sentences into phrases, locate stressed syllables, and classify the phonetic constituents all based on the principle that phonetic segments were more easily identified when contained in stressed syllables. Jones and Woodland [98] used the strength and stress of a syllable as additional constraints in a large vocabulary continuous speech recognizer to obtain a significant word error rate improvement.

Although not a focus of this thesis, keyword spotting can employ syllable-based representations towards improving garbage-modeling and reducing the implementation time for changing applications, particularly in large-vocabulary spotting tasks. "Garbage modeling" refers to the generic modeling of non-keywords and other sounds. This can be done in a variety of ways, including using averages of phone probability estimates [18] and explicit modeling of extraneous sounds with fully connected models. In a syllable-based word-spotter, the keywords are represented as concatenations of syllables and garbage models can be represented as different concatenations. This means that training the syllable mod-

els builds the keyword models and the garbage models at the same time. Altering the task and changing the list of keywords does not require the syllable models to be retrained, only the words are redefined. A syllable-based keyword spotting system is reported in [112]. Syllables or syllable-based features have also been used to improve garbage-modeling in word-spotting tasks in Spanish [121].

Some speech recognition researchers who work with languages other than American English have moved more quickly to embrace the use of syllables and parts of syllables, particularly in languages that are more clearly syllable-based and less stress-timed. Specific examples of speech recognizers with syllables or syllable parts include those for Chinese [119, 116], German [154, 162], Hungarian [195], Japanese [126], and Spanish [16]. These projects reported encouraging levels of success with methods that may be applicable to the recognition of English.

The 1997 Johns Hopkins ASR Summer Workshop [67] also explored the idea of using syllables in automatic speech recognition. Syllable models of varying complexity for the most frequent words were integrated with more conventional phone models for the remaining words. The project achieved only modest gains in accuracy, but the workshop participants concluded that the syllable approach showed promise.

2.3.5 A Few Words About Hyphenation

There is often the misconception that syllabification and hyphenation are very similar, when in fact the two operate on substantially different criteria. Hyphenation, the splitting of words for typographical efficiency, is governed by morpheme boundaries in a different way than syllables; the hyphenation of a word can differ considerably from the syllabification of the word. For example the word "booking" is hyphenated as "book-ing," respecting the morpheme boundary between the parts "book" and "-ing," but the word is usually syllabified as "boo-king" or "boo[k]ing," reflecting an ambisyllabic /k/. The algorithm for hyphenation initially developed by Liang [118] and used in TeX, usually produces good results and is widely accepted as state-of-the-art. Nevertheless, automatic hyphenation remains an active area of research. Although hyphenation is a separate avenue of research from syllabification, the research communities can interchange useful ideas and inspiration.

2.3.6 Summary

Because the syllable may have a primary role in human speech recognition, researchers have suggested the use of the syllable as a basic processing unit for automatic speech recognition for machines. The syllable may optimize trade-offs between the word-level modeling of longer-time span coarticulation and finer detail at the level of the phoneme. It may also correspond naturally to speech properties like stress, energy and pitch. The syllable unit may confer other benefits for ASR as an organizational unit; syllables may help reduce redundant computation and storage in speech decoding. Syllables have been used successfully in speech recognizers for other languages, and although there are obstacles to overcome in developing American English syllable-based speech recognition, some researchers have already reported positive and encouraging results using syllables in pilot experiments. By

using carefully applied engineering-motivated definitions of syllables, systems can capitalize on syllable-based information without becoming mired too deeply into unresolved problems.

2.4 Conclusions

Although some of the discussion has concentrated on advancing the syllable as a basic unit of speech recognition at the expense of the phoneme, the intent is not to suggest that phonemes are dispensable. The evidence of alphabetic writing systems, the existence of rhyme and alliteration in poetry, phonemic spoonerisms, and historical changes in language that can be described most easily using phonemic terms are testimony to the importance, at some level, of the phoneme [173]. One possibility for integrating phones and syllables in speech is to regard them both as expressed attributes or features of the syllable-length time interval to which they belong [78].[12]

The importance of the syllable in human speech perception is still vigorously contested. Some researchers believe that the phoneme is sufficient to describe the human speech perception process. No matter what resolution eventually prevails, the fact that a debate has raged for so long indicates in itself that long-time-span units, such as syllables and syllable-sized units, have some kind of influence in human speech recognition. The above studies suggest that the syllable or a similar long-time-span component may be a basic unit of speech perception and that the syllabic-length interval may be a temporal unit for speech recognition. Consequently, researchers have investigated and continue to explore using the concept of the syllable in speech recognition by machines.

American English has considerable potential for convoluted construction and complex syllabic structures, but everyday, conversational speech is fairly simple. These patterns are amenable to current stochastic techniques for automatic speech recognition. The experiments in this thesis work were concerned with relatively simple syllable structure types and a small number of distinct syllables in the syllabary. Chapter 7 discusses possible extensions and the issues involved in incorporating more complex syllable types and using a larger syllabary.

This chapter discussed the motivation and background behind the focus on syllable-based information in this thesis. While there are persuasive arguments for the explicit incorporation of syllable-based elements into speech recognition, there are many potential problems that could confound the effective use of such information. Whether the advantages outweigh the disadvantages can be explored most directly through experiments. The work described in Chapters 4, 5 and 6 provides some empirical support for the advantages of using syllable-based information.

[12] This supposition is discussed further in the syllable-level combining experiment described in Section 6.4.

Chapter 3

Automatic Speech Recognition

Speech recognition has arrived in the commercial, publicly accessible marketplace. In the past decade researchers have made great advances; there are a number of popular ASR-based products. The ultimate goal of *robust*, continuous, large-vocabulary speech recognition, usable by the general public, however, is still a number of years away. There are many unresolved problems and unanswered questions.[1]

In this thesis the syllable is used as a facilitator for understanding the problem of speech recognition and an avenue to viable approaches for answering some of these questions. A discussion of the strengths and weaknesses of the state of the art will establish the context for the possible contributions of syllable-based methods to the advancement of speech recognition. The field of speech recognition is very broad, however, so the overview in the first section of this chapter will be comparatively brief and cover only details relevant to the report of experiments in this thesis. As background for the work to be discussed in Chapters 4, 5 and 6, Sections 3.2 and 3.3 in this chapter give a detailed summary of the Numbers task and ICSI's speech recognition system, which serves as the platform for all the experiments discussed in this thesis. One set of experiments focuses on an approach to constraining hypothesis creation in speech decoding, so Section 3.4 in this chapter gives a summary of current thinking in speech decoding. Another set of experiments concentrates on several methods of combining syllable-level information with a baseline, phoneme-oriented speech recognition system. Section 3.5 gives a brief discussion about various combination methods.

3.1 The State of the Art

At the time of writing, speech recognition systems in the marketplace are just beginning to be usable by general users and to gain mass acceptance.[2] Naive users employ limited-task systems with success. For example, the AT&T Universal Card customer service system

[1] It has been observed that true speech recognition has been estimated as "5-10 years away" since the 1950s.

[2] The list of commercial products in this chapter is by no means exhaustive. New products and services are being introduced continuously. A recent survey can be found in [158].

accepts spoken continuous digits (credit card numbers) over the telephone. Less severely limited tasks require the user to have some training in the use of the system and a fairly clean acoustic environment. Wildfire, a telecommunications assistant service, purports to allow a user to speak naturally using a limited vocabulary and a constrained range of constructs to instruct the system in dealing with phone calls and phone messages. In these command-and-control systems the user must conform to the format of the system's interface. Older dictation applications (e.g., Dragon Dictate) usually required the user to pause between each word and to train the system– as well as the speaker– for maximum performance. Both Dragon and its competitor IBM recently released products for the recognition of *continuous* speech. Speakers still need to wear close-talking microphones and to train with the system individually. As the complexity of speech recognition applications increases, more sophistication and training on the part of the user is required.

In spite of large gaps in the understanding of human speech perception and technological obstacles, researchers in speech recognition technology have made particularly rapid advances in the last decade.[3] As the state of the art in speech recognition advances, applications for speech recognition rapidly increase in scope.

The published Defense Advanced Research Project Agency (DARPA) benchmarks for evaluating speech recognition performance have progressed considerably in size and difficulty[4] since 1971, when the evaluation task consisted of a 1000-word vocabulary task spoken by only a handful of different speakers [111]. Resource Management, a 1000-word vocabulary task used extensively in the 1980s, included read speech from hundreds of different speakers of many different U.S. dialects [156]. In 1993, the evaluation task was speech read by speakers using a 20,000-word vocabulary (North American Business News) [141]. At the same time, a 26,000-word vocabulary, spontaneous, human-to-human conversational speech corpus called Switchboard was developed [186, 70]. In 1997, Broadcast News, speech taken from television and radio news programs [71] with the attendant variety of interfering background conditions, became the latest DARPA Continuous Speech Recognition (CSR) evaluation challenge. Researchers have also been tackling other corpora with much larger vocabularies (64,000 – 100,000 words). They are also working towards improving the execution of these systems to near real-time performance.

Despite recent advances, unconstrained speech recognition usable by naive users in less than ideal acoustic environments is still very challenging. Cole *et al.* list one of the unattained goals of speech recognition as robustness at all levels, including robustness to background or channel noise, unfamiliar words, accents, differences in users and unantici-

[3] Comerford *et al.* attribute recent advances more to improvements in hardware price/performance, rather than to breakthroughs in speech research [32].

[4] One aspect of "difficulty" is quantified as perplexity, which is usually measured as

$$P = 2^{-\frac{1}{N}\sum log_2(p)},$$

where N is the total number of words in the test set, and p is the probability of the observed transition as calculated from the training data [94], which roughly corresponds to the average number of branches at any decision point in a process. Speech researchers generally approximate perplexity according to vocabulary size; one rule of thumb is that the difficulty of a recognition task increases with the logarithm of the size of the vocabulary. As noted in [32], however, if a large vocabulary task has few possible branches, accuracy can be rather good, but a small vocabulary task with many possible branches can be difficult.

SD Baseline	1.5%
SI Baseline	3.0%
Channel	12.0%
Transducer	10.0%
Speaking Rate	15.0%
Language Model	70.0%
Noise	30.0%
Dialect	20.0%
Non-Native Speaker	45.0%
Noise + Non-nativeness	85.0%
Combining All Effects	98.0%

Table 3.1: Word error rates showing abrupt degradation in recognition accuracy due to introduction of various effects [65].

pated input [31]. They define robust speech recognition as "...minimal, graceful degradation in performance due to changes in input conditions caused by different microphones, room acoustics, background or channel noise, different speakers, or other small (insofar as human listeners are concerned) systematic changes in the acoustic signal." Laboratory systems that perform well in constrained conditions show a tendency to experience sudden, relatively large decreases in accuracy. For instance, a laboratory system described by [65] achieved a 3% word-error rate for the Resource Management task in ideal conditions. After adding in a variety of acoustic variations common in realistic field conditions, such as channel differences, changes in speaking rate, changes in dialect, noise, accents from non-native speakers and a poor language model, the error rate increased to 98%, as shown in Table 3.1. Yet, human beings cope well under the same conditions, with little or no degradation in recognition. Many of the problems encountered in the field can be resolved using additional data collection, training and analysis [188]. The labor involved, however, is substantial and ideally such post-deployment effort should not be necessary.

The size of these applications and the increasingly intricate algorithms they require for robust performance present a complexity-management problem for the engineering of speech recognition software. Researchers would like to add additional sources of knowledge or do extra processing to address the challenges of open problems in speech recognition. Nevertheless, a practical speech recognition application must fit in available, affordable machines and be able to process utterances in near real-time for user comfort and acceptability.

Using the syllable as a tool of organization and understanding can help approach these fundamental issues for speech recognition for machines. In Chapters 4, 5 and 6 this thesis describes syllable-oriented attempts to address these issues through efforts to improve speech recognition accuracy and robustness for numbers spoken naturally over the telephone.

zero	oh	ten	uh
one	eleven	hundred	um
two	twelve	twenty	
three	thirteen	thirty	
four	fourteen	forty	
five	fifteen	fifty	
six	sixteen	sixty	
seven	seventeen	seventy	
eight	eighteen	eighty	
nine	nineteen	ninety	

Table 3.2: The list of vocabulary words in subset of Numbers used for experiments.

3.2 The Task: Numbers

For the speech recognition experiments discussed in this thesis, it was necessary to select a corpus that was neither too large (which would have introduced impractical development cycle times) nor too small (which might not be representative of actual, conversational speech). The Numbers corpus is sufficiently varied that a number of the effects of naturally spoken speech are in evidence. These include factors such as differences in speakers, variations in speaking rate, and reduced syllables. The samples also show effects from channel and environmental interference, for instance babies crying in the background. The Numbers task is fairly small yet non-trivial, so findings with this corpus are likely to be extensible to less constrained tasks.

Researchers at Oregon Graduate Institute (Center for Spoken Language Understanding, or "CSLU") collected the Numbers corpus as part of a larger assemblage of data for the purpose of providing challenging corpora for speech recognition research [30]. This corpus contains continuous, natural speech from many different people in response to prompts from an automated census system over telephone lines (digitized at 8 kHz). The Numbers utterances were cut from longer speech waveforms of people reciting their addresses, telephone numbers, zip codes or other miscellaneous items. OGI labelers phonetically hand-transcribed about half of the complete Numbers corpus.

A subset of the Numbers corpus was chosen for the experiments in this thesis.[5] The "core subset" contains only utterances with accompanying phonetic hand-transcriptions. The set is further limited to utterances in which the words at each end of the waveform are still intelligible (rather than being largely clipped), and which also contain only words that could strictly be called "numbers." These criteria eliminated utterances such as "Sears one day sale." Utterances where words on the boundary of the waveform was only partially recorded were also subtracted, for example when "seven" was represented by just "s-." The complete vocabulary of the core subset is 32 words, as listed in Table 3.2.

A sample utterance from the corpus is "eighteen thirty one." Since the utterances were

[5]The core subset was defined with Michael Shire (at ICSI).

excised from longer recordings, acoustic information from speech commenced immediately and broke off sharply. In order to allow for the start-up time in the recognition process, each wavefile was padded with 100 ms of artificially-created silence on both ends. The core subset contains about two hours of training data (3500 utterances, about 700,000 frames) and 40 minutes each of development test set (1,206 utterances, total of 4,673 words, about 230,000 frames) and evaluation test set (1,227 utterances, total of 4,757 words, about 230,000 frames) data. Any parameter tuning for the training and recognition systems in the experiments in this thesis involved only the training data, of which 10% is used as the cross-validation set.

3.2.1 Reverberation

Some of the experiments used artificially reverberated versions of the development and evaluation test sets as representative samples of one specific form of distortion. Reverberation manifests in sound propagating through a room due to the reflectivity of the walls and other solid objects. It also gives human listeners an impression of a room's size and general attributes. Human listeners prefer some reverberation when listening to music in concert halls. In statistical terms, reverberation is characterized as a transient, nonstationary, fairly slow response of sound in rooms.[6]

Reverberation can degrade speech intelligibility by masking direct sounds with reflected energy. When such environmental effects are not represented in the training data, they can increase the word error rates of speech recognition systems by an order of magnitude or more. Kingsbury *et al.* produced mildly reverberant speech for these experiments in connection with other research [80, 106, 107, 105]. The original speech from the Numbers database was digitally convolved with a real room impulse response using a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 db [107].[7] Artificially reverberated speech differs from actual reverberant speech in two significant ways: 1) Speakers compensate for perceived interference by modifying vocal effort. Since the reverberation was added after the speech was recorded, this effect is not reflected in the reverberant speech used here. 2) The impulse response used reflects a particular room model with a single source and microphone location. Recordings from actual rooms almost certainly will vary.

3.2.2 Human Recognition Performance

An informal speech understanding experiment with two human subjects (conducted in conjunction with Brian Kingsbury's thesis [105]) on 200 sentences of the Numbers development test set showed that humans can understand both the clean and the reverberant Numbers utterances with near perfect accuracy. The subjects had an average word error rate of 0.3% on both the clean and reverberant portions, in sharp contrast to the capabilities of automatic speech recognition systems.

[6] More details regarding the nature of room reverberation can be found in [132]. A statistical characterization of reverberation in rooms can be found in [191].

[7] The room impulse response used was part of a collection by Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, now at the University of California, Davis.

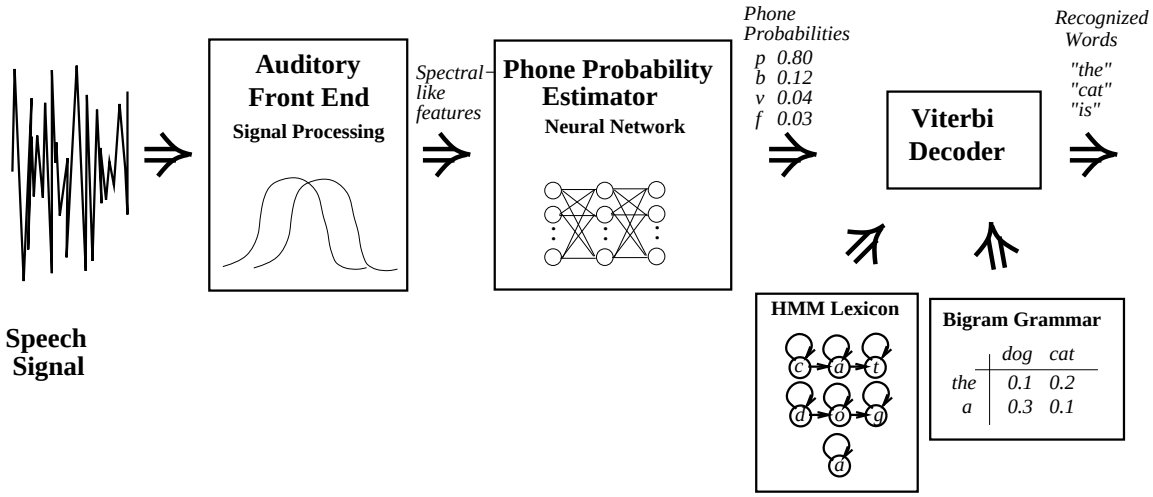

Figure 3.1: ICSI's speech recognition system. (Dan Jurafsky and Nikki Mirghafori)

3.3 The ICSI Speech Recognition System

A typical, current speech recognition system applies statistical pattern recognition techniques, first applied to this problem by Baker [11] and Jelinek, Bahl and Mercer [95]. The system first processes raw acoustic data into features by an analysis technique that characterizes the short-term spectral envelope (e.g., mel-frequency cepstral analysis [40],[8] or PLP analysis [85][9]).

A probability estimation technique, such as multivariate Gaussian mixtures [160, 90] or artificial neural networks [20], further processes these features. Standard 1990s ASR technology, as described in [17], usually refers to systems with hidden Markov models and multiple Gaussian mixtures. A classic example is the SPHINX system [115]. Young describes a fairly typical, current HMM-based system in [211]. The combination of HMMs with neural networks is commonly referred to as the "hybrid" approach. All statistical approaches typically involve extensive model training on large databases.

Figure 3.1 [100] is an illustration of the speech recognition system in primary use at the International Computer Science Institute (ICSI)[10] and is used as the baseline system for the experiments in this thesis. Although similar in form to the speech recognition systems in general use in the research community, the ICSI system has two less-common aspects: (1) acoustic probabilities are estimated by a neural network instead of by mixtures of Gaussians and (2) the system uses context-independent phones instead of triphones. These differences can be loosely characterized by noting that the ICSI system uses a comparatively

[8] Mel-frequency cepstral coefficients (MFCCs) are calculated by warping the speech signal spectrum to approximate the spatial-frequency scaling characteristic of human hearing. The process takes the logarithm of the warped spectrum and uses an inverse Fourier transform to generate features.

[9] PLP analysis estimates the auditory spectrum using several concepts from psychophysics and an autoregressive all-pole model. PLP is described in more detail later in this chapter.

[10] A much more complete description of the theory and mechanics underlying this type of recognition engine can be found in the book *Connectionist Speech Recognition– a Hybrid Approach* [20] or in [133].

large number of neural network parameters to estimate the density function for each of relatively few sound categories. The more common approach in the research community uses mixtures of Gaussians, estimators that use comparatively few parameters to estimate density functions for each of a relatively large number of sound categories. Systems that use mixtures of Gaussians usually have many times the number of sound categories as neural network-based systems. As a result, Gaussian-based systems typically have many more parameters in total.

3.3.1 Feature Extraction: RASTA-PLP

The ICSI system first analyzes sound waveforms with an acoustic processing technique, represented in Figure 3.1 as the "Auditory Front End," or the "ear" of the speech recognition system. The process segments the acoustic waveform input into overlapping "frames," which are 25 ms long with a 10-ms overlap for the experiments in Chapters 4, 5 and 6. The work reported in this thesis used RASTA-PLP features [86], roband features [184] and modulation spectrogram features [80, 105, 107]. Roband features are not intended for phonetic determination, unlike RASTA-PLP and modulation spectrogram features. Roband features are described in more detail in Section 4.1.2 and modulation spectrogram features are described in more detail in Section 5.1. All three features sets use principles from human speech perception to improve the representation of the speech signal.

RASTA-PLP[11] was derived from an older feature extraction method, Perceptual Linear Predictive (PLP) analysis [85]. PLP estimates the auditory spectrum using engineering approximations to the psychophysics of hearing. The process maps critical-band power spectra into a perceptually-based loudness domain. Features are generated using an autoregressive all-pole model. The results are converted into cepstral coefficients. Hermansky states that PLP is more consistent with some of the important properties of human hearing than conventional linear predictive (LP) analysis. Additional properties of human hearing are incorporated in RASTA-PLP.

The temporal characteristics of environmental noise or channel frequency response often differ from those of speech. This observation prompted Hermansky and Morgan to develop the RASTA-PLP speech processing method [135, 131, 86]. RASTA stands for relative spectra, a class of representations based on filter methods designed to exploit the differences between the temporal qualities of environmental noise or channel frequency response and those of speech.

The RASTA technique suppresses the components of the input spectral trajectories that change more slowly or more quickly than the statistically observed behavior of speech [86]. This has its foundation in human auditory perception, where researchers have observed that humans are sensitive to changes in an input in relative rather than absolute values. For instance, humans appear to be fairly insensitive to slowly changing background noise. Procedurally, Hermansky and Morgan altered the PLP speech analysis method: instead of the usual short-term critical-band spectrum in PLP speech analysis, RASTA-PLP has a spectral estimate where the temporal trajectory of each frequency channel is band-passed

[11] Also referred to as "log-RASTA-PLP" or simply "RASTA."

filtered with a sharp spectral zero at the zero frequency. This suppresses constant or slowly varying components in the input speech signal. Of further note is that RASTA uses several contiguous frames in its analysis, amounting to integrating information over about 150 ms. Thus, RASTA processing has more reliance on previous context than "vanilla" PLP.

One result of this processing is that transitions between speech segments are emphasized. That is, the RASTA analysis technique is less sensitive to slowly varying components. The bandpass filtering has the effect of passing modulations between 1 and 12 Hz. Experiments indicate that the RASTA-PLP processing method produces roughly the same word error rate as PLP alone on "clean" speech and significantly improves accuracy with speech in the presence of spectral interference (e.g., changed channel characteristics).

The RASTA-PLP feature analysis method transforms each window of the sampled waveform into a numerical representation, as a vectors of numbers. For the experiments described in Chapters 4, 5, and 6, "delta" features, which represent an approximation to the instantaneous rate of change of each feature, complemented the vectors of features produced by these feature extraction methods. Historical experience at ICSI has found that eighth-order RASTA-PLP is suitable for kind of recognition task described in this thesis. With energy and delta features, eighth-order RASTA-PLP, gives a total of 18 features per frame.

3.3.2 Probability Estimation: Neural Network

Equation 3.1 expresses the speech recognition process in mathematical terms. For a sequence of acoustic vectors $Y = y_1, y_2, \cdots, y_T$, where T is the number of individual observations (frames), and the series of actual words in an utterance is $W = w_1, w_2, \cdots, w_n$, the speech recognition process produces the most probable word sequence $\hat{W}$. Bayes' rule is used to decompose the desired probability into factors that are computable by the decoding process.

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ P(W|\mathbf{Y}) = \underset{W}{\operatorname{argmax}}\ \frac{P(W)P(\mathbf{Y}|W)}{P(\mathbf{Y})} \tag{3.1}$$

$P(W)$ is the *a priori* probability of the word sequence W, regardless of the acoustic input, and $P(\mathbf{Y}|W)$ is the probability of observing the vector of acoustic features, $\mathbf{Y}$, given that the word sequence W occurs.

To evaluate this equation and find the most likely sequence of words, the feature vectors generated by the front end are transferred to a phone probability estimator as depicted in Figure 3.1. The ICSI system and variants used in these experiments use a fully-connected, feedforward, multilayer perception with one hidden layer. The neural network uses the input features, plus additional context from 8 to 16 surrounding frames of features, to estimate the probability that the input corresponds to each of the defined categories. The network outputs represent estimates of posterior probabilities from which data likelihoods are calculated via Bayes' Rule (i.e., dividing by prior probabilities). Experience at ICSI has shown that 8 surrounding frames (i.e., a total context of 9 frames) performs well for typical recognition tasks, as in those described in this thesis. Using 16 surrounding frames (a total of 17 frames) can be useful for speech where the acoustic data is smeared over a longer time-span than 9 frames. The neural network was trained using simple, online error-back-propagation and softmax normalization. For the experiments in this thesis, the

neural networks typically had a hidden layer size of 400 units. To prevent overtraining, a technique known as early stopping was used. This technique reserves about 10% of the training data, referred to in this thesis as the cross validation set, for checking the progress of the procedure. Early stopping periodically assesses and maintains the generalization abilities of the neural network by testing on the cross validation set to decide when to stop training. With this technique the neural networks typically trained for seven or eight epochs (iterations through the training data) using a post-threshold, adaptive exponential decay learning rate. That is, the learning rate is held constant until the performance on the cross-validation set no longer improves. Then the learning rate is divided by two for each succeeding epoch until the performance on the cross-validation set again no longer improves. The cross validation subset also serves as a testbed for empirically determining various system parameters in these experiments.

3.3.3 Recognition Unit: Phonemes

The neural network converts feature vectors into estimates for the posterior probability of each phone, which are input into the decoding stage of the recognition system. The phoneme-based recognition units used in these experiments are fairly conventional, consisting of 56 context-independent phones based on the TIMIT phone set. This set is fairly complete for English. The "ICSI56" set of phones, listed in Appendix A.1, is composed mostly of phonetically representative exemplars of phonemes, with the addition of phones with acoustic distinctions such as stop closures, flaps and reduced vowels. The inclusion of these distinctive phones was historically found to promote the discriminative abilities of the artificial neural networks.

The baseline system was initially trained from the original phonetic labels taken from the manually-produced transcriptions. The labelers [30] who phonetically transcribed the Numbers task used a superset of the ICSI56 phone set. To derive phonetic targets for training the speech recognition system the original phonetic labels were mapped into the ICSI56 set. Since the ICSI56 set is a more limited group of phones, certain phonetic variation details in the original labelings were discarded. Some of these, such as aspiration, might have been useful for syllabification.

3.3.4 Lexicon

Dan Gildea (of ICSI) examined the hand-labeled, phonetic transcriptions of the training set and generated multiple pronunciations for each of the Numbers words, which covered approximately 90% of the pronunciation variations actually occurring in the training set. Scripts written by Eric Fosler-Lussier (also of ICSI) converted these pronunciations into a working lexicon. The mapped labels and the derived lexicon were used directly in the experiments in Chapter 4, since these were believed to have a close relationship to the actual acoustic manifestation of syllable features such as onsets.

For the experiments and analysis in Chapters 5 and 6, one iteration of forced alignment further refined and matched these labels and the derived lexicon to the learning capabilities of the neural net, using the baseline system with 400 hidden units. The forced-alignment

process is discussed later in this chapter. A one-time automatic adjustment of the lexicon revised word pronunciations and phone durations.[12] This eliminated a number of the hand-derived pronunciations as unused. The lexicon matched the corresponding neural network and training labels and vice versa. The resulting system exhibited a significant performance improvement over the original system.

3.3.5 Decoder

The probabilities from the neural network for each frame are input to a "Decoder," depicted in Figure 3.1. The decoder generates words and sentences by finding the maximum likelihood path through the probabilities, constrained by the pronunciation models. The ICSI system employs Viterbi decoding (a variant of dynamic programming) and, in some cases, stack decoding (a variant of the A* algorithm) to find the best path through the sequence of probabilities and thus the most likely words and sentences. The decoder uses a lexicon of hidden Markov models to enumerate the different pronunciations of the words and attach *a priori* probabilities to each version in the vocabulary. It also uses a language model that describes the way the words potentially fit together in a utterance. Decoding algorithms are discussed in more detail later in this chapter.

The ICSI system usually uses one of two decoders; a Viterbi decoder, Y0 (pronounced "why not") [88], and a start-synchronous stack decoder, called NOWAY [164, 163, 165]. The two decoders in general give comparable results, though they can have slight variations in the resulting sentence hypotheses. These dissimilarities are largely due to differences in the pruning strategies and choice of parameters. A set of hidden Markov models represents multiple pronunciations for each word. The state-to-state transition probabilities were untrained for these experiments and remained at a uniform $1/T$ where T is the number of transitions out of a particular state.

An N-gram language model provides the probability that some word, w_j, follows some sequence of words w_i through w_{i+N-1}. The language model used for the experiments described in this thesis was a simple bigram model (i.e., $N = 2$). Bigram probabilities (the probability that a certain word follows another word) can be calculated from the training set by counting the number of occurrences of each pair. "Backoff" methods estimate the probabilities for bigrams that do not occur in the training set [27].

Decoders generally use an empirically-determined value called a "language model" or "acoustic model" scaling factor to weight the influence of the language model over the acoustic information. A multiplicative value is applied to the log probabilities of sound classes or N-gram models. Typically, system builders use the language model scaling factor to balance the proportion of word insertion errors to word deletion errors. The relationship between the contribution made by of acoustic information and the language information is not well understood, but recognition system performance can be somewhat sensitive to the value of this parameter.

[12] The lexicon of this recognition system could have been adjusted with every iteration. Pilot studies, however, suggested that additional changes to the lexicon would yield minimal further improvement and serve only to obscure the experimental procedure.

Some of the experiments in this thesis used forced alignment (also called forced Viterbi). This procedure provides the correct word string to the Viterbi decoder, Y0, which uses the string to render and constrain the mostly likely path to the supplied words, given some acoustic input. It generates a new set of time-aligned labels for the utterance that can be used as targets in a subsequent neural network training. Using multiple, iterative applications of this procedure with optimization updates to a lexicon is sometimes called "embedded training." Researchers generally use this technique to automatically label acoustic input files when word transcriptions, but not phonetic transcriptions, are available. Even when phonetic transcriptions are available, the forced alignment technique, particularly in conjunction with lexicon updating, can help optimize the learning capabilities of the system by realigning phonetic segment labels. The resulting relabeling can help the neural network learn the distinction between the labeled patterns more effectively. Iteratively applied, this method converges to a training set labeling that the recognition system identifies most accurately. Since the labels are shifted automatically, however, the new labels may not entirely agree with the acoustic evidence in a way that is obvious to a human researcher.

3.3.6 Evaluation

To evaluate the performance of a speech recognition system, the most commonly accepted measure is word error rate. Specifically, word scoring for these experiments used a dynamic programming algorithm that computed the minimum number of substitutions, insertions and deletions between the reference (correct) string and the output of the speech recognition system. While universally applicable, simple word scoring does not fully examine differences between one system and another, so some of the experimental systems in this thesis were evaluated by additional criteria besides word-error rate, in order to provide some insight into how accuracy might be improved. These are described in more detail in Chapters 5 and 6.

3.4 Speech Decoding

Speech decoding is the process of finding the most probable sequence of words given a sequence of probabilities based on acoustic representations and other knowledge sources, also governed by Equation 3.1. For decoding, the sequence of observation vectors $\mathbf{Y}$ is defined to be the vector of acoustically-based probabilities. For the systems used in this thesis, artificial neural networks generated these probabilities from acoustic observation vectors.

Practical concerns often constrain the quest for algorithms and heuristics that produce the highest possible recognition accuracy. These factors conflict with one another, affecting the allocation of both human and computer resources, and requiring tradeoffs. Many different algorithms exist, each with varying implementation details, parameters, inputs, outputs, target tasks and performance.

Template matching, or "dynamic time warping," (DTW) is a decoding method that was successfully employed for small vocabularies in the 1960s through the mid 1980s. Dynamic

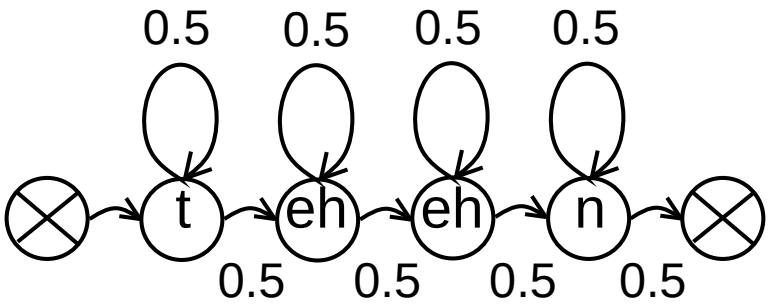

Figure 3.2: An example of a typical HMM, for the word "ten." The phone /eh/ has a minimum duration of two states.

programming matched acoustic data directly to templates of whole words [172, 43]. Since the templates modeled the entire word in detail, the storage cost of these templates and the computational cost of the search are impractical for large vocabularies. Researchers attempted to use sub-word models, but found that coarticulatory variations became more significant with smaller units and more difficult to model under DTW. Rosenberg, Rabiner, Wilpon and Kahn concluded that whole word prototypes provide superior results to models based on demisyllables, but acknowledged that storage and computation limitations made the sub-word units attractive [167].

Stochastic methods were initially explored in the 1970s and achieved wide acceptance in the 1980s as a way to represent coarticulation and other variations in speech. Principal among these methods is the stochastic, finite-state automaton known as the hidden Markov model (HMM). Today, researchers typically concentrate on HMM-based, stochastic model methods. HMMs provide more facility for incorporating coarticulatory effects into sub-word models than do comparable templates used in dynamic time warping. The HMM [157, 11, 95] and its variants for decoding with more general sub-word units have proven successful for probabilistically representing speech for large vocabularies with sub-word units, while still keeping the decoding process computationally tractable. In HMMs, words are represented as a sequence of sub-word units (usually based on phonemes), thus representing a reduction in the computational and storage costs over whole-word models.

Figure 3.2 shows a representative HMM for the word "ten." Each circle represents a state and is typically associated with a phone, or a part of a phone; this state is the "hidden" part of the HMM. In the production view of the HMM, the model generates observables (i.e., feature vectors) as it progresses from one hidden state to the next. Decoding speech reverses the process. Decoders use the observables to hypothesize the hidden state of the HMM. The model includes probabilities of transition, both back to the same state and on to other states, that govern the likelihood of particular paths. If the sub-word unit is too small and the categories are not carefully defined, this method may poorly model coarticulatory effects within a word. While HMMs have limitations and associated problems with modeling speech, it is not the purpose of this thesis to explore this issue or HMM alternatives.

Several HMM decoding algorithms are currently popular.[13] The algorithms chiefly focus

[13]Deller's book, *Discrete-Time Processing of Speech Signals* [43] contains a review of HMMs and decoding.

on selecting the most likely HMM sequence given a set of acoustic information vectors. The algorithms may also be combined to capitalize on each method's strengths.

These decoding algorithms include:

- Decoding with maximum likelihood probabilities [12] [43]: This algorithm involves the calculation of the total probability that a particular model produces a given observation sequence (i.e., a set of acoustic-feature vectors). This requires the summation of the effect of any and all paths through the particular model, usually via the forward algorithm.
- Decoding with an approximation to the maximum likelihood: This less optimal, but computationally more tractable approach, uses the likelihood value of the single best state sequence through any HMM that produces the given observation sequence as an approximation to the total probability of the model. This may not produce the same path as decoding with true maximum likelihoods. The two standard algorithms for computing the best state path are:
  1. Dynamic programming [172, 10, 138]: Also called Viterbi decoding,[14] the algorithm is characterized by a "breadth-first" search strategy, where all relevant paths are extended simultaneously for each time step.
  2. Stack decoding [93, 102, 150, 170]: Often used in combination with a Viterbi criterion, stack decoding[15] is characterized by a "depth-first" search strategy. A single hypothesis is extended until search directives dictate that it is no longer the most likely solution. Then the decoder chooses a new hypothesis to be extended and the process repeats.

The above search algorithms are provably optimal only for complete searches which are usually extremely slow and memory intensive in practice. Implementation details are often the key difference between a usable decoder and an impractical one. For example, one important technique used to improve execution time is the use of a tree-structured lexicon or network [145]. Combining the shared prefixes of the words in a lexicon reduces redundant computation and memory usage without sacrificing accuracy.

Suboptimal variations are used to elicit the maximum performance for a reasonable amount of computing resources. "Suboptimal" heuristics are so named because they can not be guaranteed to lead to the best solution. Often it can be proven that these heuristics will definitely miss the best available solution under certain pathological conditions. Despite this disadvantage, suboptimal methods usually use a small fraction of the computing resources of optimal algorithms and deliver high accuracy for a large percentage of situations. Researchers find this property a compelling reason to develop and use suboptimal methods.

Various techniques for recovering from the search errors resulting from suboptimal algorithms have received considerable attention. Popular techniques include:

Comerford *et al.* give a short overview of HMMs in [32].

[14] "Viterbi decoding" is a term from the digital communications field.

[15] Also known as A^* in the artificial intelligence field, and related to the Fano algorithm in digital communications. It is interesting to see how similar algorithms emerge from different fields.

- Multiple passes, including backwards processing [4]: By varying the way the system performs decoding during each pass, researchers hope to compensate in the sum of all passes for deficiencies in individual passes. Progressive search techniques [137], in particular, start with crude but cheap decoding techniques and use more refined and computationally intensive decoding schemes for later passes.

- N-best lists of sentences [177, 176]: By generating N of the most likely word sequences rather than just the single most likely one, the correct answer is more likely to appear somewhere in the resulting output set. Systems can then use subsequent processing to re-rank the various hypotheses.

- Word graphs and lattices [146, 139, 140]: A word graph is defined to be a directed acyclic graph where each edge corresponds to a word with a score and each node is a point in time. There is less agreement on the definition of a word lattice. Some authors use the word lattice in a manner consistent with a word graph. Others suggest that word lattices contain only word-order information and allow the possibility of temporal overlaps, or even gaps, between words. In this thesis, the terms "word lattice" and "word graph" are synonymous and both refer to the directed acyclic graph. Similar to N-best lists, graphs and lattices contain more information than just the best sequence. Systems use postprocessing, perhaps incorporating additional kinds of information, to rescore the information in the graphs and lattices.

- Combinations and variants of the above: One example of this is the work of Soong and Huang, who used a forward search with a Viterbi criterion, then a backward search with a stack decoding scheme to produce an N-best list [185]. Researchers have combined various techniques and methods in order to capitalize on as many advantages as possible.

In addition to these basic algorithms, system engineers use many heuristics singly or in combination to improve search performance, trading accuracy for speed or memory usage. New strategies, variations and combinations of existing techniques appear frequently. Pruning heuristics can improve the computation times, but have the potential of introducing search errors, where the decoder discards the correct answer due to the direction taken by the pruning strategy. Practical decoders make extensive use of pruning. The art in the design and implementation of decoders is in quickly discarding as much of the search space as possible without losing the correct answer.

Examples of current popular pruning heuristics include:

- Beam search [122] [43] (sometimes with multiple beams): "Beams" limit the extent of the search space considered during Viterbi decoding so that the decoder expends search effort in a narrow region where the most probable hypothesis is likely to reside.

- Fast match [9, 8]: Typically used with stack decoding, fast match helps select the next candidates for adding to the current hypothesis by looking ahead in the acoustic probability stream and performing quick, coarse phonetic matches.

- State- or phone-level pruning, such as deactivation pruning [164]: The decoder completely deactivates phones (i.e., assigns a probability of zero) that appear to have comparatively low probability. These phones and their corresponding words are not considered during the decoding process, improving execution speed.

ASR systems apply higher-level knowledge at various levels and in a variety of forms. The experiments described in Chapters 4, 5 and 6 used a bigram language model to describe the probabilistic word structure of the Numbers utterances. N-gram models [10, 43] are a popular method of adding language constraints into the decoding process. Bigram ($N = 2$) or trigram ($N = 3$) models are common. Other researchers have been studying a variety of long-span language models and different ways to incorporate more knowledge-based sources of information.

Clearly, there has been great effort expended in speech decoder design and implementation. Many decoders for current tasks (e.g., for 64,000-word vocabularies) perform recognition in close to real-time on consumer personal computers. Yet there is still much to improve in machine speech recognition. As researchers discover more about the nature of human speech recognition and incorporate new processing techniques and sources of speech information, the demand on computing resources by decoders will continue to increase.

The work described in this thesis required a pool of several decoders, since each had unique capabilities. The four different decoders used were:

- A simple, small-vocabulary, special-purpose decoder with an explicit syllabic level (Chapter 4).
- The Y0 decoder [88], a general purpose, Viterbi decoder with forced alignment capabilities (Chapters 5 and 6). Y0 uses beam search to limit the number of simultaneous hypotheses.
- The NOWAY decoder [164, 163, 165] (Chapters 5 and 6) . NOWAY is a start-synchronous stack decoder using a Viterbi criterion. NOWAY incorporates phone deactivation pruning and limits the creation of new hypotheses with beam-search-like techniques. The NOWAY decoder has the added capability of producing word lattices.
- The lattice decoder LATTICE2NBEST [166] which uses a NOWAY-like stack decoding algorithm. This decoder determines the best sequence of words from a word lattice (Chapter 6).

3.5 Combination of Multiple Streams

The combination of information from multiple sources is an attractive approach to the problem of speech recognition. Merging information has the potential to exceed the summed performance of the individual parts. The Hearsay system [48] of the 1970s and early 1980s attempted to combine information from different knowledge sources to first make a hypothesis and then correct it. More recent, anecdotal evidence suggests that combining information at the feature extraction level from even slightly different analysis methods

can lead to increased recognition performance. Chapters 4 and 6 describe the exploration of combining two speech recognition systems, one oriented towards the phoneme, and the other incorporating syllable-based information. While the phoneme-based system is well established and highly optimized, the more recently developed syllable-based systems are less optimized. By combining the two, the advantage of the maturity of the phoneme-based can potentially be enhanced by the innovation of the syllable-based system for a more accurate overall result.

The combination of multiple sources of information can occur in a free-form manner at arbitrary levels in the recognition process. The experiments in Chapters 4 and 6 concentrate on merging streams of information from two recognition systems at the frame, syllable and whole-utterance level. The pattern recognition community has proposed quite a few algorithms for combining classifiers, particularly for handwriting recognition. Since the focus of this thesis is not on the science of combining classifiers, but rather on the combination of specific speech recognition systems, the literature overview here will be brief and will only discuss the combination algorithms relevant to the work described in this thesis.

A speech recognition system normally uses a single form of feature analysis and a single decoding method to generate phonetic probability estimates. As noted by Ho, Hull and Srihari, a perfect form of feature analysis or method of decoding is difficult to define for problems with a large number of classes and noisy inputs. Classifiers using different feature analyses and different decoding paradigms can result in different errors, even if each separate classifier achieves about the same overall percentage error. The crux of the combination problem, then, is to determine the ideal combination algorithm to take advantage of each classifier's strengths ("classifier correlation") [87].

There are two fundamental approaches to combining the outputs of more than one classifier: (1) merge the outputs of each single classifier acting in parallel over some input, in a uniform way to produce a global output that represents a group consensus, or (2) chose for each target one the classifiers in a group, acting in parallel, to represent the whole. These are called, respectively, "classifier fusion" and "dynamic classifier selection" [206]. Woods, Kegelmeyer and Bowyer approached the problem of combining multiple classifiers by using a dynamic classifier selection algorithm with a local accuracy criterion. That is, when the classifiers differed in their outputs, the algorithm assessed local accuracy estimates of each of the classifiers from the "nearby" examples in the training data in order to determine which classifier to use for the final output. In preliminary experiments with the Numbers task, the computation of local accuracy appeared to be a poor estimate of the reliability of the classifier, particularly when the test case was not represented in the training set (e.g., in the presence of noise).

Other classifying techniques include the simple majority vote. Preliminary experiments with the Numbers data supported the intuition that when each classifier's accuracy was fairly high, voting eliminated a significant number of errors. When the individual classifiers each had a large error rate, however (for example, with the addition of reverberation), the voting method did not significantly improve the overall error rate. This was due to the large variance in recognition answers when the input was noisy.

There are also methods based on confidence measures. Unfortunately, it is difficult to

define a confidence measure that is comparable between different recognition paradigms. Similarly, other numerical scores such as distances and estimates of posterior probabilities are difficult to use directly because of the basic incompatibility of the assumptions in the data to be combined [87].

Aside from the question of how to combine classifiers is the question of which classifiers to include in the merging. If estimates of classifier performance were exact, no such choices would be necessary. Estimates of accuracy are flawed, however, especially for unexpected inputs, so the choice of classifiers to be combined must be carefully considered. Experimental evidence from Woods *et al.* further illustrates this point. They found that certain subsets of four classifiers outperformed a combination of all five classifiers [206].

In the field of automatic speech recognition, the decoding stage adds an additional level of complexity in the combination process. As a result of the dynamic programming in decoders, there is only an indirect relationship between probability estimation and improved accuracy. Common approaches to combining multiple sources of information include N-best list rescoring and word-lattice rescoring.

For the work in Chapter 4, the combining took the form of constraining the decoding process of one system with the output of another. The combination methods used in Chapter 6 are based on the linear combination of the log probability outputs of each recognition system, a standard classifier fusion technique. These proved to be more successful than attempts at classifier selection. Chapter 6 further describes investigations of differing frameworks for combination where the granularity of the combination unit was systematically varied from the whole sentence to the phone/frame level.

For combining at the frame level (i.e., at the output of the neural network), simply multiplying the corresponding probabilities for each frame was effective. The result can then be passed into the decoding process as usual. A method recently reported by Bourlard, Dupont and Ris provided an avenue for experimenting with combining systems at the syllable level. Bourlard *et al.* experimented with what they term HMM-recombination [19], a variant of the HMM decomposition technique [193, 194, 66] more commonly used to statistically decompose noise and speech (independent sources of sound information). Dupont, Bourlard and Ris [45] have begun investigating combining speech information from several streams, each representing a different time-scale (e.g., phones and syllables) with some asynchrony permitted between recombination points. Potential advantages listed by Dupont *et al.* included better robustness to noise. The work presented in Chapter 6 used their technique to combine hypotheses at the syllable level during decoding. The combination procedure at the utterance level used N-best rescoring to determine the best utterance overall. Each method had distinct advantages and disadvantages for implementation and optimization. These issues are discussed further later in this thesis.

3.6 Summary

While some speech recognition applications are currently enjoying some measure of commercial success, accurate and robust speech recognition, particularly for naturally spoken, conversational speech, is still a considerable challenge for ASR systems.

The Numbers corpus is a small vocabulary, naturally spoken speech database ideal for these exploratory studies. For evaluation purposes, the Numbers corpus provides a manageable but nontrivial recognition task. An artificially reverberated version of the test set was used to model one kind of adverse environmental condition. Humans can recognize both the clean and the artificially reverberated test set with very few errors. Machines, on the other hand, usually produce dramatically more errors than humans [120]. Thus, there is considerable room for improvement in ASR. The work reported in Chapters 4, 5 and 6 used the ICSI speech recognition system (a hybrid neural network/hidden Markov model paradigm) as a starting point.

The experiments in this thesis involve manipulating the decoding stage of the speech recognition process, either by introducing syllable onsets or by combining two streams of recognition output. There are many possible combination methods. The experiments described in later chapters focus on linear combinations of probabilities at the frame-, syllable- and utterance-level. Each of these combination strategies has distinct advantages and disadvantages.

The rest of this thesis focuses on using the decoding and combining techniques outlined in this chapter to incorporate syllable-based information into speech recognition.

Chapter 4

Integrating Syllabic Onsets

Accurate estimation of the beginnings of spoken syllables can reduce the number of viable utterance hypotheses and thus improve automatic speech recognition performance. In the work described in this chapter, we[1] explored the integration of syllable onsets into the speech recognition process via a specially-designed decoder. The first set of experiments used artificial onsets derived from advance knowledge of correct syllable boundaries. The results of these trials showed that onset information could be useful in improving recognition accuracy. The second set of experiments used onsets estimated directly from acoustic information. This added information produced a reduction (10% relative) in the word-error rate for the Numbers task. The latter experiment also suggested additional study of coordinating acoustic and lexical representations of speech. From this arose the inspiration for the work described in Chapters 5 and 6.

This chapter begins with a review of the background and previous work with syllable boundaries. Section 4.2 describes the special purpose decoder used for these experiments. The recognition system is further outlined in Section 4.3. Section 4.4 reports the results of the experiments with syllable onsets.

4.1 Detecting Syllable Boundaries

Approaching the question of speech recognition from the syllabic level, rather than from the phonetic level, may confer several benefits, as discussed in detail in Chapter 2. Statistical, structural regularities suggest that the boundaries of syllables may be more precisely defined than that of phonetic segments in both speech waveforms and spectrograms. This effect is particularly visible during conversational speech. Research by Cutler, Butterfield and Norris indicated that humans perceive word-initial clusters of phones as integral units [37]. Statistics gathered by Greenberg show that syllabic onsets are expressed in canonical form far more frequently than the rest of the syllable. The syllabic onset exhibits more stability than either the nucleus or the coda [78]. Figure 4.1 shows roughly regular patterns at

[1] The study detailed in this chapter was the result of collaboration between Michael Shire and myself, with additional input by Steven Greenberg and Nelson Morgan. We described parts of this work during a presentation at the International Conference on Acoustics, Speech and Signal Processing, 1997 [210].

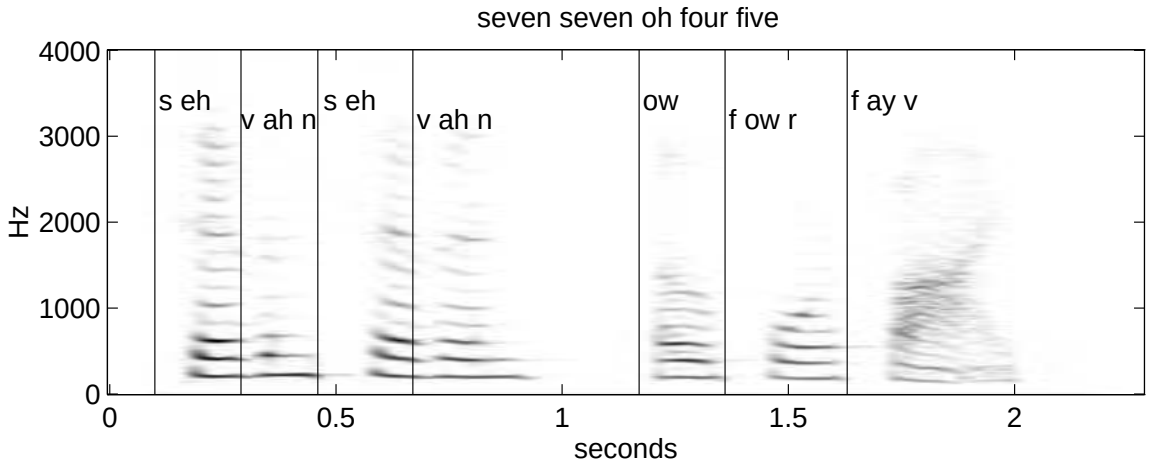

Figure 4.1: Spectrogram of the utterance "seven seven oh four five" with syllabic onsets marked as vertical lines [184].

the syllabic level in a representative example spectrogram with the beginnings of syllable onsets marked. Attempting to detect syllable boundaries and nuclei is not a new idea. The approach reported in this chapter differs from previous work in that these experiments are perceptually-oriented and focus on the recognition of spontaneous, naturally spoken speech.

4.1.1 Syllable Nuclei and Boundaries in Speech Recognition

Researchers have documented studies focused on detecting syllabic properties such as boundaries and nuclei in speech research literature since 1975, as described in Chapter 2. The two most closely related projects are discussed in further detail below.

In the mid-1970s, Mermelstein described a method for the automatic segmentation of speech into syllabic units using a loudness criteria [129]. Hunt, Lennig and Mermelstein incorporated this method into a speech recognition system [92, 91]. As mentioned in Chapter 2, Mermelstein calculated their loudness function over the entire power spectrum. In their recognition experiments they used a single speaker for both training and testing the system. The test set comprised the same word sequences as the training set, re-recorded by the speaker. They concluded from their experiments that a syllable segmentation system provides sufficiently encouraging results as to warrant further investigation.

In the experimental work described in this chapter, the focus is on syllable onsets rather than on boundaries because of the perceptual evidence that syllabic onset structures are better preserved in spontaneous conversational speech than syllabic coda structures. Syllabic onsets were estimated from distributions of energy in 9 separate bands. The recognition system used in these experiments was tested in a speaker-independent manner; speakers and utterances from the training set did not appear in the test set. These experiments have also had the benefit of more experience with stochastic methods and decoding strategies which has become available since the time of the experiments reported by Hunt *et al.*

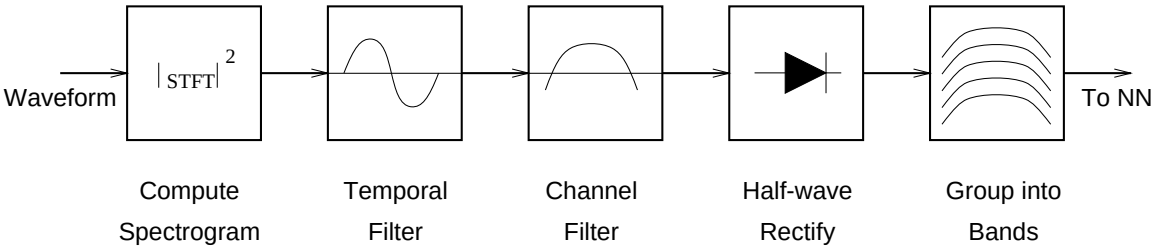

Figure 4.2: Major processing steps for deriving the syllable onset features [184].

More recently, Green, Kew and Miller [73, 72] have worked on estimating syllable onsets for incorporation into the SYLK project, ultimately for the recognition of TIMIT utterances. They focused on improving the discriminative abilities of their system through refinement of training techniques. The SYLK project reported phone and phone-class results, but did not report word recognition scores.

The work described in this chapter differs from the SYLK project in that the experiments focus more exclusively on using specialized acoustic features. The perceptually-based method of estimating syllable onsets, developed by Shire and Greenberg [210, 184], allows the incorporation of onset information into the decoding process and produces results with word recognition.

Estimating syllable nuclei has also been an active area of interest. In the mid-1970s, the Hearsay system's [48] word sequence generation process started with a syllable-nucleus-identification function. The recognition system generated hypotheses based on the positions of the nucleus. For German, several researchers have used syllable nuclei position estimates [203, 169, 154, 155, 162], sometimes in combination with demisyllable-based speech recognition systems. The syllable onset detection method discussed here could potentially be combined with nuclei detection for overall better segmentation, but so far work by Shire towards this end remains inconclusive.

4.1.2 Detecting Syllable Onsets

This investigation used a perceptually-oriented method of estimating syllable onsets developed by Shire and Greenberg for spontaneous, conversational speech [184], also described in [210].

In these experiments, artificial neural networks estimate the probability of syllable onsets and thereby automatically provide online calculation of syllable boundaries. Patterns of synchronized rises in subband energy spanning adjacent subbands typically characterize syllable onsets [74]. The time course of these coordinated changes in energy level roughly correlate with the length of syllables in naturally-spoken English, i.e., about 100-250 ms.

Figure 4.2 illustrates the signal processing procedures designed to enhance and extract these observed acoustic properties. The process decomposes the speech waveform via short-time Fourier analysis into a narrowband spectrogram, and then convolves the result with both a temporal filter and a channel filter, effectively creating a two-dimensional filter.

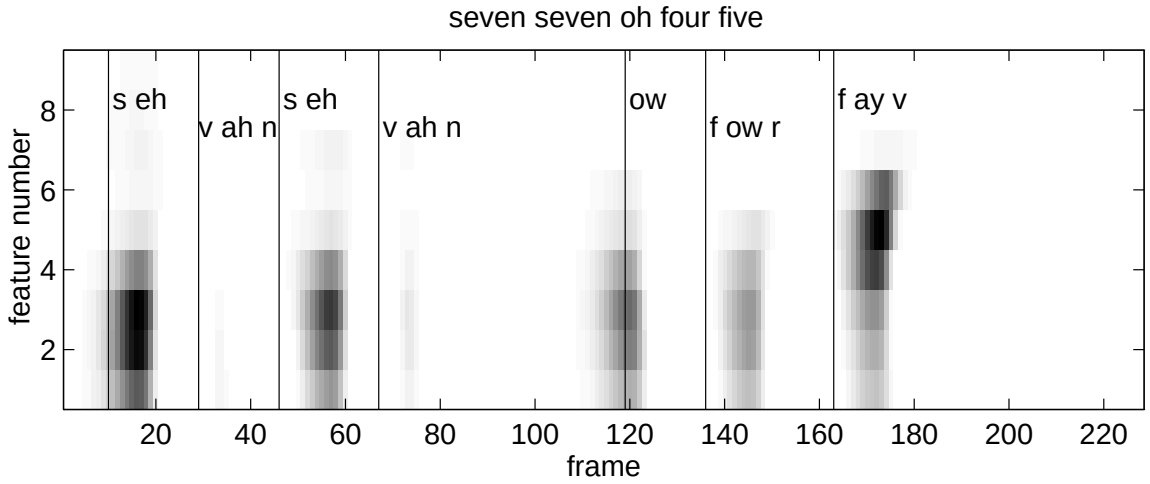

Figure 4.3: Example of onset features derived for the utterance "seven seven oh four five." The vertical lines denote syllable onsets as derived from hand-transcribed phone labels. [184]

The temporal filter (a high-pass filter analogous to a Gaussian derivative) was tuned for enhancing changes in energy on the order of 150 ms. The filter smoothes and differentiates the waveform along the temporal axis. The (Gaussian) channel filter performs a smoothing function across the channels, providing weight to regions of the spectrogram where adjacent channels are changing in coordinated fashion.[2] Half-wave rectification preserves the positive changes in energy, thus emphasizing the syllable onsets.

Large values in this representation correspond to positive-going energy regions where hypothesized syllable onset characteristics occur. The channel outputs are subsequently averaged over a region spanning 9 critical bands, the result of which is referred to as "roband" features, illustrated in Figure 4.3.

The process produced updates of these features every 10 ms. The resulting vectors were concatenated with eighth-order RASTA-PLP [86][3] features computed over a 25-ms frame every 10 ms. This combination formed the input to a neural network for estimating the locations of syllabic onsets. For the given acoustic patterns described above, a trained, single-hidden-layer, fully-connected, feedforward multilayer perceptron with 400 hidden units estimated the probability that a given frame was a syllable onset. For the purposes of training, a series of 5 frames represented the syllable onset (as derived from automatic segmentation of phonetic hand-transcriptions), where the initial frame corresponded to the actual beginning of the syllable.

A simple numeric threshold applied to the probability estimates generated by the neural net determines the identification of any given frame as a syllabic onset. The choice of the threshold value primarily optimized correct identification of onsets and secondarily minimized insertions, on the cross-validation set. This procedure correctly detected 94%

[2] More details concerning the filter specifications of this system can be found in [184].
[3] RASTA-PLP is described in more detail in Section 3.3.1.

of the onsets computed from phonetically transcribed data (within the 5-frame tolerance window defined for training). The procedure also mistakenly inserted syllabic onsets where there were none (false positives) in 15% of the frames outside the tolerance window of any onset. A syllable-based decoder uses these onset decisions as frames corresponding to the beginnings of syllables.

4.2 Speech Decoder With Additional Syllabic Level

This study involved the design and implementation of a special-purpose speech decoder, suitable for small vocabulary tasks. Its most notable departure from standard decoders is that it incorporates an intermediate, syllabic level of abstraction between the level of the phone and the level of the word or sentence.[4]

This decoder processes phonetic probabilities from a neural network using a conventional Viterbi algorithm with hidden Markov models. Using a bigram syllable grammar, the decoding process creates a syllable graph (a derivative of the word graph described in Section 3.4) from the phonetic information. Trials without a syllable grammar showed that the grammar plays an important role in the efficient pruning of hypotheses. Each arc in the graph represents a single syllable hypothesis, to which the decoder assigns a likelihood value. The endpoints of the arc indicate the beginning and ending times of the syllable hypothesis. The next stage, the program's stack decoder, uses this syllable graph as input along with a bigram word grammar. The stack decoder[5] determines the most likely sequence of words given the syllable graph. This procedure is a type of multiple-pass decoding method and is conceptually similar to the two-level dynamic programming algorithm [171]. The additional complexity of the decoder design permits the explicit representation of the relationship of phones to syllables and syllables to words. The algorithm's representation of the syllable as an intermediate stage in the design allows easier expansion and experimentation at the syllabic level. Syllable onset information appears as an additional input at the level of the syllable graph, as illustrated in Figure 4.4.

To validate the design and implementation of the special-purpose decoder, we compared the performance of the recognition system with this decoder to the performance of the same system except with YO and NOWAY performing the decoding function. Without the introduction of syllable onset information, the special purpose decoder produces word-error rates on the Numbers corpus roughly comparable to the more established decoders available, given similar input.

4.3 Recognition System

The recognition system for these experiments was derived from the ICSI hybrid HMM/MLP system, described in Section 3.3, with extensions for incorporating syllable onsets. The baseline system used the following elements:

[4] Future work that was planned for this decoder was eventually subsumed into other directions which did not require the use of this specially-designed decoder.

[5] Stack decoding is also discussed in Section 3.4.

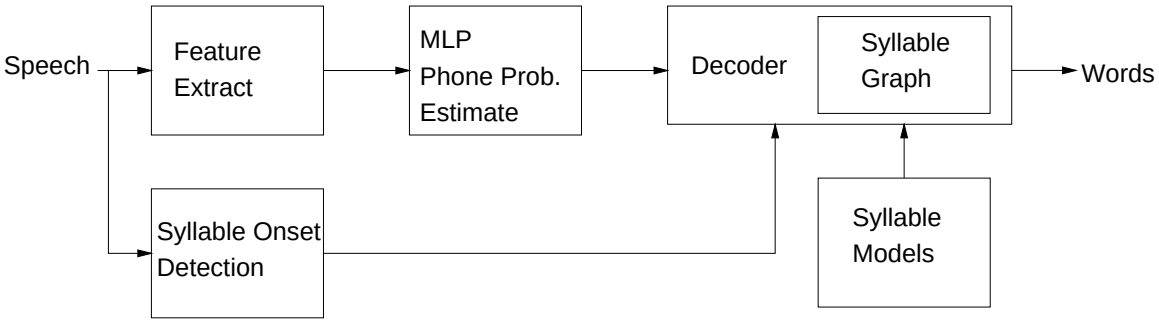

Figure 4.4: Illustration of recognition system incorporating syllable onset information into decoding [184].

- RASTA-PLP8 features, 25-ms frames, calculated every 10 ms. (A total of 18 features per frame.)
- Phone-based recognition units.
- A 400 hidden unit, fully-connected, single hidden layer neural network with 9 frames (nominally 105 ms) of neural network input context.
- A special-purpose decoder with an explicitly represented syllable level, as discussed in the previous section.
- A bigram backoff grammar[6] derived from the training set.
- A single pronunciation lexicon for the initial pilot experiments with artificially derived onsets and a multiple-pronunciation lexicon for the later experiments with acoustically-derived onsets.

As described above, the syllable-onset estimation system used the following elements:

- RASTA-PLP8 features, 25-ms frames, calculated every 10 ms. (A total of 18 features per frame.)
- Roband features, calculated every 10 ms.
- Onset/no onset training targets.
- A 400 hidden unit, fully-connected, single hidden layer neural network with 9 frames of neural network input context.

4.4 Experiments

The experiments with syllable onsets were performed with the "core subset" of the OGI Numbers corpus described in Section 3.2. In each set of experiments, decoding parameters,

[6]The grammar is described in Section 3.2.

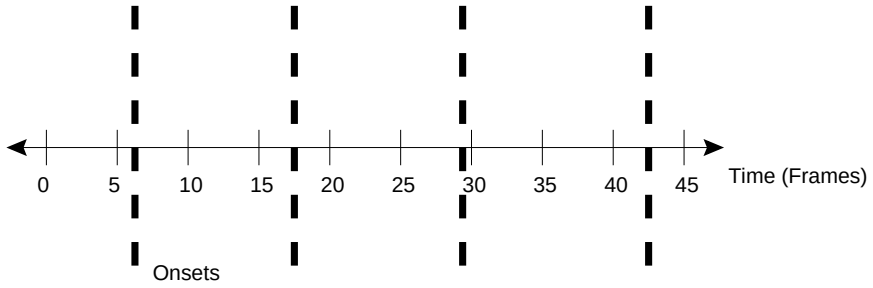

Figure 4.5: Illustration of where the decoder hypothesized the beginning of syllable models in the first set of pilot experiments (at vertical, dashed lines only).

such as word-transition penalty and language model scaling factor, were independently derived from a series of experiments with the cross-validation subset. Any parameter tuning that was needed involved only this cross-validation subset of the training set. The experimental results below are reported for the Numbers development test set.

4.4.1 With Previously Determined Syllabic Onsets

An initial set of experiments (a pilot study) using onset information derived from *advance* knowledge[7] of the true word-transcriptions of the test utterances helped to ascertain the potential value of incorporating syllabic onsets into decoding.

The lexicon for this set of experiments included 32 single-pronunciation words, comprising 30 different syllables. These pronunciations were derived from the Carnegie Mellon University (CMU) dictionary [202] and syllabified according to standard dictionary principles. In general, the pronunciations reflected orthodox pronunciations of the words. For example, the word "twenty" was defined (in ICSI56 phone orthography, listed in Appendix A.1) as "t w eh n t iy" even though the word is often pronounced in actual speech without the middle "t," as in "t w eh nx iy." An embedded training process calculated context-dependent phonetic durations from the training data.

A forced-alignment procedure[8] with the lexicon described above generated phone alignment labels based on word transcriptions, which were provided for all the utterances in the test set. An automatic process inferred syllable boundaries from the phone labelings using the syllabified lexicon. Due to this top-down process, the resulting syllable boundaries corresponded to a word-level idealization of the utterance. Artificial syllabic onsets with a duration of one 25-ms frame were then derived from these forced-alignment labels.

During recognition the decoder hypothesizes a syllable model only when its beginning frame is identified as an onset frame by advance information, as depicted in Figure 4.5. In these experiments, the decoder contained no restriction on the end-of-syllable-model location. It was therefore possible for one model to overrun later-indicated onsets. The decoder had access to only syllabic onset information from the test set and not to any other

[7] That is to say, this was a "cheating" experiment.

[8] Section 3.3 discusses the advantages and disadvantages of forced alignment.

System	Word Error Rate
no onset information	10.8%
with known syllable onset times, Total frames/onset = 1	6.7%

Table 4.1: Performance results (word error rates) for decoding using a single-pronunciation lexicon, with and without artificial syllabic onsets derived from forced alignment. Represents ideal conditions.

prior knowledge from the test set, such as phonetic information. Therefore, changes in the recognized output can be associated directly with the onsets provided.

If the dynamic programming decoding procedure and the speech input were ideal, and if the available phonetic information were sufficient to resolve ambiguities, the addition of artificially derived syllabic boundary information would, in theory, provide little or no improvement in recognition performance. In principle, the decoding process assumes that models can begin at any frame, including the ones specified as syllabic onsets. In this experiment, however, the incorporation of the artificially-derived syllable segmentation information reduced the word error rate from 10.8% to 6.7% (Table 4.1), a substantial relative reduction of 38%. When the system's phone probability estimator, in conjunction with the decoder, hypothesized incorrect word sequences as the most likely recognized phones, these sequences often had syllable onsets that did not match the beginnings of syllables in the correct utterance. Supplying syllabic onsets compensated for this kind of error by allowing the recognition system to discard misaligned hypothesis. The decoder is able to override the phones erroneously recognized as most likely, resulting in greater word accuracy. The large reduction in word error observed suggests that correct syllabic boundary information can significantly improve speech recognition performance when incorporated into the decoding process. This may be due to the syllable onset information providing a separate dimension of knowledge about the speech signal from the phonetic information. If the estimation of syllable onsets from acoustic information can be performed accurately enough, this pilot experiment shows that the syllable onsets can overcome shortcomings in the phonetic estimates to produce a significant reduction in error rate.

A second series of experiments focused on assessing the precision required for syllable onset estimates to be of significant benefit in decoding. Multiple 25-ms frames, with a 10-ms step between the beginnings of adjacent frames, were associated with each onset, instead of just one frame. The decoder hypothesized the beginning of syllable models at any of the expanded onset frames, as shown in Figure 4.6. As the window of frames for each onset widened from 5 to 13 frames, the word error rate increased, as shown in Table 4.2. In the last experiment each onset encompassed up to 13 frames where syllables could be hypothesized and the word-error rate was still 21% better (10.8% versus 8.5%) than without the onset information. This suggests that some erroneous word sequences recognized by the system had corresponding syllable onsets that were more than 13 frames from the actual onsets of the utterance. Providing the onset information, even with 13-frame precision, allowed the decoder to discard these misaligned hypotheses. Thus, syllabic onset information of even

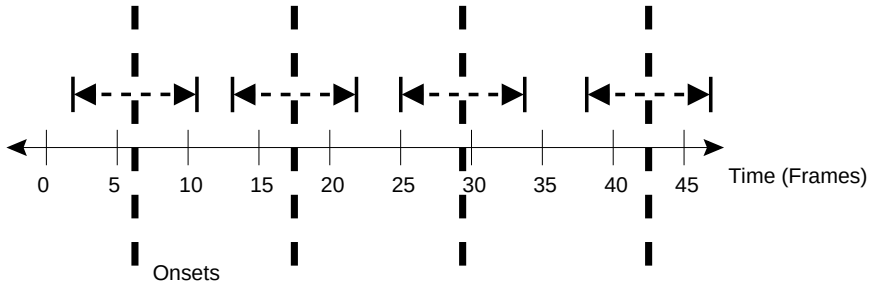

Figure 4.6: Illustration of where the decoder hypothesized the beginning of syllable models in second set of pilot experiments (at vertical, dashed lines plus a fixed interval to the left and right of the onset).

Number of Frames Centered on Each Onset	Error Rate
Total frames/onset = 5	7.3%
Total frames/onset = 9	7.8%
Total frames/onset = 13	8.5%

Table 4.2: Performance results (word error rates) for single-pronunciation decoding, using syllable hypotheses that were allowed to begin within several frames of artificial onsets derived from forced alignment.

limited precision can be beneficial in decoding in speech recognition systems. Fairly broad hints as to the location of syllable boundaries were sufficient to overcome faulty phonetic recognition and improve recognition accuracy in many utterances. These results indicate that syllable onset information, if reasonably accurate, has high value, separate from that of phonetic information.

4.4.2 With Acoustically Determined Syllabic Onsets

Since speech recognition systems do not usually have access to true syllabic timing information, systems must infer syllable boundaries from other sources. In the next series of experiments, the decoding process was constrained by *acoustically-derived* syllable onset estimates from the procedure outlined in Section 4.1.2. The trials described below did not incorporate any advance information from the test utterances.

The subset of the Numbers corpus used for these experiments was phonetically transcribed at OGI [30]. Dan Ellis' (at ICSI) adaptation [47] of Bill Fisher's (NIST) syllabification program TSYLB2 [51] automatically generated syllable boundaries for the training data from the phonological interpretations of the phonetic transcriptions. The neural network training procedure in Section 4.1.2 used these onsets. These syllable boundaries do not necessarily respect word boundaries, unlike the syllabifications used in the pilot experiments.

The neural network was trained on targets derived directly from the phonetic hand-

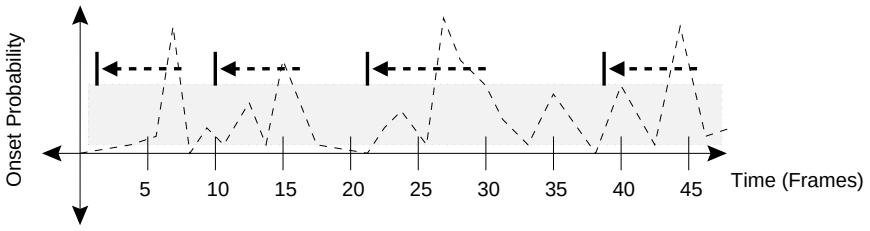

Figure 4.7: Illustration of where syllable models were hypothesized to begin in experiments with acoustically-determined probabilities from an onset- detection neural network (at frames covered by horizontal, dashed arrow).

transcriptions to preserve the relationship between the acoustic properties of syllable onsets and the linguistic assignment of phones as much as possible. To provide a closer match between the phonetically transcribed material and the acoustically-determined syllabic onsets from the neural network, a new lexicon, with attendant grammars, was needed. Dan Gildea and Eric Fosler-Lussier (both at ICSI) created the new lexicon using the phonetic transcription data from the training set of the Numbers corpus. The resulting lexicon included 32 words (and their range of 178 possible pronunciations), comprising 118 different syllables [69]. This lexicon included approximately 90% of the pronunciation variations in the corpus, as reflected in the hand-transcriptions.

The incorporation of multiple pronunciations based on the actual manifestations of words in the training set improved the performance of the baseline system relative to the pilot experiments, which used only single, canonical pronunciations for each word. Words can vary in pronunciation depending on factors such as phonological context and individual speaker characteristics. Multiple-pronunciation models derived from training data more accurately characterize the representative acoustic information for each word than canonical definitions. This results in somewhat higher accuracy, a documented effect in the ICSI system [207].

Forced alignment and embedded training techniques were not used to optimize the training labels and lexicon for this experiment. Forced alignment and embedded training typically cause the training labels and lexicon to become closely matched and tuned to the stochastic properties learned by the neural network. The work in Chapter 5 demonstrates that this considerably improves recognition accuracy rates for Numbers. For the experiments in this chapter, however, it was judged that forced alignment and embedded training could obscure the simple relationship between the phone labels and acoustically determined syllabic onsets.

Auxiliary language and word modeling files were modified in accord with the new lexicon. Context-dependent phone durations were computed for the lexicon from the transcription material. The word grammar, derived from the word-level transcriptions of the training set, was the same as in the experiments in Section 4.4.1. A new syllable grammar was developed that matched the syllables in the lexicon.

The decoder applies a simple, empirically determined threshold to the output of the onset detection neural network to determine the possible occurrence of a syllabic onset, as

System	Error Rate
with data-derived lexicon, no onset information	9.1%
with data-derived lexicon, with onsets derived from threshold	8.2%

Table 4.3: Word-error rates for multiple-pronunciation (data-derived) decoding, with and without acoustically-derived onsets.

described in Section 4.1.2. In the illustration in Figure 4.7, the lightly filled rectangular box represents the threshold. If the value of the neural network's onset-detection output is greater than the threshold for a frame, then the decoder considers the frame to contain an onset. The decoding algorithm begins syllabic models at frames containing hypothesized onsets. The decoder also starts syllabic models in the 5 frames before a frame identified as an estimated syllable onset. That is, if the onset-detection neural network indicates an onset at frame number 17 via the threshold criterion, then the syllable-based decoder hypothesizes syllable models as beginning at frames 17, 16, 15, 14, 13 or 12. This reduced the number of potential starting frames for syllabic models by 58% in the Numbers development test set.

When the decoder incorporated acoustically-derived syllabic onset estimates into the decoding process, the recognition performance improved slightly. The word-error rate decreased by 10%, from 9.1% to 8.2%, as shown in Table 4.3. Since this improvement was produced by the addition of only syllable onset estimates, this result indicates there is potential performance benefit to be gained from this method. The syllable onset information may better encapsulate certain properties of the speech signal than phonetic probabilities.

4.5 Discussion

The process of constraining decoding with syllable onsets can be interpreted from the viewpoint of hypothesized syllabic interval units. The syllable onset estimates are hints as to where hypothesized syllable intervals begin. Decoding of phonetic information is then restricted to fitting syllable models into plausible intervals between syllable onsets. The decoding process can be thought of attaching the syllable onset hints and the most likely phonetic realizations as features to hypothesized syllabic intervals. This interpretive model, based on the framework of the syllable, is consistent with the discussion of the human speech perceptual process in [78]. This model will be discussed again in the context of the work with combining recognition systems at the syllable level in Section 6.4. The syllable interval interpretation provides an instructive connective fabric between these experiments and those of Chapters 5 and 6.

Resyllabification Phenomena

The experiments described in Section 4.4.2 illuminated certain limitations in the recognition system used for this study that necessarily impacted its performance. One such limitation

in the experimental paradigm used was the mismatch between the acoustic-phonetic and phonological representations of the syllable forms employed for word recognition. The syllabic segmentation method depended largely on acoustic-phonetic criteria, where the input was streams of phones composing a multi-word utterance. The syllabification of the lexical items used for decoding came from the phone sequence of a word in isolation. Thus, this method did not account for cross-word effects in the lexicon used for decoding. An instance where this distinction was of particular significance in word sequences was one in which the syllable coda of the first word was consonantal and the onset of the following word was vocalic, as in "five eight." The phonological representation of such a sequence would be /f ay v/ /ey t/, while the phonetic realization was more typically /f ay/ /v ey t/, where the /v/ resyllabified to the /ey t/ syllable. Such "re-syllabification" phenomena are not easily accommodated within the syllabic representational framework used in the decoder in a generalizable fashion. One possible solution, which increases the complexity of the decoding considerably, is to use multiword clustering, as described in [50]. By modeling multiple words in sequence together, alternate syllable segmentations can be modeled. Re-syllabification, however, can happen in a large number of word combinations, so the multiword set may become very large.

Extension to Larger Vocabulary Tasks

Subsequent to the main body of work in this chapter, a strategy was defined that allowed the incorporation of onset information into the input to standard decoders. This eliminated the need for a special-purpose decoder[9] and allowed for easier extension to larger vocabulary tasks. The dominant decoding technologies at ICSI, Y0 and NOWAY, both define lexicons via HMM states and accept input from neural networks in the form of one probability value per neural network output per frame. By taking an "outer product" of the phonetic neural network output and the onset estimation neural network output, and marking states at the beginnings of syllables in the lexicon as special, this modified input can be used with Y0 or NOWAY to perform recognition constrained by onset estimates.

The method essentially adds phone probabilities conditioned on whether the phone is also a syllabic onset or not to the original set of phone probabilities. This can also be thought of as altering the first state of each syllable in the HMM. Functionally, the strategy uses a phonetic neural network output stream double the original in size. Instead of 56 phones per frame, the scheme uses 112. Outputs 0-55 are the original phonetic values and outputs 56-111 represent the same 56 phonetic values, gated by whether the onset-detection neural network considers the frame as containing a syllabic onset or not. The lexicon representation needs modification only in that the initial state of the phone at the beginning of a syllable is converted from a regular phone output to a phone output conditioned on syllabic position. Thus, during recognition the decoder implicitly synchronizes the syllabic onsets in the lexicon with onsets indicated in the modified neural network output stream without modifying the internal code of the decoder.

This scheme has the potential of affecting the decoder's pruning and other optimization strategies because the decoder performs recognition in a manner for which it was not

[9]The method was developed with Philip Faerber (then at ICSI).

intended. Pruning parameters and other user-defined arguments can be used to mitigate the effect. While a special purpose, syllable-based decoder has advantages for ongoing, highly experimental work, this manipulation of the decoding input which allows the use of a standard decoder can facilitate the limited use of syllable onset information into large vocabulary tasks. The two methods are functionally equivalent, but the second involves considerably lower implementation effort since existing decoders can be used unmodified. Preliminary trials with this scheme showed that for the Numbers test set the error rate was not negatively impacted. The HMM-recombination work in Section 6.4 used a similar, but more elaborate paradigm.

We shared this strategy with Cook and Robinson, who incorporated syllable boundary information into an experimental version of their ABBOT recognition system [34] for the DARPA Hub-4 Broadcast News task [71]. Their system included a trigram language model and a 65,000-word vocabulary. Using a very similar methodology to detect syllable onsets and the aforementioned scheme for incorporating onsets into their system, Cook and Robinson found their error rate improved from 31.5% to 28.8%, a 8.6% relative reduction in word error rate [33].

4.6 Summary

Detecting syllable boundaries and nuclei has the potential to improve recognition accuracy by helping to accurately segment speech signals. Estimates of syllable onsets were used as constraints in a special-purpose decoder that explicitly represented the syllable as an intermediate stage between phones and words.

Pilot studies with "cheating" information indicated that considerable potential improvement could be achieved by accurate syllable-level segmentation. With the artificial boundaries in the cheating experiment, the system showed a 38% relative improvement over the baseline system. The pilot study also showed that the system required only modest precision from the onset detection mechanism to produce significant improvements in performance.

Further experiments used acoustic segmentation estimates derived from a signal processing method based on general principles of auditory analysis ("non-cheating"). The word-error rate was reduced by 10% for the boundary information derived from the acoustic segmentation method. We assisted Cook and Robinson in implementing these ideas for their large vocabulary system and they also found a roughly 10% improvement through the use of syllable onsets.

4.7 Conclusions

Incorporation of syllabic onset information has the potential to significantly increase the accuracy of word-level recognition. The onset information has been used in these experiments to hypothesize syllable-length intervals in the speech signal. Phones were then used for decoding on a syllable-by-syllable basis. These results with the small Numbers database indicate the potential utility of incorporating syllable boundary information in

future speech recognition systems. Although the improvement seen with the Numbers test set was slightly too small to be statistically significant at the 0.05 level,[10] experiments by Cook and Robinson showed a similar result with a larger test set for a different task which was indeed statistically significant. Furthermore, the results by Cook and Robinson used a large vocabulary, demonstrating the extensibility of these ideas.

[10]Significance testing used normal approximations to binomial distributions and used a Z-score to test whether the two distributions were significantly different.

Chapter 5

Incorporating Syllable Time Scales

The work with syllable onsets, described in Chapter 4, indicated that syllable-based information has the potential to provide meaningful improvements to speech recognition technology. This suggested that the incorporation of additional pieces of syllable-based information may be helpful as well. This chapter describes the development of an experimental speech recognition system that incorporates syllable-timed information at three different stages of the recognition process: at the feature extraction level, at the input to the neural network and in the statistical representation of the pronunciation models. In this system, selected elements of the baseline system setup were replaced by new, syllable-based elements, with a focus on the long-time span properties of speech (on the order of the length of the syllable, i.e., about 200 ms). The development of three additional systems, each incorporating some subset of these syllable-based elements, provided additional context for analysis and comparison. Each of the four experimental systems represents one of the possible combinations of a feature analysis method (RASTA-PLP or modulation spectrogram) and a recognition unit (phone or half-syllable). This chapter and the one following concentrate primarily upon the system with the maximum number of syllable-based elements, and use the other three systems for understanding more completely the effects of introducing syllable-based information. For simplicity, the system with the maximum number of syllable-based elements is referred to as the "focus" experimental system. Word error rate results showed that the more syllable-oriented experimental systems underperformed the baseline in many cases. All of the experimental systems, however, were still fairly good recognizers, as judged from the Numbers test sets.

Examination of the recognition outputs showed that the errors made by each system differed considerably from the errors made by the baseline system, particularly in the case of the focus experimental system. This suggested that combining the focus system with the baseline system may be advantageous, as discussed and demonstrated in Chapter 6. The following chapters recount investigations into how longer time intervals can affect speech recognition performance through combination with a mature, phoneme-based system.

This chapter describes the development of the baseline and the experimental, syllable-

based systems in detail. This exposition begins with a brief review of the background of the modulation spectrogram features [80, 107, 105]. Next, Section 5.2 describes the syllable-based training and recognition targets and lexicon. The system components and individual recognition performance results for the focus experimental system and each of the other experimental systems are reported for the Numbers task, for both clean and reverberant versions of the test sets in Sections 5.3 and 5.4. The chapter ends with a brief summary and conclusions.

5.1 Feature Extraction: Modulation Spectrogram

As described in Chapter 3, the baseline phoneme-oriented system for these experiments processed raw acoustic input with RASTA-PLP features. Two of the experimental systems developed, including the focus system, used modulation spectrogram features to incorporated syllable-timing at the feature extraction level. The modulation spectrogram features supplanted the RASTA-PLP features used in the baseline recognition system.

This section summarizes the work of Greenberg, Kingsbury and Morgan [80, 105, 107] on the modulation spectrogram features, as used for the work described in this thesis. Greenberg began looking towards the modulation spectrum as a means for explaining the effects of many sources of acoustic variation in speech, such as speaker differences and adverse environmental conditions. A stable representation of speech that encapsulates the most important features of speech can be an invaluable tool for the investigation of pronunciation variability.

Current speech recognition applications reduce the problems of reverberation[1] and environmental noise by gathering speech input with close-talking microphones. Ultimately, however, ideal speech applications should be accessible without the need to speak directly into a microphone. In this case, the problem of recognizing reverberant speech must be handled. Currently, as discussed in Section 3.2, speech recognition systems with high accuracy rates for clean speech make many more errors in the presence of moderate to high reverberation. Human subjects, when asked to transcribe the words in moderately reverberant speech, achieve a performance level that is vastly better than the best automatic speech recognizer, as illustrated with Numbers in Section 3.2 and in [106].

Kingsbury implemented and refined the original modulation spectrogram proposal and focused his attention on improving the accuracy of speech recognition systems in the case of reverberant speech[2] and in the presence of additive noise. Greenberg and Kingsbury observed that many unexpected variations in speech, such as speaker differences, and distortions, such as moderate levels of background noise and reverberation, that have little effect on the intelligibility of speech for humans, can dramatically affect the most popular speech representation, based on either the narrowband or wideband spectrogram. The modulation spectrogram appears to more stably represent speech by reducing the presence of parts of the speech signal that are not important in determining phonetic identity.

[1] Chapter 3 also discusses reverberation.

[2] Avendano, Tibrewala and Hermansky discuss another approach to improving recognition accuracy for reverberant speech [5].

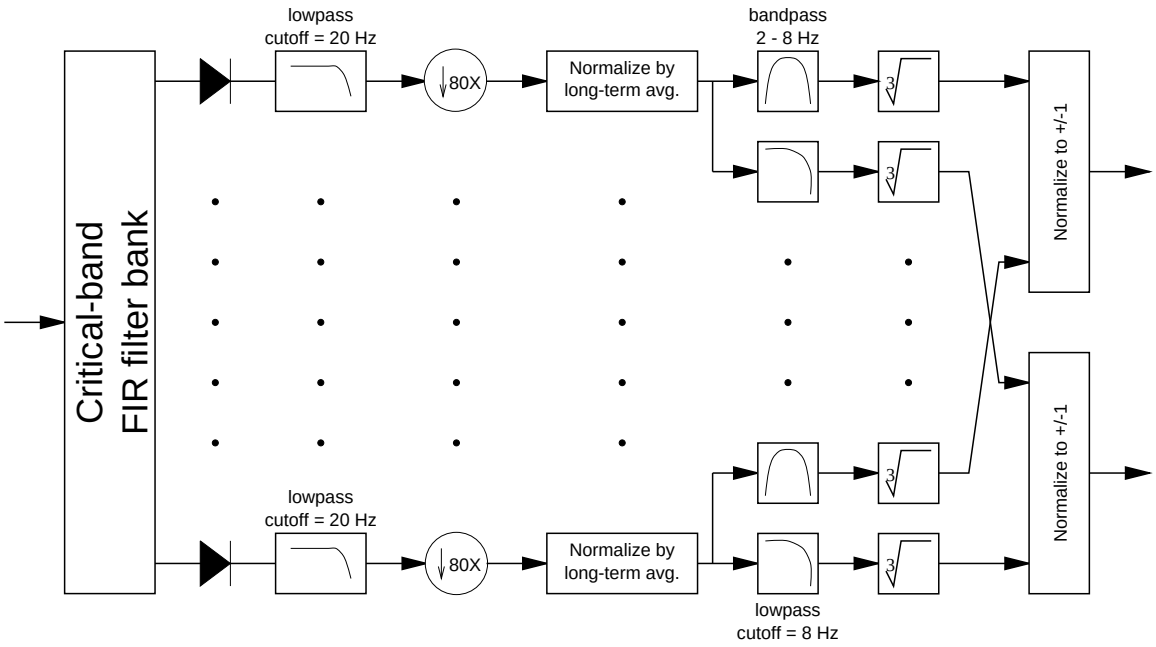

Figure 5.1: Modulation spectrogram feature extraction method [104].

The primary hypothesis behind the modulation spectrogram is that phonetic information is encoded in the speech signal as relatively slow changes in the spectral structure of speech.[3] Such a hypothesis matches the timing properties of the articulators and auditory cortical neuron activity [80]. The modulation spectrogram represents the speech signal as a distribution of slow modulations, from 0 to 8 Hz with a peak at 4 Hz, across time and frequency. The 4-Hz sensitivity corresponds roughly to syllabic frequencies. This serves as a matched filter that passes only signals that share the same modulation properties as speech. The features are computed in critical-band-wide channels in order to expand the representation of the low-frequency, high-energy portions of speech and match the characteristics of the human auditory system. The modulation spectrogram incorporates a simple automatic gain control and emphasizes spectro-temporal peaks. The specific signal processing details required to produce the modulation spectrogram features are described in [80, 105, 107] and illustrated in Figure 5.1. These steps improve the relative stability of these features in comparison to the conventional spectrogram. The signal processing results in 15 features plus 15 delta features for a total of 30 features per frame.

As a result of the signal processing steps involved, the modulation spectrogram loses some of the fine details of speech that are evident in the conventional narrowband or wideband spectrogram, such as harmonic structure and onsets. The coarse picture of the distribution of energy, however, provided by the modulation spectrogram seems to have greater stability in the presence of noise and reverberation than does the narrowband or wideband spectrogram. Greenberg *et al.* demonstrated that syllable durations coincide with energy

[3] RASTA-PLP also restricts the signal to low frequency modulations, but the filter characteristic is less extreme than in the modulation spectrogram case.

Features	Clean W. E. R.	Reverb W. E. R.
PLP	6.4%	37.6%
RASTA	6.4%	26.0%
modulation spectrogram	8.5%	27.3%
RASTA combined with PLP	5.7%	26.9%
RASTA combined with modulation spectrogram	5.5%	20.1%

Table 5.1: Partial list of performance results (word error rate) from original experiments with modulation spectrogram features [107].

distribution in the modulation spectrum [79] for conversational American English. Arai and Greenberg showed a similar pattern for Japanese [1]. Kingsbury and Morgan speculated that the performance improvement with reverberation provided by the modulation spectrogram features could be attributed to a more robust representation of syllabic segments leading to fewer deletions [106]. Modulation spectrogram features seem to emphasize the high-energy portions of the speech signal usually associated with syllabic nuclei. Kingsbury *et al.* pointed out again in [107] that most of the energy that appears in displays of the modulation spectrogram falls between onsets, and the observed energy appears to correspond roughly with syllabic nuclei.

The performance of a recognition system based on modulation spectrogram features is compared to the performance of systems based on other feature analysis methods in Table 5.1, reproduced from [107]. Kingsbury *et al.* produced these experiments with an ANN/HMM hybrid system similar to the paradigm used for this thesis work (i.e., both are based on the ICSI system). Their systems had somewhat different parameters, but used the same Numbers task for evaluation. In the experimental trials reported by Greenberg and Kingsbury [80], they found that the modulation spectrogram features performed slightly worse than a more conventional front-end method, PLP, for clean speech, but better for reverberant speech, by a statistically significant margin, as shown in Table 5.1. While the modulation spectrogram by itself did not outperform RASTA-PLP, a simple, frame-level combination of the modulation spectrogram system with the RASTA-PLP system produced statistically significant improvements on reverberant speech and a range of additive noise conditions. In this simple approach, the scaled phone likelihoods from the pair of MLPs were multiplied as they were output from the neural network. The same combination procedure applied to PLP and RASTA-PLP did not produce a similar improvement. Kingsbury *et al.* speculated that the improvement was due to the modulation spectrogram features complementing the RASTA-PLP features; the modulation spectrogram emphasized syllabic nuclei while the RASTA-PLP analysis emphasized the onsets of speech sounds. This simple approach shows that combining different features with disparate properties is a promising paradigm. The same frame-level combination approach is analyzed more thoroughly later, in Chapter 6.

Kingsbury *et al.* has continued to develop the modulation spectrogram beyond the

version used for the work in this thesis; the details about his continuing work can be found in his Ph.D. thesis [105]. It was necessary to choose a version of the modulation spectrogram features for this thesis work, however, and it was not practical to continuously update the work with revised features. The older version of the modulation spectrogram features, used in this thesis, contains many of the major features of the current work of Greenberg and Kingsbury.

5.2 Recognition Unit: Syllables

In some of the experimental systems described in this chapter, the recognition units were oriented towards the syllable by using syllable-based units as an alternative to phone units as training targets (the output of the neural network) and in the definition of word pronunciations (statistical representations of pronunciation models). The syllable-based units typically spanned one to four distinct, consecutive phones in the Numbers experiments and therefore tended to cover a larger contiguous length of the speech input then was typical for individual phones.

Each syllable was represented with 2 distinct states.[4] The syllables were divided in the middle of each syllable's nucleus. The halves are referred to in this thesis as "half-syllables." These units are not called "demisyllables" as defined by Fujimura [60, 61], though conceptually they are similar, in order to avoid the additional context and meaning carried by the term "demisyllable" in the research literature. Typically, demisyllables are formed from syllables divided just after the initial CV transition, not in the middle of the nucleus as with these half-syllables.

Half-syllables have as boundaries the syllable features most likely to be easily identified: syllable beginnings (or endings) and the syllable nucleus. Since there are many more syllables than there are phonemes in English, this scheme allowed parts of syllables to be shared between syllables, for example, the beginnings of "-ty" and "-teen" for some pronunciations. This reduced the total number of recognition units over using whole syllables.

Each syllable could have been represented with more than 2 states, each with independently derived probability densities, as in the pilot study by Schiel [174].[5] The half-syllable unit appeared to be a reasonable starting point for the investigations in this thesis. The 2-state framework minimally reflects the heterogeneous structure of the syllable; syllable onsets are generally preserved and but syllable codas are often deleted [78]. The boundary between the 2 states is initialized to be the nucleus, which usually maintains its vocalic nature through transformations.

Representing each half-syllable as a single unit is functionally similar to the representation of phones with a single unit in the baseline systems and sufficient for these experiments; these units are easily mappable into the ICSI frame-based HMM/MLP system without creating an explosion in the number of states. In the ICSI system, where each 25-ms frame is assigned a phonetically classification, a sequence of frames attributed to the same phone can be either multiple, separate instantiations each with short duration or a single instan-

[4] My thanks to Steve Greenberg for this suggestion.

[5] Florien Schiel's syllable-oriented work is discussed briefly in Chapter 2.

tiation with a long duration. Usually, but not always, the use of word models resolves this uncertainty. Representing each syllable with only a single probability density state could similarly lead to confusion as to whether the neural network was indicating repeated short outputs or one long output, for example in the utterance "one one one," but without readily available means to distinguish between the two cases. Using a minimum of 2 states per syllable is one convenient avenue for avoiding this problem and also remains consistent with the observations of syllabic structure noted above.

In this thesis the set of syllables includes only those occurring in the Numbers corpus, a tiny fraction of the syllables occurring in the English language. Fisher's automatic syllabifier TSYLB [51] (via a Tcl/Tk interface created by Dan Ellis (at ICSI)) partitioned the pronunciations in the lexicon. The process uses pronunciations defined in terms of the ICSI56 phone set, to produce a set of corresponding syllables. Another automatic process partitioned the resulting syllables into "halves," where the procedure took the "middle" of the syllable to be the durational middle of the nucleus, usually a vowel. The target labels were then either the first half or the second half of a syllable. Multiple instantiations of the same half-syllable represented minimum durations for each unit, just as in the phoneme-based system.

Multiple iterations (max of 3) of forced alignment matched the created labels to the trained system. The best iteration was, as usual, selected according to the systems' performance on the clean version of the cross-validation set. After the forced alignment process, the boundary between the two halves of a syllable may no longer be strictly the middle of the nucleus due to compounded shifts during the automatic labeling process. The placement of this boundary, however, represents the best location according to the usual global likelihood criterion.

5.3 Recognition System

This section describes the specific parameters of the automatic speech recognition systems used for these experiments. Chapter 3 reviews the general system description. The following description outlines the baseline system, the focus experimental system and the three supplemental systems more specifically. As mentioned previously, the system with the largest number of syllable-relevant attributes is the focus of this chapter and the next (hence the "focus" system). The other variant systems provide context and contrast to promote more thorough understanding. Each of these systems differed from the next only in a small way. With two different feature analysis methods, (i.e. RASTA-PLP and modulation spectrogram), and two different recognition units, (i.e. context-independent phones and half-syllables), there were four unique systems that were different from the baseline system, for a grand total of five systems:

- RASTA-PLP feature analysis with phonetic recognition units, 9 frames (nominally 105 ms) of neural network input context, the baseline system.
- RASTA-PLP feature analysis with phonetic recognition units, 17 frames (nominally 185 ms) of neural network input context.

- RASTA-PLP feature analysis with half-syllable recognition units, 17 frames of neural network input context.
- modulation spectrogram analysis with phonetic recognition units, 17 frames of neural network input context.
- modulation spectrogram analysis with half-syllable recognition units, 17 frames of neural network input context, the focus system.

The baseline system was patterned after the established systems at ICSI, used for both small and large vocabulary speech recognition tasks. This system differed from the second system in the list only in that the baseline system used a 9-frame context window. Each of the experimental systems used a 17-frame neural network context window, in keeping with the emphasis on long-time span approaches. A 17-frame span of speech is more likely to contain sizeable parts of syllables than a 9-frame segment.

With these systems, these experiments explored two test conditions, namely clean speech and speech with artificially added reverberation.[6] Since the object is to develop a general speech recognition system, not one tuned specifically toward the reverberant speech category, decisions about the recognition system components used only error rates representing clean speech performance, not data collected with the reverberant sets.

5.3.1 Experimental Procedure

The baseline system, the focus system and each supplemental experimental system variant were very similar and used the following elements:

- A 400 hidden-unit, fully-connected, single hidden-layer neural network, for frame-level probability estimation.
- A Viterbi decoder, Y0 [88]. Some experiments also used NOWAY [164, 163, 165], a stack decoder using a Viterbi criterion, for its lattice generation capability, and LATTICE2NBEST [166], for its N-best list generation function.
- A backoff bigram grammar[7] derived from the training set.
- A multiple pronunciation lexicon represented as a set of HMMs, with simple minimum duration modeling.

A set of optimization trials with the cross-validation set empirically determined the language model scaling factor (i.e., a weighting that adjusts the relative influence of the language model and the acoustics).[8]

In addition, each system had one or more of the following:

[6] Reverberation and the creation of the reverberant test sets are discussed further in Chapter 3.

[7] The grammar is described in Section 3.2.

[8] Language model scaling factors are empirically determined values that weight the influence of the language model over the acoustic information during decoding. This is discussed more fully in Chapter 3.

- RASTA-PLP features, 25-ms frame size, calculated every 10 ms. Includes 8 features, 8 delta features, 1 energy feature and 1 delta energy feature.
- Modulation spectrogram features, calculated every 10 ms. Includes 15 features, 15 delta features.
- 9 or 17 frames of MLP context.
- Phone recognition units (56 total, 31 active).
- Half-syllable recognition units (124 total, all active).

The two standard techniques for stochastically optimizing speech recognition systems, forced alignment and embedded training that includes updating the system lexicon, are appropriate for optimizing each of the experimental systems discussed in this chapter. Each of the recognition systems involved in these experiments underwent an initial training and then a maximum of three iterations of forced alignment without lexicon updating. As before, with the phone units, this served to closely match the recognition capabilities of the system with its training labels given a fixed lexicon. Recognition trials with the cross-validation set indicated the system from the best performing iteration; the selected system in each case was used for the rest of the performance figures in this chapter and the analysis and combining work in Chapter 6.

For each system, pilot studies tested the idea of optimizing the lexicon after the initial training. In these early trials, the hidden Markov models were updated once, in order to more closely match the lexicon to the learning abilities of the neural network. This independent optimization of the experimental system lexicons proved to have negligible effect on the accuracy of the resulting systems. Additionally, the resulting systems were less amenable for combination, due to mismatches in training and lexicon formulation between systems. In the interests of simplifying the experimental procedure, the lexicon-adaptation was discarded. The experiments reported used only systems without this particular optimization. Because the work in this thesis involved combining like hypotheses, over-optimization of individual lexicons was undesirable. This led to the selection of the best experimental systems without lexicon adjustment for the comparison and analysis work later in this chapter and in Chapter 6.

5.3.2 The Impact of Enlarging Hidden Layers

Because the systems varied in the number of input features and in the number of outputs, but not in the number of hidden units, the number of parameters in each system also varied. The baseline system had 77,600 neural network weights and the focus system, the largest of the experimental variants, had 253,600 parameters, as shown in Table 5.2. The table also shows the number of parameters associated with each of the other experimental systems. Keeping the number of hidden units the same (at 400) in every system reduced the number of variables affecting the training behavior and helped to keep the abilities of the hidden layer roughly the same between systems. If the largest of the variants, which had 510 input units (30 modulation spectrum features per frame over 17 frames) and 124 output units

System Description	Total Number of Parameters
RASTA + phones, 9 frames **baseline**	77,600
RASTA + phones, 17 frames	135,200
RASTA + half-syllables, 17 frames	172,000
modulation spectrogram + phones, 17 frames	216,800
modulation spectrogram + half-syllables, 17 frames **focus**	253,600

Table 5.2: Number of parameters for each of the baseline and experimental systems. Each had either 18 RASTA-PLP features or 30 modulation spectrogram features per frame.

(124 half-syllable units), were equalized in the number of parameters with the baseline by reducing the size of the hidden layer, it would have had only 137 hidden units. From early trials, indicated in Table 5.3, this appeared to be too few for this amount of data.

The role of the hidden layer in a neural network is to effect a nonlinear transformation on input data towards maximizing a discrimination measure [14]. This can be informally thought of as carving the input space with hyperplanes. The number of hidden units is related to the granularity of the pieces of the input space compartmentalized by these hyperplanes. The larger the number of hidden units, the finer the granularity. Empirically, researchers have observed that improvements in discrimination due to increasing the hidden layer size eventually asymptote. One interpretation of this effect is that the granularity eventually reaches an optimum for the given discrimination task. Adding complexity to the nonlinear mapping function does not necessarily translate to improved discrimination.

How much does a simple increase in the number of parameters affect recognition results? Experiments with several systems that differed only in the number of hidden units used, addressed this issue. No other properties in these systems were altered. Each used RASTA-PLP input with delta features, 17 frames of context and phone recognition units. Each underwent up to 3 iterations of forced alignment, in addition to the initial training, to optimize the training labels, but the lexicon was not modified. The best of the iterations for each system was chosen based on the system's performance on the clean version of the cross-validation part of the training set. These development set experimental results, listed in Table 5.3, indicated that because the training set was fairly small and highly constrained, increasing the number of parameters did not significantly affect the performance of the system on the clean version of the test set once the system contained about 100,000 parameters. The word error rate remained close to 6.5%, the asymptote, for systems with 400 hidden units or larger. This suggested that a simple increase in the number of parameters would make only a minor contribution to improving recognition performance for clean speech. Apparently, the complexity of the nonlinear mapping represented by the hidden layer reached an optimum level for the given input. For the main body of these experiments, 400 units was used as the size of the hidden layer.

As the number of hidden units increases, Table 5.3 shows a modest reduction in word error rate for the reverberant version of the development test set, from 27.6% to 24.3% going

Hidden Layer Size	Total Number of Parameters	Clean W.E.R.	Reverb W.E.R.
100	33,800	9.4%	31.7%
200	67,600	7.7%	27.6%
400	135,200	6.5%	27.6%
600	202,800	6.7%	26.8%
800	270,400	6.4%	26.0%
1000	338,000	6.4%	24.3%
2000	676,000	6.3%	26.0%

Table 5.3: Performance results (word error rates) for both the clean and the reverberant versions of the Numbers development set. Each system used about 17 frames of neural network input context and differed only in the size of the hidden layer.

System Description	Total Number of Parameters	Word Error Rate
RASTA-PLP (original)	94,672	6.4%
RASTA-PLP (doubled in size)	189,344	5.9%
modulation spectrogram (original)	99,056	8.5%
modulation spectrogram (doubled in size)	198,112	8.2%

Table 5.4: Performance results (word error rates) showing the effect of doubling the number of parameters by increasing the number of hidden units, from 488 to 976 (RASTA-PLP) and from 328 to 656 (modulation spectrogram) [107].

from a hidden layer size of 400 to 1000 units. Increased complexity in the mapping provided by the hidden layer of the neural network produced some benefit. This is not unreasonable in view of the added variation introduced by the artificial reverberation. A trial with 2000 hidden units showed that the downward trend in the word error rate does not continue indefinitely. Reverberant speech is used as an exemplar of realistically distorted speech in order to test the robustness of the system. Since the object is not to customize the speech recognition system for reverberant speech, the system parameter decisions account only for performance on the clean version of the speech data. The word error rate scores for reverberant speech are provided in Table 5.3 for comparison purposes with other experimental systems with similar numbers of parameters.

Kingsbury *et al.* report similar effects when the number of parameters in their systems were doubled on the same task [107]. Although the systems in the paper by Kingsbury *et al.* use the same ICSI methodology, their systems had 9 frames of neural network input context instead of 17 frames. The results of Kingsbury *et al.* with doubling the number of hidden units in the neural network from 488 to 976 (with RASTA-PLP features) and from

328 to 656 (with modulation spectrogram features) are shown in Table 5.4, reprinted from [107]. The performance results with double the number of hidden units are not statistically different from the original system.

As has been observed in the past by others, the technique of merely adding more parameters eventually produces diminishing returns and requires more complex training algorithms. For practical purposes, this method reaches a limit with respect to how much performance improvement can be attained and considerably increases the amount of time required for training and recognition. A better organization of data, however, has the potential to use additional parameters more effectively.

5.4 Recognition System Performance

This section describes the recognition results of each system individually on both the clean and the reverberant versions of the development and evaluation test sets. Each system was evaluated under each condition at the sentence, word, syllable and frame level. The following sections show the performance results.

To evaluate the performance of each of the systems, it is necessary to have a notion of the "right answer" or "ground truth." That is, some reasonable assignment of possible answers to questions that can be considered to be correct by some consistent and, hopefully, meaningful interpretation. As discussed by Chase [24], what the "right" answer is for the output of a recognizer or a stage of a recognizer can depend on many factors and can vary across levels of analysis. For each of the analyses discussed below, the notion of the "truth" for the relevant scoring method is discussed. Consistent application of a reasonable assignment of truth can yield a performance description that furthers the understanding of syllable-based recognition systems.

Word error rate has been and continues to be the dominant assessment criterion for ASR systems. This scoring method or an analogous procedure is universally applicable and allows comparisons between diverse applications; all speech recognition tasks consist of words or analogous tokens. The algorithm is generally simple and easily applied. Basic word scoring, however, has a number of conceptual shortcomings that limit its diagnostic value. For example, word error rate calculations treat all words equally, but this may not produce the most useful assessment in practice. Some word classes, such as nouns and verbs, are often more important for understanding than such classes as articles. Thus, this measure is of limited utility for assessing refinements that address more accurate *understanding* of speech. With the Numbers task, however, all of the vocabulary words are nouns that refer to numerical values. Deleting any single word affects the correct decoding of the utterance meaning so the word error rate measure is probably a fair metric of evaluation.

The word error rate score is typically calculated using dynamic programming to determine the minimum number of insertions, deletions and substitutions required to reconcile a recognized string with a given correct string of words. The algorithm has been standardized; the scores reported here use the SCLITE [142] scoring utility and the ICSI-local WORDSCORE [161] utility. Both are consistent with the NIST standard. The simple dynamic programming method of scoring has the disadvantage of not applying higher-level

knowledge or time-alignment information. The algorithm blindly finds insertions, deletions and substitutions without regard for the phonetic content or the temporal alignment of the errors.[9] For example, the scoring algorithm can theoretically find a large number of insertions between one word and the next, even though this situation is unlikely or is not consistent with the time alignments of the correctly recognized words.

These same scoring utilities can compute syllable error rates. In this case, canonical syllables are the basis for performance assessment, *not* the syllabary used in the recognition experiments. The syllable level generates a slightly finer granularity of analysis. For example, in the case of comparing "forty" to "fourteen," a syllable-level analysis accounts for the similarity in the first syllable and the difference in the second. Each word has a single canonical, ("dictionary") pronunciation composed of canonically defined syllables (Appendix A.3). After first replacing each word by its canonical syllable-based pronunciation, the same scoring algorithm as for words (SCLITE and WORDSCORE) calculated a syllable error rate for each system.

Using canonical syllables instead of the syllables in the recognition syllabary avoids the difficulty of dealing with mismatches between the syllables in the reference string and the recognized string where the corresponding words are actually the same. For example, if the word was "seven," and the reference string contained the syllables /s-eh/ /v-ih-n/ and the recognized string contained the syllables /s-eh/ /v-ix-n/, and both are acceptable pronunciations of "seven," the syllable string should be marked as correct. The syllable deletion effect in spoken speech can skew the accuracy scoring, but complete syllable deletion is very infrequent; usually the syllable onset, at the very least, is preserved [78]. In these experiments, syllable error rates varied moderately from word error rates, but, in general, word error rate is a good predictor of syllable error rate and vice versa.

It was possible to use a simpler procedure to compute frame-level error scores since every frame has a label and has been assigned phonetic probability estimates by the neural network. A "correct frame" is defined as one where the maximum-valued output of the neural network matches the label assigned to that frame. In this case multiple, successive frames can have the same phonetic assignment, but they are considered to be separate instances for the purposes of statistics gathering. Alternatively, the analysis method could consider segments of varying lengths of the same recognition unit as one instance. For the purposes of this thesis, however, and for the combination methods explored, the frame-level procedure was sufficient. In this case, no dynamic programming is necessary, a simple one-to-one comparison yielded the performance score. This avoids the additional effect of the decoding procedure and removes any time alignment problems between different tokens, but does introduce a strong dependence on correct labels. The procedure takes the labels generated by the fully trained baseline system using forced alignment to be the "correct" labels of the test set. When the experimental systems incorporate forced alignment, the ground truth labels become somewhat mismatched, so the frame-level error rate reported may be slightly inflated. In most cases, however, the best systems selected were the ones without additional forced alignment, so the "ground truth" labels are adequate for evaluation.

[9] Lin Chase suggests some solutions to these issues [23, 24].

	Frame-level	Syllable-level	Word-level	Utterance-level
Number of Tokens	230,000	5703	4673	1206

Table 5.5: Number of recognition tokens at each level for the Numbers *development* test set. The number of frames is approximate, since this can change depending on the feature extraction method and context window size.

The scoring utilities also provide an utterance-level score, where an utterance is correct only if the recognized words exactly match the reference string. The programs calculate the utterance level error scores by matching recognized sequences of words with the correct sequence. Alternatively, the utterance-level error scores can be calculated using the sequences of syllables. For Numbers, since the vocabulary is highly constrained, these two methods yield the same result; however, for large vocabulary tasks, the methods may generate disparate numbers. At such a large granularity of analysis, detailed effects are lost. The analyses in Chapter 6 showed, however, that considering the outputs of recognition systems at the level of the entire utterance still has some utility, particularly for the combination of systems.

5.4.1 Clean Speech

"Clean" speech in the case of Numbers is a misnomer; the original acoustic data were collected over the telephone, so the acoustic signal includes a variety of line and background noises. For example, one utterance has the sound of a wailing baby in the background. For the purposes of this thesis, "clean" refers to this relatively pristine version of the recorded speech.

Table 5.6 shows the word error rates of each of the selected five systems on the original task's development test set. The total number of recognition tokens in each category is shown in Table 5.5. In addition, Table 5.6 also shows frame error rates, syllable error rates and whole-utterance error rates. The error rates between each column are strongly correlated and are reported to provide context for the comparisons in the next chapter. For direct comparison with the experimental results in Chapter 6, Table 5.7 shows word error rate scores on the evaluation test set. Tuning of any parameters thus far was performed using only the cross-validation portion of the training set, not the development test set. The evaluation test set was reserved, according to usual practices, until all system design decisions were fixed and was not used for empirical determination of any parameters.

As can be seen from Table 5.6, the baseline word error rate for the clean speech version of the Numbers development test set was 6.8%. The experimental system variants had word error rates ranging from 6.9% to 10.6%. The word error rates show that the experimental system variants incorporating modulation spectrogram features and/or the half-syllable unit, including the focus system, clearly performed worse than the more established forms using only RASTA-PLP and phoneme-based recognition units, though all the recognition

System Description	Frame-level	Syllable-level	Word-level	Utterance-level
RASTA + phones, 9 frames **baseline**	14.9%	6.6%	6.8%	20.1%
RASTA + phones, 17 frames	15.7%	6.5%	6.9%	19.7%
RASTA + half-syllables, 17 frames	23.9% †	8.2%	8.3%	23.1%
modulation spectrogram + phones, 17 frames	20.1%	8.6%	9.2%	26.1%
modulation spectrogram + half-syllable, 17 frames **focus**	28.3% †	10.1%	10.6%	27.9%

Table 5.6: Performance results (error rates) for the baseline and experimental systems on the complete, clean Numbers *development* test set. Frame-level error scores labeled with a † are not directly comparable to the other values in the same column, due to a difference in recognition unit.

systems achieved a reasonably high level of accuracy. At a word error rate of 10.6%, the focus system was the least accurate.

The influence of the decoding (dynamic programming) step in recognition accuracy can be seen in that the frame-error scores in Table 5.6 are categorically larger than the syllable- or word-level error scores. The process of constructing words and sentences from phones smoothes the errors that occur at the phone level. The syllable-level scores are essentially the same as the word level scores, an indication that with Numbers, there is a strong correlation between correct syllable recognition and correct word recognition. Since the Numbers vocabulary is highly restricted, this is not surprising. A large vocabulary task may produce a somewhat different relationship. The larger error values at the utterance level indicate that word and syllable errors are not concentrated in particular sentences, but are rather dispersed. That is, the word errors are not the result of a few very difficult sentences.

A more detailed consideration of Table 5.6 reveals that the first and second rows are roughly the same at each level. This means that the baseline system and the experimental variant most similar to the baseline produced approximately the same performance. Therefore, extending the length of the neural network window did not affect overall word error rate for clean Numbers speech.

The RASTA-PLP system with half-syllable recognition units achieved a comparable performance to the modulation spectrogram system with phoneme-based units when measured at the syllable or word level. Using half-syllables as the recognition unit had approximately the same degradation in performance as using the modulation spectrogram. Combining the modulation spectrogram with half-syllable units caused significant further degradation in

System Description	Word Error Rate
RASTA + phones, 9 frames **baseline**	6.7%
RASTA + phones, 17 frames	6.6%
RASTA + half-syllables, 17 frames	8.1%
modulation spectrogram + phones, 17 frames	8.6%
modulation spectrogram + half-syllables, 17 frames **focus**	10.0%

Table 5.7: Performance results (word error rates) for each system for the complete, clean Numbers *evaluation* test set.

performance.

Since the RASTA-PLP system with half-syllable recognition units had a higher frame error rate, perhaps the half-syllable units were more easily confused by the neural network. Examination of the half-syllable units showed that a number of the automatically created units have confusable characteristics. The multiple pronunciation lexicon partially compensated for some confusions. The lexicon included variants that are very similar, often differing only in the identity of the vocalic segment. Therefore, differing frame-level estimates could map to the same, correct word. Since there are 124 half-syllable units as opposed to the 31 phone units, there were fewer training patterns for each half-syllable unit than for each phone unit. The half-syllable unit also spans a longer contiguous segment of speech and therefore maybe subject to increased variability in training patterns compared to the shorter phones, making training more difficult.

Coarser units and features led to somewhat lower accuracy rates on clean speech by themselves. Frame-level evaluations suggest that using the modulation spectrogram features led to a lower frame-level accuracy than using RASTA-PLP features, possibly due to increased confusions of blurred featural details.

The performance results from the reserved evaluation set, Table 5.7, reflect the same word-level performance characteristics observed with the development set. This independent validation using the evaluation set is repeated throughout this series of experiments.

The focus system had multiple sources of long-time span smearing, so it is reasonable to expect that the focus system has the worst error rate. The comparison here may not be entirely fair, since the lexicon and phonetic labels used as a starting point for training were optimized for the baseline, a well-established system. It is possible that additional optimization of the experimental systems may help close the gap in error rates. As mentioned previously, the ultimate goal of combining systems limits the utility of individually optimizing the experimental systems. In the next chapter, experiments will determine if long-time span information can help in combination with finer grain features and units.

System Description	Frame-level	Syllable-level	Word-level	Utterance-level
RASTA + phones, 9 frames **baseline**	43.7%	28.6%	29.3%	65.5%
RASTA + phones, 17 frames	39.9%	25.5%	25.6%	60.8%
RASTA + half-syllables, 17 frames	47.8% †	31.3%	30.5%	66.6%
modulation spectrogram + phones, 17 frames	38.8%	27.0%	26.6%	60.2%
modulation spectrogram + half-syllables, 17 frames **focus**	45.7% †	30.0%	30.1%	65.0%

Table 5.8: Performance results (error rates) of each system for the complete, reverberant version of the Numbers *development* test set. Frame-level error scores labeled with a † are not directly comparable to the other values in the column, due to a difference in recognition unit.

5.4.2 Reverberant Speech

The tables in this section show the error rates of each of the systems on the Numbers development test set where the speech was artificially made reverberant with a 0.5-second reverberation time.[10] The number of recognition tokens in each category is the same as in Table 5.5.

The relationships between the columns of Table 5.8 reflect those observed with the clean speech case. As was found by Kingsbury *et al.* [107], the modulation spectrogram system with phoneme-based recognition units performed almost as well as a comparable RASTA-PLP system, also with phoneme-based recognition units. Both resulted in a word error rate of around 26%. Using a larger input window to the neural network longer acoustic context) always seemed to help. All other systems showed some amount of degradation in performance compared to the best system, with the focus system once again producing the worst error rate of 30.1%. The difference in accuracy from the best to worst system, however, is within 5% absolute. Error rates at several stages of consideration for each system are given in Table 5.8 for the development test set.

Word error rates are given in Table 5.9 for the reserved evaluation test set, again to validate the word-level trends observed in the development set.

[10] Reverberation and the creation of the reverberant test set is discussed in more detail in Chapter 3.

System Description	Word Error Rate
RASTA + phones, 9 frames **baseline**	28.0%
RASTA + phones, 17 frames	25.1%
RASTA + half-syllables, 17 frames	30.9%
modulation spectrogram + phones, 17 frames	25.8%
modulation spectrogram + half-syllables, 17 frames **focus**	30.1%

Table 5.9: Performance results (word error rates) of each system for the complete, reverberant version of the Numbers *evaluation* test set.

5.5 Summary

These experiments explored the effects of substituting syllable-based processing elements into the established baseline speech recognition system. A series of experiments with a focus experimental system and three supplemental systems using the Numbers database showed that a wider context window (17 frames vs. 9 frames) had no effect on clean speech, but improved speech recognition accuracy moderately for reverberant speech. Using modulation spectrogram instead of RASTA-PLP features resulted in significant degradation in the clean speech case and did not help in the reverberant speech case.[11] Using half-syllable targets caused significant degradation in both clean and reverberant speech. The individual performance of the focus system was the worst overall for both test sets.

5.6 Conclusions

It is not surprising that the baseline and the experimental system most similar to the baseline should outperform the other experimental systems for clean speech. The baseline system was a mature system, the result of optimization and improvement efforts over several years. A comparable effort for the focus system and each of the three supplemental system variants was not possible; independently optimizing the syllable-oriented systems could potentially confound the combination experiments, as discussed in the next chapter. Some of the combination methods required the two constituent systems to have closely coordinated recognition behavior. While syllable-based elements did not seem to help much, the syllable-oriented systems were still reasonably accurate; the difference between the best and the worst performance was less than 5%, absolute. Chapter 6 discusses how the coarser, syllable-oriented focus system can be combined with the finer-grain phone-oriented baseline system to produce a lower error rate overall.

[11] Kingsbury has recently refined the features further; they now produce an improvement for the reverberant case. Future work will incorporate this revision.

Chapter 6

Combining Systems

The results described in Chapter 5 showed that substituting syllable-based elements into the baseline system could increase word-error relative to the phoneme-based system, though the resulting systems were still fairly accurate. Studying the recognition outputs revealed that syllable-oriented systems tended to make different errors, complementary to the phoneme-oriented baseline. The introduction of syllable-based information promoted a divergence in error patterns. This effect was most marked with the focus system, the system variant that incorporated syllable-based time-span elements at the feature extraction level, in the context window of the neural network and at the recognition unit level. The focus and baseline systems both had reasonably good performance and a low degree of error correlation. These observations suggested that a combination of the two systems may produce a system more accurate than either of the constituents alone if a suitable combination method could be found.

This chapter first describes the error analysis method used and discusses one case study in detail– the analysis of word errors. Section 6.2 discusses the combination methods used for integrating the outputs of the experimental systems with the baseline system at three stages of the recognition process. The following three sections discuss the analysis and combining results at each of the stages: at the frame level (at the output of the neural network), at the syllable level, and at the whole-utterance level. The analysis of the experimental results pointed out the different strengths and weaknesses of the focus and the baseline systems. In all cases combining systems improved the overall recognition performance moderately for clean speech, and more substantially for reverberant speech. The chapter concludes with a summary and discussion of the results.

6.1 Analysis

This chapter focuses not on individual performance, but rather on finding complementary attributes of experimental systems that can be used to improve speech recognition performance through a combination scheme. For this objective, comparative error analysis can quantify some of the differences in behavior between systems.

6.1.1 Background and Methods

As discussed in Chapter 5, there are many ways to assess speech recognition performance, and to analyze and compare different recognition systems. Although word error rate is the most universally applied criterion, it has its limitations. It does not convey a complete picture of performance and is not enlightening for comparing two systems with similar error rates. A simple word error rate metric does not reveal deviations in behavioral properties from the baseline system. More specific analysis and comparison methods that can expose the differences in performance characteristics among systems are more helpful. At each level of consideration (i.e., frame, syllable and whole-utterance) there are many possible analysis methods. The primary tool used in this chapter is a coarse error analysis method that has roughly analogous procedures at each level. By using similar procedures, trends that appear across recognition stages can be made more apparent. At some stages, additional evaluation was helpful, as discussed in the later sections of this chapter.

The work of Farrell, Ramachandran and Mannone on combining systems in the context of speaker verification [49] inspired the error analysis method described in this chapter. In their paper, Farrell *et al.* evaluated four commonly-used models for the speaker verification task and three ways to combine the four scores. They calculated the error correlation among model outputs and used this measure to select models for their combination experiments. Their experimental trials supported their hypotheses about the best combination pair, based on the error correlation information.

Speech recognition has outputs that are considerably more complex than those of speaker verification. Typically the speaker verification task has only "accept" and "reject" outputs for a given input, possibly along with an estimate of confidence. Therefore, some adaptation of the above general idea is necessary. The adopted analysis method compares the baseline speech recognition system with the focus system and with the other variants, one pairing at a time. For each pair, the method considers only those outputs for which at least one of the systems makes an error, since outputs that are correctly identified by both systems do not bear on the error analysis. The analysis procedure uses a simple characterization of errors that does not distinguish degrees of error. Hence, for each output, only four possible outcomes are measured in the analysis:

1. The baseline system was correct and the experimental system variant was wrong. ("Baseline Only Correct")
2. The baseline system was wrong and the experimental system variant was correct. ("Variant Only Correct")
3. **Both systems were wrong and reported the same erroneous value. ("Identical-Incorrect")**
4. Both systems were wrong, but reported different erroneous values. ("Different-Incorrect")

The goal of this analysis procedure is to represent, as simply as possible, the coarse differences in system behavior that can affect combination strategies. Word error rate is

not sufficiently informative for this. The error analysis procedure outlined above contains considerably more information than word error rate; nonetheless it is only one way to describe these details. The possibility that some types of errors are more egregious than others is ignored in favor of simplifying the analysis as much as possible. For example, mistaking "four" for "forty" might be regarded as less of an error than mistaking "four" for "hundred," but both errors are equivalent in this analysis.

Of these four different categories of comparison results, the category that contributes the least straightforwardly toward a better combined error rate is (3) Identical-Incorrect errors. It is very tricky for combination algorithms to correct errors of this type. Most combination schemes enhance accuracy by balancing the outputs of different systems against each other. If the different systems both agree on the best answer and that answer is wrong, designing an algorithm that can both detect and correct the error is a much more formidable problem. One possible method is to use local accuracy estimates [206] or measures of confidence [204] for both systems. The reliability of such methods, however, can degrade when the recognition system encounters speech that is markedly divergent from the training samples.

In these experiments, the differences between systems are distilled down to the proportion of exactly alike errors in the respective recognition outputs. For ease of discussion, this percentage value is referred to as the "Identical-Incorrect" measure; these percentages are highlighted in boldface type in the relevant tables of this chapter. The system pairs with the greatest potential for increased accuracy through simple combination methods are likely to have good individual performance and a small number of Identical-Incorrect errors. This measure is one way to characterize the behavioral differences between systems as it affects combination strategies. It can indicate, in a coarse way, how "different" a system variant is from the baseline.

At the frame-level, each system generates an output for each 25-ms frame of input. This one-to-one correspondence makes the comparison between systems simpler. Two of the experimental recognition systems, however, have output values based on half-syllables which can not be compared directly (at the frame level) to the phoneme-oriented baseline system.

In analysis at the syllable or word level, recognized values can potentially encompass variable-length, unaligned sections of speech. In this case the analysis procedure first compares the output word or syllable strings to the correct reference string and then compares each of the identified errors on a one-to-one basis. For example, imagine that for the utterance "one two three four five," System 1 recognized "one oh six nine five" and System 2 reported "two six oh five," as illustrated below:

	one	*two*	*three*	*four*	*five*	(correct string)				
System 1:	one	oh	six	nine	five	[cor]	[sub]	[sub]	[sub]	[cor]
System 2:	—	two	six	oh	five	[del]	[cor]	[sub]	[sub]	[cor]

The scoring programs first generate an assessment of the recognition output using dynamic

programming. These programs output words correct ([cor]), deleted ([del]), inserted ([ins]) and substituted ([sub]). The System 1 error evaluation generated by the scoring would be "[cor] [sub] [sub] [sub] [cor]" and System 2 would have "[del] [cor] [sub] [sub] [cor]," as illustrated above. After discarding words that were correctly recognized by both systems, the analysis procedure then considers the differences between the words corresponding to "[cor] [sub] [sub] [sub]" and the words corresponding to "[del] [cor] [sub] [sub]," marked by boxes in the example. The procedure compares the recognized strings "one oh six nine" and "[del] two six oh" by pairs of words. For exposition purposes, each pairing of words between the recognized strings, filtered in this way, is referred to as an "error token." In the example, there are four such error tokens, outlined by boxes.

Using dynamic programming to match the recognized word strings with the correct sequence avoids the problem of comparing two streams of words in which correctly recognized words are offset temporally, a situation that should not affect the error analysis. Each system included some independent training so it is reasonable for the beginning and end times of the recognized words to shift from system to system. Not using the exact temporal values for the beginning and end points of each word exposes the possibility that, due to the vagaries of the scoring algorithm, two words are compared that do not actually share the same acoustic segment. This simple analysis method, however, proves useful for revealing some general trends, despite this potential drawback. For the frame-level and whole-utterance-level analyses there is no time-alignment discrepancy problem between compared recognition outputs. The trends at these two levels validate the similar trends observed at the syllable and word level.

The error analysis described should be viewed as providing comparative information about two systems, rather than predictive data about the combining process. The simple analysis method does not account for specific combination strategies so it can not be used to derive concrete performance expectations. Also, as mentioned previously, the method does not incorporate the magnitude of error, which can be a factor in the behavior of the combination method. This analysis is useful for distilling comparison information into a small number of values that can be used to guide experiments in combination. The analysis charts show generalized characterizations; large discrepancies in values between recognition systems are meaningful.

The charts can be used to design new combination strategies that capitalize on the trends displayed. For example, if a chart shows that between two systems, at least one is always right (that is, there are no instances where both are wrong), the system designer should consider concentrating on recognizer selection combination schemes. Such strategies are typically based on confidence values and assessments of the relative correctness between two systems. On the other hand, if the chart shows that two systems always recognize the same tokens erroneously (that is, there are no instances where one system recognizes the token correctly and the other does not), then the system designer would be better rewarded by using a combination method that melds recognition results and ameliorates wrong answers. The chart coarsely summarizes the potential for improvement through combination, if selection and fusion between two recognizers could be performed perfectly.

Variant (paired with Baseline)	Clean		Reverberant	
	Number of Error Tokens	Percentage of Total Tokens	Number of Error Tokens	Percentage of Total Tokens
RASTA + phones, 17 frames	422	8.8%	1,646	33.4%
RASTA + half-syllables, 17 frames	527	11.0%	1,898	38.4%
modulation spectrogram + phones, 17 frames	578	12.0%	1,965	39.2%
modulation spectrogram + half-syllables, 17 frames **focus**	652	13.6%	2,125	42.6%

Table 6.1: Number of tokens contributing to the error analysis along with the percentage of total tokens that these error analysis words represented in the *clean* and *reverberant* speech versions of the Numbers development test set.

6.1.2 A Case Study

Data generated by this procedure at the word level are shown in Tables 6.1, 6.2, and 6.3. These tables illustrate the error analysis method in more detail. Table 6.1 shows how many word tokens the analysis used and what fraction of the total number of word tokens these error tokens represented. As expected, a much larger number of word tokens contributed to the error analysis under reverberant conditions than under clean conditions. Also, each comparison between an experimental system and the baseline resulted in a different number of word error tokens. Nevertheless, the number of error tokens generated by each pairing is roughly similar, particularly for reverberant speech.

Table 6.2 displays the distribution of error tokens between the four error categories for the clean version of the Numbers development test set. The percentages corresponding to the Identical-Incorrect measure outlined previously are shown in boldface type. The word-level exhibits several trends that are also reflected in the frame-, syllable- and utterance-level analyses. These are examined in detail below:

Clean Speech: Relative Accuracies

The first two analysis columns in Table 6.2 illustrate the relative accuracy of each system individually by showing how often one system was correct and the other in error. This information is most relevant for recognizer combination by selection methods. In any pairing the baseline tended to be more correct than the experimental variant system. This reflects the relatively lower word error rate of the baseline compared to the word error rates of the experimental variants. The variant in each pairing was correct where the baseline was in

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	23.9%	22.7%	**43.6%** (184)	9.7%
RASTA + half-syllables, 17 frames	39.1%	26.0%	**24.9%** (131)	10.1%
modulation spectrogram + phones, 17 frames	44.5%	25.1%	**21.5%** (124)	9.0%
modulation spectrogram + half-syllables, 17 frames **focus**	50.8%	23.5%	**14.1%** (92)	11.7%

Table 6.2: Distribution of error tokens across the four analysis categories in the *clean* speech version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

error for about a quarter of the error tokens.

The similar values in the top row of the first two columns correspond to the comparable accuracy levels of the two systems. The considerable disparity in the values of the same two columns in the bottom row demonstrates the superior accuracy of the baseline system over the focus system. The similarity in values for these first two columns between rows 2 and 3 suggests that the use of the half-syllable unit results in a degradation comparable to the use of the modulation spectrogram features. The degradation increases when the two are used together, as in the focus system.

Clean Speech: Relative Differences

The last two analysis columns provide crude information about the errors made when both systems recognize the input incorrectly. Such information is useful for recognizer fusion strategies.

The data show that when the baseline and the experimental variant were both wrong, the focus system (i.e., the system with the most syllable-based elements) was the least correlated with the baseline. The Identical-Incorrect value between the baseline and the focus system (14.1%) was smaller than in any other pairing. That is, the percentage of tokens where both the baseline system and the focus system recognized the same erroneous values was smaller than the corresponding percentage in any of the other pairs.

Syllable-based elements helped produce more complementary behavior. The large proportion of Identical-Incorrect errors and the relatively small number of Different-Incorrect errors in the top row suggests that the behavior of the two systems is similar– the two systems in the first pairing tended to make the same errors. In contrast, the relatively smaller

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	16.8%	27.3%	**38.9%** (640)	17.0%
RASTA + half-syllables, 17 frames	27.8%	24.8%	**25.7%** (488)	21.7%
modulation spectrogram + phones, 17 frames	30.3%	36.6%	**13.6%** (267)	19.5%
modulation spectrogram + half-syllables, 17 frames **focus**	35.5%	33.6%	**11.0%** (234)	19.9%

Table 6.3: Distribution of error tokens across the four analysis categories in the *reverberant* speech version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

number of Identical-Incorrect errors for the comparison of the baseline with the focus system suggests that these two systems acted differently. Again, the similar values shown in rows 2 and 3 of these two columns suggest that the half-syllable unit and the modulation spectrogram unit had comparable influence on the recognition process. The use of the two in tandem, in the focus system, created a larger effect. This suggests that the half-syllable unit affects recognition in a distinct, complementary manner when compared to the signal processing features.

Reverberant Speech: Relative Accuracies

The reverberant version of this analysis shows many of the same trends.

In the first two columns of Table 6.3, the baseline was found to be more often in error than two of the variant systems: the RASTA-PLP system with phoneme-based units and a 17-frame input window (row 1), and the modulation spectrogram system with phoneme-based units and 17-frame input window (row 3). In other baseline/variant comparisons, the baseline was more correct, though the gap between the values in these columns is considerably reduced compared to the clean case. This suggests that using the 17-frame input window was an advantage for recognizing reverberant speech, along with using the modulation spectrographic features.

Reverberant Speech: Relative Differences

Examining the last two columns in Table 6.3 shows that the modulation spectrogram alone had a considerable effect in causing errors to diverge between systems. Moving from clean to reverberant conditions widened the difference between the information provided by the

modulation spectrogram and that derived from RASTA-PLP. Using the half-syllable units had a smaller, but still noticeable effect. As with clean speech, when both systems were in error, the focus system (the most syllable-oriented of the variants) was the least correlated with the baseline, as reflected in the 11.0% Identical-Incorrect measure.

Trends

Subsequent sections of this chapter continue this analysis and comparison of recognition systems more briefly at the frame, syllable and at the whole utterance level. Some commonalities appear at every level in the systems tested:

- The total number of error tokens is much greater for reverberant speech than for clean speech.
- Systems with RASTA-PLP features and phone recognition units are more often correct than the other experimental variants for clean Numbers speech.
- Systems with syllable-based features and recognition units perform less poorly relative to the baseline on reverberant speech than on clean speech.
- The 17-frame context window for the neural network improved the recognition of reverberant Numbers speech.
- The use of the modulation spectrogram and the half-syllable units minimizes the number of Identical-Incorrect errors.

As mentioned previously, informal inspection of recognition outputs suggested a concentration on combining the baseline system with the modulation spectrogram system including half-syllable units. The more detailed error analysis described above supports this decision by demonstrating that this pairing minimized the number of word errors in which both systems made the same error (the Identical-Incorrect measure). The analysis sections of the rest of this chapter more briefly describe results that reflect trends similar to those observed at the word-level. These findings support the hypothesis that there are benefits in combining the baseline system with syllable-based systems.

6.2 Combining: Background and Methods

Many researchers have espoused the idea that the human speech perception system integrates information over several levels. Greenberg suggests that human speech recognition relies on temporal dynamics in coarse spectral patterns [74]. For efficient communication, human beings rely on the use of multiple, redundant, coarse patterns to obtain the robustness to noise and other nonlinguistic sources of variability [75]. The human brain may employ a rather sparse representation that exhibits most of the temporal dynamics of the speech signal. Todd incorporates the temporal, or rhythmic nature of auditory processing via what he refers to as dynamic spatio-temporal receptive fields [189, 190]. Todd and Lee

also discuss the combination of simultaneous information into some sort of multimodal, multi-scale sensory representation in the human brain. Todd postulates the existence of neurons that combine primary inputs into higher level features, possibly via a cascade of secondary receptive fields that process information from primary units. Although the combining methods described in this thesis bear only a passing similarity to the physiologically-motivated perspectives of Greenberg and of Todd, such perspectives from human auditory processing support the basic idea of combining multiple sources of information for recognition.

Inspection of the word recognition outputs of the baseline system and the focus system suggested that if a recognizer could dynamically select the best system for each input then a higher performance could be attained overall. This "dynamic recognizer selection" requires some sort of numerical evaluation of the relative accuracy of the systems for a given input. As discussed in Section 3.5, such evaluation is currently regarded as difficult. Some obvious methods include using confidence measures based on likelihoods, posteriors or lattice densities, for example. But some preliminary trials along these lines for the Numbers corpus were not successful in choosing the better outputs based on simple calculations. Other methods, not addressed here, might involve training neural networks to perform the selection (e.g., mixtures of experts).

The dual to recognizer selection is "recognizer fusion," the merging of the outputs of multiple systems. Combining the outputs of multiple neural networks is an open research topic that this thesis does not fully address. Included among the many possible techniques is the neural network boosting algorithm AdaBoost [57][1] and parallel consensual neural networks [13]. Because ASR includes a crucial decoding step subsequent to the pattern classification stage, there is an added level of complexity when considering the combination of neural network outputs. Combining methods for multiple recognition streams are discussed more generally in Section 3.5.

In this chapter the combining method is a simple multiplication of probabilities, that is, an unweighted linear combination of log probabilities (effectively an average). Greenberg and Kingsbury first demonstrated the value of this method for a phone-based recognizer employing modulation spectrogram features combined at the frame level with a RASTA-PLP system [80]. The combining experiments summarized below replicate the original findings, and extend the strategy to larger granularity combinations with the focus system (such as at the syllable and whole-utterance level).

Combining two recognition systems increases the total number of parameters involved. This complicates the comparison between the combined system and the performance of the individual, constituent systems. Nevertheless, as mentioned in Section 5.3.1, merely increasing the number of parameters beyond the default 400 hidden units of the baseline system did not increase the system's accuracy for the clean version of Numbers. Even for the reverberant speech test set, enlarging the size of the hidden layer of the neural network only moderately improved performance. The combined systems described in this chapter did not exceed the parameter count of the 1000-hidden-unit neural network system in Table 5.3. Therefore, the further performance improvements in the combined systems must result from

[1] Schwenk has implemented a version of AdaBoost for the Numbers corpus [179].

the additional structuring of the parameters.

Combining multiple recognition streams exhibited performance advantages over the individual constituent systems at each of the three stages of decoding. Each level, however, (i.e. at the frame, syllable and whole-utterance levels) has separate interpretations and implementation properties. The remainder of this chapter describes in detail these issues and the results of the analysis and combination.

6.3 The Phoneme/Frame Level

This section considers the combination of recognition output streams at the frame level, that is, at the output of the neural network before the decoding process. Since frame level integration entails combining similar outputs in a one-to-one manner, dissimilar output units can not be directly combined. The subsequent decoding essentially uses probability estimates from each stream in a lockstep manner. For example, in each time frame, the probability of an /ah/ from System 1 is combined with the probability of an /ah/ from System 2 and nothing else. This constraint ensures that both streams are decoded to be in the same HMM state at the same time. Because the focus system uses different HMM states (half-syllables) than the baseline (phones), this pairing can not be combined at the frame level. For this reason, the variant system used for combining with the baseline was the system with modulation spectrographic features and phone recognition units. The frame level is relatively advantageous in that it may be easier to isolate short-time trends, such as patterns in phone identification. Since this analysis reflects information prior to the decoding stage, however, these findings can be somewhat decoupled from the final word error rate.

6.3.1 Analysis

This section reports on the analysis of the raw neural network outputs in each system. As mentioned in Section 3.3.3, a forced alignment process used the baseline system (400 hidden units, phone recognition units, 9 frames of neural network context) to generate "correct" frame-level phone labels for the development test set, given knowledge of the true word transcriptions. These analyses used this labeling as the "ground truth," as discussed previously in Section 5.4. Because the experimental variants included systems which were further optimized using forced alignment, comparing the frame-level output of these systems with the baseline may overstate the number of mismatches.[2] These analyses compare only systems with the same recognition unit; it is not clear how to compare systems with different targets, i.e., a phoneme-based system and one based on the half-syllable model. Even with these caveats, the analyses below are in keeping with the general trends observed; the baseline and the system variants had roughly similar percentage accuracies, but tended to make different errors.

Modulation spectrographic features may help systems recognize certain sounds more accurately than RASTA-PLP. Evaluating the frame-correct values according to training

[2]The issue of forced alignment is discussed in more detail in Section 5.4.

Variant (paired with Baseline)	Clean		Reverberant	
	Number of Error Frames	Percentage of Total Frames	Number of Error Frames	Percentage of Total Frames
RASTA + phones, 17 frames	47,286	22.9%	105,988	51.2%
modulation spectrogram + phones, 17 frames	56,073	29.9%	114,375	54.9%

Table 6.4: Number of frames which contributed to the error analysis and the percentage of total frames these error analysis frames represented in the *clean* and *reverberant* speech versions of the Numbers development test set.

target type produced inconclusive results; the modulation spectrogram system did not produce clear patterns of performing well on some sorts of targets and poorly on others. A comparison of the two variants that differed only in their input features– RASTA-PLP or modulation spectrogram– while keeping the phoneme-based targets and 17-frame context window the same, showed that the modulation spectrogram system consistently underperformed the RASTA-PLP system for clean speech. At best, the modulation spectrogram system equaled the performance of the RASTA-PLP system for some targets. For reverberant speech, however, the modulation spectrogram system significantly outperformed the RASTA-PLP system for 13 out of the 31 phones used. Inspection of the values revealed no obvious pattern among these 13. The Numbers task may be too limited to illuminate phonetic trends.

The error correlation analysis method helped characterize the differences between systems. In a procedure analogous to the word-based analysis described above, the phoneme-based variant systems were paired with the baseline (having the same recognition unit type). Frames where both systems generated the correct value were discarded. This left frames where one system was correct and the other was not, or where both systems were incorrect. Table 6.4 shows the number of frames, out of approximately 210,000 in the development set, remaining after this pruning process for both the clean and the reverberant versions of the Numbers development test set.

Tables 6.5 and 6.6 show the percentage of the error frames in each of the 4 categories defined above for the clean and the reverberant versions of the development test set. Many of the trends observed in the word-level case study (Section 6.1) appear in this frame-level version.

For the clean case the baseline was again correct more often than the experimental variants. For reverberant speech, the situation was reversed and the variant systems were more often correct, indicating that under reverberant conditions the modulation spectrogram and the wide neural network input window were beneficial to recognition performance. The Identical-Incorrect columns of Tables 6.5 and 6.6, highlighted in boldface, show that the

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	34.6%	31.3%	**20.9%** (9,883)	13.3%
modulation spectrogram + phones, 17 frames	44.8%	25.3%	**14.8%** (8,299)	15.2%

Table 6.5: Distribution of error frames across the four analysis categories in the *clean* version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	14.7%	22.2%	**38.2%** (40,487)	24.9%
modulation spectrogram + phones, 17 frames	20.8%	29.4%	**19.1%** (21,846)	34.9%

Table 6.6: Distribution of error frames across the four analysis categories in the *reverberant* version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

modulation spectrogram features produced a divergence in the recognition behavior. The features apparently introduced a variation in the errors produced by the neural networks, especially with reverberant speech, causing a decrease in the correlation between errors committed by each system.

The results of these analyses show that the system variants achieved a fairly good overall percentage correct at the frame level, but were correct for a somewhat different set of frames from the baseline system, more notably in the presence of reverberation. These analyses support the general conclusion that combining the systems at the frame level may capitalize on the individual system strengths while diluting weaknesses, resulting in an improved error rate overall.

6.3.2 Combining

The combining method for merging two recognition systems at the frame level involved multiplying the probabilities as output from each system's neural network. The Y0 decoder [88] (no modification necessary), used these probabilities as input and produced words and

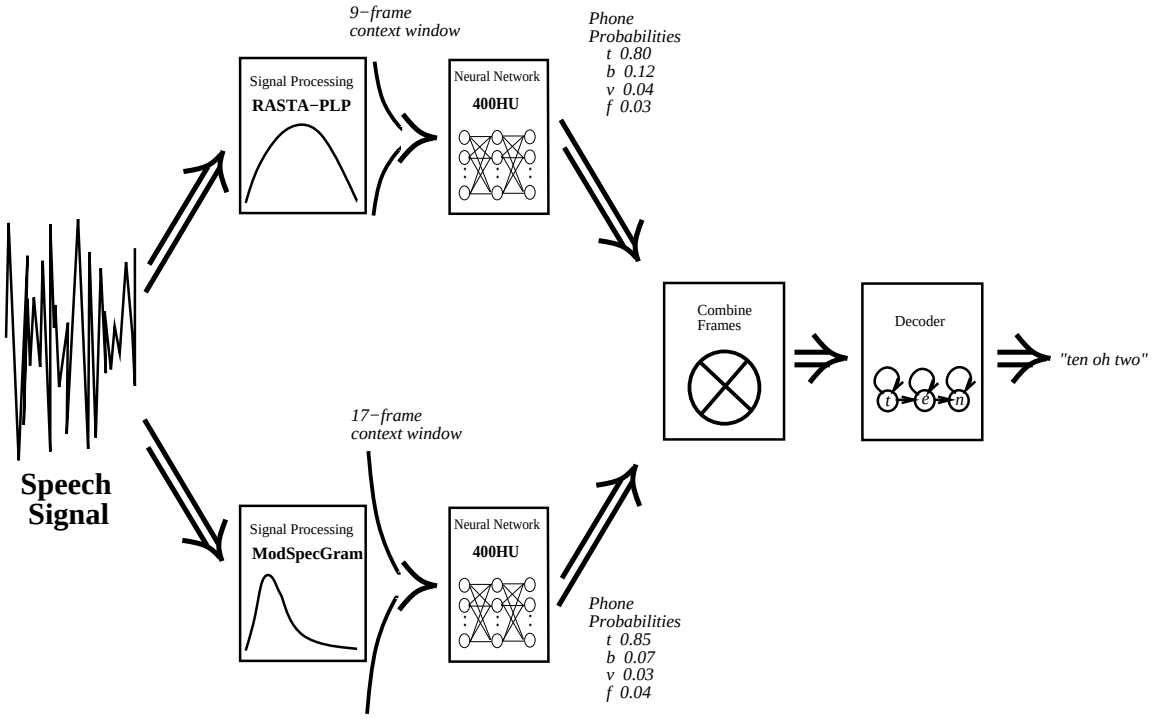

Figure 6.1: Combination of systems at the frame level.

sentences, as illustrated in Figure 6.1. The only change made to the decoder parameters was that the language model scaling factor[3] was doubled. The logarithms of the acoustic probabilities were, on average, twice the magnitude that they had been in the original, independent systems.

As noted earlier, since the combining method multiplied output probabilities on a one-to-one basis, only systems with matching outputs could be combined with this simple method. Therefore only systems using phoneme-oriented units could be combined with the baseline system. Since this precluded a combination with the focus system, the modulation spectrogram system with phone units was used for these experiments, being the closest to the focus system while still meeting the matching-output requirement. This system incorporated the most syllable-based information given the constraints on choice of recognition unit. Error analysis with this system, detailed in the previous section, supported this choice by showing that it produced the smallest proportion of identically-incorrect errors with the baseline. The resulting error rates for this combination, on both clean and reverberant Numbers test sets, are listed in Tables 6.7 and 6.8, along with the error rates of the systems individually.

The simple combining of the systems at the frame level resulted in a significant decrease in word error rate, 13.4% relative for clean speech and 36.8% for reverberant speech. These

[3]Language model scaling factors are empirically determined values that weight the influence of the language model over the acoustic information during decoding (and is discussed more fully in Chapter 3). Trials with the cross-validation set that doubled the language model scaling factor for the baseline alone showed no appreciable difference in performance.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *phones*, 17 frames	clean	6.8%	9.2%	5.9%
		reverb	29.3%	26.6%	19.4%

Table 6.7: Performance results (word error rate) scores of each system independently and after combining, at the *frame* level on clean and reverberant versions of the Numbers *development* test set.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *phones*, 17 frames	clean	6.7%	8.6%	5.8%
		reverb	28.0%	25.8%	17.7%

Table 6.8: Performance results (word error rate) scores of each system independently and after combining, at the *frame* level on clean and reverberant versions of the Numbers *evaluation* test set.

findings are consistent with those previously reported [107]. They are also consistent with the suggestions of the frame level analyses and comparisons. Although the method enforces a lockstep synchronization between the two recognition systems and prevents the use of the focus system in this experiment, the results show that much improvement can be gained with a relatively simple combining scheme, and with minimal changes to existing systems.

6.3.3 Discussion

The frame-level analysis examines the abilities of the baseline and variant systems to classify individual frames. The scores showed that the experimental variants were more accurate than the baseline for reverberant speech. Since reverberation tends to smear spectro-temporal information in time, it was reasonable to expect that the integration of information over a longer time window could be helpful. The introduction of modulation spectrogram features produced not only greater frame-level accuracy for reverberant speech, but also more complementary patterns of recognition behavior, a boon for combining strategies.

The frame-level combination can be thought of as two independent perceptual systems interfacing at the level of the phone, with no other intermediate structure, to produce words and sentences. The speech was essentially statically segmented into 25-ms frames; then each system used different criteria to associate phones with each interval. A subsequent process then simply accepted these hypotheses with equal weighting and formulated a unified recognition output. The baseline system, with RASTA-PLP features, emphasized phonetic segment transitions and integrated this information over 9-frame neural network windows. The modulation spectrogram system emphasized the energy over syllable-timed intervals, integrated over 17-frame neural network windows. The segment onsets apparently supplied information that was somewhat orthogonal to the syllable-length analysis, resulting in higher accuracy overall when combined.

6.4 The Syllable Level

This section considers combining recognition system output streams at the syllable level. The error rate calculations are in terms of canonically defined syllables, not the syllables in the task's syllabary; this was done for ease of scoring, as discussed in Section 5.4. The analyses and comparisons are based on canonical syllables as well, and indicate that syllable error rates are closely related to word error rates and exhibit very similar trends.

Combining at the syllable level allowed the decoder to use information from each of the streams in a more desynchronized manner. In the version of the HMM recombination strategy [45] implemented at ICSI,[4] the decoding could use a phone probability from one stream and a half-syllable (from the syllabary) probability from the second stream subject only to the constraint that the two streams have common syllable beginning and end points. While this removed the limitation experienced with frame-level combinations of requiring that the two systems have the same output unit, matching phones with their corresponding half-syllables does require that pronunciations in the two streams be the same at the syllable

[4] With Nikki Mirghafori.

Variant (paired with Baseline)	Clean		Reverberant	
	Number of Error Tokens	Percentage of Total Tokens	Number of Error Tokens	Percentage of Total Tokens
RASTA + phones, 17 frames	496	8.5%	2,016	32.9%
RASTA + half-syllables, 17 frames	636	10.7%	2,392	38.1%
modulation spectrogram + phones, 17 frames	674	11.5%	2,438	39.0%
modulation spectrogram + half-syllables, 17 frames **focus**	771	13.0%	2,627	41.7%

Table 6.9: Number of tokens which contributed to the error analysis and the percentage of total tokens these error analysis syllables represented in the *clean* and *reverberant* speech version of the development test set of Numbers.

level. The focus system fits within these constraints, so the combination experiments in this section involve merging the baseline with the focus system.

6.4.1 Analysis

The analysis procedure first decomposes recognized words into their canonical (i.e., dictionary) syllable components with a single pronunciation per word. The list of canonical syllabic pronunciations used in these analyses is provided in Appendix A.3. These differ from the recognition system lexicon's syllabary; for example, the word "twenty" is analyzed as the syllables /t-w-eh-n/ /t-iy/ (ICSI56 orthography), rather than as one of the word's 10 pronunciation alternatives in the Numbers' lexicon.

As described in Section 5.4, the canonical pronunciations were chosen for two reasons: 1) This method smoothed over the recognition of different syllables that correspond to the same word; for these experiments, recognizers should not be penalized for recognizing different syllables if the output is word-consistent. 2) Human listeners probably perceive speech in terms more similar to canonical lexical units than to syllabary units. Moreover, they do not generally perceive small variations in syllables in conversational speech, if the word is correctly understood. For example, it was found by transcribers in the Switchboard Transcription Project [76] that the word "problem" was often pronounced in a reduced manner, "proem," though the word was initially perceived, before careful examination of the spectrogram, as the fully expressed, two-syllable word. While whole syllables can occasionally be deleted, as discussed previously, the rate of complete deletion is very small compared to the deletion rate of phones [78].

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	23.8%	25.2%	**42.7%** (212)	8.3%
RASTA + half-syllables, 17 frames	40.6%	26.6%	**23.3%** (148)	9.6%
modulation spectrogram + phones, 17 frames	43.9%	26.9%	**20.5%** (138)	8.8%
modulation spectrogram + half-syllables, 17 frames **focus**	51.0%	25.3%	**13.4%** (103)	10.4%

Table 6.10: Distribution of syllable error tokens across the four analysis categories in the *clean* speech version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

The analysis process, analogous to the one described for words, computed the error correlation between systems in terms of syllables. The procedure compared the focus system and each of the supplemental systems to the performance of the baseline system. Table 6.9 shows the total number of syllable error tokens used in the analysis. Tokens identified correctly by both systems in a pairing were eliminated.

The comparisons reflect those in the word-level case study, examined in detail in Section 6.1. The syllable analysis level exhibits very similar trends.

Table 6.10 shows that the systems with modulation spectrogram features or half-syllable recognition units were less correct than the baseline by a considerable margin. When both systems of a pair were wrong, the errors were more likely to be of the Identical-Incorrect variety than its complement (Different-Incorrect). The focus system, with both modulation spectrogram features and syllable-based recognition units, had the lowest proportion of Identical-Incorrect errors.

With reverberant speech, illustrated in Table 6.11, the modulation spectrogram and the half-syllable unit had the effect of narrowing the gap between the number of error tokens identified correctly only by the baseline system and the number identified correctly only by the variant. Further, when the two systems both produced erroneous outputs, the focus system again had the lowest Identical-Incorrect value. These analyses further motivated combining the systems.

6.4.2 Combining

Combining at the syllable level entails merging the probabilities of different hypotheses at the end of each syllable, a process illustrated in Figure 6.3. The desired functionality is to

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	19.1%	27.7%	**37.0%** (746)	16.2%
RASTA + half-syllables, 17 frames	31.8%	25.3%	**22.5%** (538)	20.4%
modulation spectrogram + phones, 17 frames	33.1%	36.7%	**12.6%** (307)	17.6%
modulation spectrogram + half-syllables, 17 frames **focus**	37.9%	34.8%	**9.7%** (255)	17.5%

Table 6.11: Distribution of syllable error tokens across the four analysis categories in the *reverberant* speech version of the development test set of Numbers. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

have two independent speech recognition processes that interact only at the endpoints of syllables. At these combination points, hypothesis scores for each path are integrated before the recognition processes continue with their individual computations. In the example shown in the upper portion of Figure 6.2, two recognition processes are depicted, each recognizing the word "ten." At the end of the syllable the paths meet and combine values before separating again for the next syllable. Between combination points the recognition behavior of the systems can diverge and desynchronize.

This combination method uses the HMM-recombination algorithm of Bourlard and Dupont [19]. The background of this method is discussed in Section 3.5. To use standard decoders without modifying them, Bourlard and Dupont combined HMM models in a many-to-many mapping before the model was input into the decoder.

As mentioned previously, the HMM-recombination scheme was reimplemented at ICSI for the YO [88] decoder. As illustrated for a simple example in Figure 6.2, the HMM-recombination scheme expands two parallel HMMs (an atypical form) into a single HMM with a conventional description. For these experiments, the amount of desynchronization between the states of the two parallel models was limited by stipulating that the current states of the two streams must share a phone constituent. For the case in which one stream involved a phoneme-based HMM and the other was syllable-based, this amounts to requiring that the syllable of the current syllable-based HMM state contain the phone of current phoneme-based HMM state. The decoder can use the /t/ of the phoneme-based stream only at the same time that the decoder uses the /t-eh/ of the syllable-based stream (Figure 6.2). The decoder, however, can use the /eh/ of the phoneme-based stream at the same time as either the /t-eh/ or the /eh-n/ of the syllable-based stream. The HMM states of the new, expanded HMM represent all of the permissible temporal synchronization conditions between the phoneme-based stream and the syllable-based stream. This HMM-

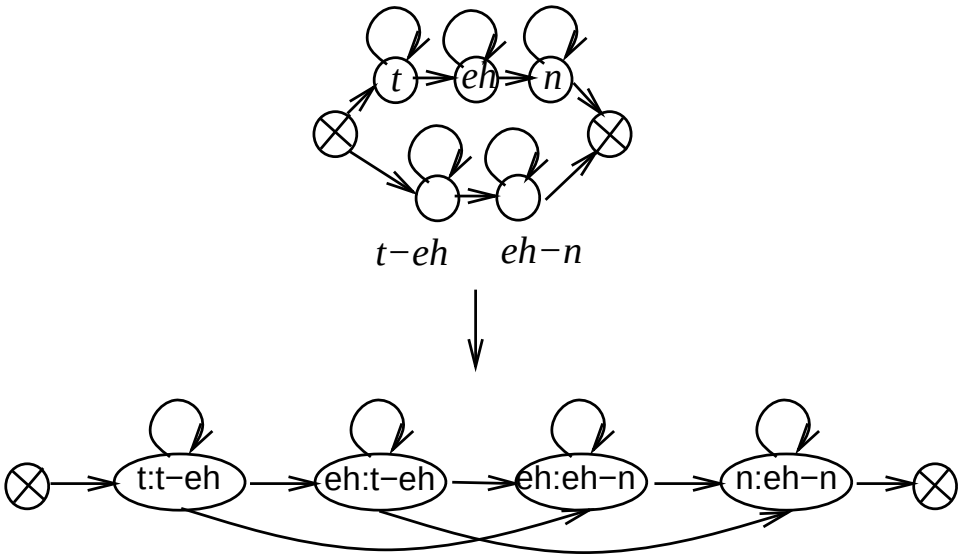

Figure 6.2: A simple example of an HMM recombination implementation for the word "ten," with desynchronization allowed only within half-syllables.

recombination scheme also creates a single probability stream from the dual probability steams of the individual neural networks. The corresponding probabilities are multiplied to generate new values for these expanded states. Word models are formed by concatenating these syllable-sized expanded HMMs. With complex syllables, minimum duration models implemented with repeated states and multiple pronunciations, the full word-length HMMs comprised hundreds of probabilities and thousands of states.

Arcs at the bottom of the lower model in Figure 6.2 show that states can be skipped, reflecting the different ways the two streams can proceed in parallel. Thus, as the decoding progresses from one frame to the next, the phoneme-based stream can proceed from /t/ to /eh/ and the syllable-based stream can proceed from /t-eh/ to /eh-n/ or stay with /t-eh/. As illustrated in Figure 6.2, the two streams meet only at the beginnings and ends of each syllable HMM. This has the effect that probabilities for the same sentence hypothesis from the two different systems are linearly combined at the end of each syllable, an enforced synchronization point.

Bourlard and Dupont combined a phonetic stream with a syllabic stream, where a single model described all syllables. In the work described in this thesis, there is a separate, unique model for each syllable. HMM-recombination was used to integrate the baseline system with the focus system, i.e., the system incorporating the most syllable-based information. Error analysis, detailed in the previous section, had showed that this pairing had the lowest error correlation values.

The word error rates for the combined system, tested on the Numbers task, are listed in Tables 6.12 and 6.13. This combination method produced a considerable improvement in accuracy over the performance of each constituent system alone, amounting to a 23.9% relative gain for clean speech and a 40.3% relative improvement for reverberant speech. The

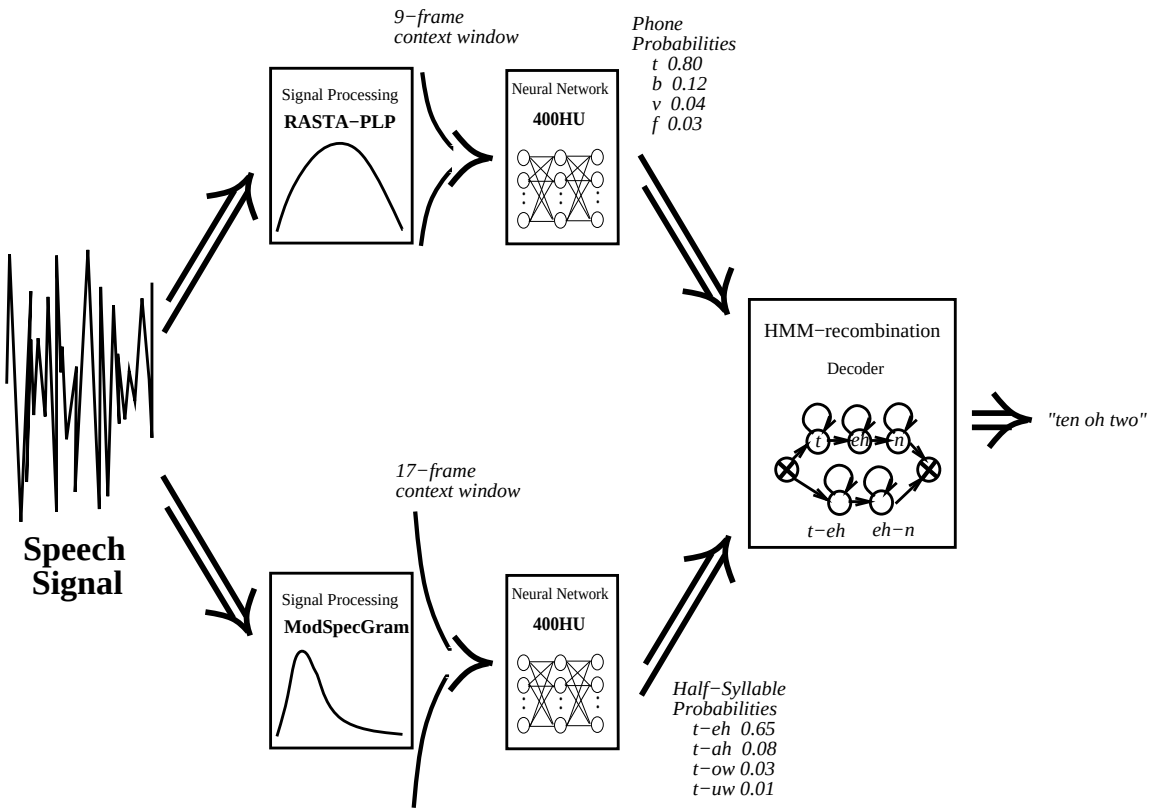

Figure 6.3: Combination of systems at the syllable level.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *half-syllables*, 17 frames **focus**	clean	6.8%	10.6%	5.4%
		reverb	29.3%	30.1%	17.6%

Table 6.12: Performance results (word error rate) scores of each system independently and after combining, at the *syllable* level on clean and reverberant versions of the *development* test set.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *half-syllables*, 17 frames **focus**	clean	6.7%	10.0%	5.1%
		reverb	28.0%	30.1%	16.7%

Table 6.13: Performance results (word error rate) scores of each system independently and after combining, at the *syllable* level on clean and reverberant versions of the *evaluation* test set.

complexity of the combination-method implementation, however, was much larger than with the frame-level method.

6.4.3 Discussion

The syllable-level error analysis compared systems based on recognized syllables. For clean speech, the introduction of syllable-based elements produced a degradation in accuracy but an increased divergence in errors compared to the baseline. The divergence in errors was more pronounced with reverberant speech. Also evident for reverberant speech, the long-time span neural network context window provided greater accuracy than the baseline for the variants with phoneme-based output units.

The combination strategy, HMM-recombination, can be likened to two separate perceptual processes interfacing at the level of the syllable. As in the discussion of Section 4.5, the combination process can be interpreted as dynamically hypothesizing syllable-length intervals in the speech stream and attaching to them information from the two separate perceptual processes. The phone hypotheses from the baseline, and the half-syllable hy-

potheses from the focus system become features of the underlying syllable-length interval. Greenberg uses such a model to explain pronunciation variation [78].

The resulting improvement in word-error rate is probably due to the successful combination of the complementary aspects of the two recognition systems. This combination strategy produced the lowest error rates of the all methods examined.

6.5 The Utterance Level

This section considers recognition system output streams at the whole-utterance level, the unit used for analysis and combining.[5] Because a recognized string of words can have both correctly and incorrectly recognized words, analyses at the whole utterance level can mask trends involving smaller speech units. An entire utterance may often be too large a unit for combining methods in general, though it is manageable for the Numbers corpus. In practical applications with large vocabularies and long input utterances, one might imagine combining recognition system output streams at the *phrase* level.

By combining at the utterance level, the combined streams can become completely desynchronized between the beginning and the end of the speech input. In the extreme, one stream can produce completely different syllables or phones from another stream for the same acoustic input. If the word strings are the same, this combining method considers the outputs from two different streams to be the same answer regardless of the internal temporal alignment of word boundaries or recognized phones and syllables. Combining enforces synchronicity only at the start and termination of the utterance. Thus, the recognition systems can be completely different and separately optimized.

6.5.1 Analysis

The error analysis method can also be used at the whole-utterance level. Table 6.14 shows the number of sentences that contribute to the error analysis. As previously, the sentences that both systems in each pair recognized 100% accurately have been removed.

The trends observed with the detailed case study at the word-level are also evident in the utterance level analysis.

For clean speech, Table 6.15 shows the percentage of sentence errors where only one system in a pairing recognized the utterance with 100% accuracy. The table also shows the percentage of sentence errors where both systems made errors in recognizing utterances. As in the other analyses, the systems with modulation spectrogram or half-syllable unit elements produced more errors than the baseline. When both systems produced erroneous output, these systems tended to produce errors different from the baseline; these recognizer pairings produced lower proportions of Identical-Incorrect errors. The focus system, with

[5] Parts of the study involving combining hypotheses at the whole-utterance level, as detailed in this section, were the result of a collaboration between Brian Kingsbury and myself, with advice from Nelson Morgan and Steven Greenberg. This work was briefly presented at the International Conference on Acoustics, Speech and Signal Processing, 1998 [209].

Variant (paired with Baseline)	Clean		Reverberant	
	Number of Error Sents	Percentage of Total Sents	Number of Error Sents	Percentage of Total Sents
RASTA + phones, 17 frames	286	23.7%	863	71.6%
RASTA + half-syllables, 17 frames	360	29.9%	920	76.3%
modulation spectrogram + phones, 17 frames	378	32.1%	936	77.6%
modulation spectrogram + half-syllables, 17 frames **focus**	419	34.7%	976	80.9%

Table 6.14: Number of sentences which contributed to the error analysis and the percentage of total sentences these error analysis utterances represented in the *clean* and *reverberant* speech versions of the Numbers development test set.

both modulation spectrogram information and half-syllable units, had the lowest proportion of Identical-Incorrect tokens.

As mentioned previously, using such large units for comparison can mask more detailed trends. Table 6.16 shows further analysis of the sentences where systems produced erroneous, but differing outputs; in this case, the experimental system variants tended to be more correct than the baseline.

For reverberant speech, Tables 6.17 and 6.18 give the analogous error analysis category values. As before, the reported figures represent the percentage of sentences where one or both systems in a pairing made errors in recognizing the utterance. As observed with the other analyses, the gap between the Baseline-Only-Correct percentages and the Variant-Only-Correct percentages was reduced compared to the clean speech case for the systems with modulation spectrogram and/or half-syllable unit elements. When both systems produced erroneous output, these systems exhibited lower error correlation values (a smaller Identical-Incorrect percentage). They also exhibited a tendency to produce different errors rather than the same erroneous sentence. The focus system, with both modulation spectrographic features and half-syllable recognition units, again had the lowest number of Identical-Incorrect tokens. Further evaluation of the sentences where systems produced erroneous, but differing outputs showed that the variant systems with the half-syllable unit were more correct than the baseline.

In Chapter 5, analyses showed that sentence error was roughly correlated with word error rate. One might hypothesize that the systems with the larger error rates would tend to perform less well *uniformly* across sentence inputs. The analyses above show that for some of the sentences the variants were more correct than the baseline. An early pilot experiment

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	15.4%	17.1%	**43.4%** (124)	15.4%
RASTA + half-syllables, 17 frames	32.8%	22.5%	**20.0%** (72)	24.7%
modulation spectrogram + phones, 17 frames	37.5%	18.6%	**19.4%** (73)	24.5%
modulation spectrogram + half-syllables, 17 frames **focus**	42.2%	19.8%	**10.0%** (42)	27.9%

Table 6.15: Distribution of error sentences across four analysis categories in the *clean* speech version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

Variant (paired with Baseline)	Baseline More Correct	Variant More Correct	Both Equally Wrong
RASTA + phones, 17 frames	26.1%	31.9%	42.0%
RASTA + half-syllables, 17 frames	19.1%	31.5%	49.4%
modulation spectrogram + phones, 17 frames	23.2%	37.9%	38.9%
modulation spectrogram + half-syllables, 17 frames **focus**	17.2%	38.5%	44.4%

Table 6.16: The percentage of sentences where one system was more correct than the other or where both systems were equally wrong for the Different-Incorrect sentences on the *clean* speech version of the Numbers development test set.

Variant (paired with Baseline)	Baseline Only Correct	Variant Only Correct	**Identical-Incorrect** (count)	Different-Incorrect
RASTA + phones, 17 frames	8.5%	15.1%	**33.2%** (287)	43.5%
RASTA + half-syllables, 17 frames	14.1%	12.7%	**17.1%** (157)	58.1%
modulation spectrogram + phones, 17 frames	15.6%	22.4%	**8.8%** (82)	53.2%
modulation spectrogram + half-syllables, 17 frames **focus**	19.1%	19.7%	**5.8%** (57)	55.4%

Table 6.17: Distribution of error sentences across the four analysis categories in the *reverberant* speech version of the Numbers development test set. The actual number of error tokens is shown in parentheses for the Identical-Incorrect column.

Variant (paired with Baseline)	Baseline More Correct	Variant More Correct	Both Equally Wrong
RASTA + phones, 17 frames	37.9%	24.3%	37.9%
RASTA + half-syllables, 17 frames	25.8%	33.1%	41.1%
modulation spectrogram + phones, 17 frames	33.3%	29.3%	37.3%
modulation spectrogram + half-syllables, 17 frames **focus**	28.1%	32.0%	39.9%

Table 6.18: The percentage of sentences where one system was more correct than the other or where both systems were equally wrong for the Different-Incorrect sentences on the *reverberant* speech version of the Numbers development test set.

Variant (paired with Baseline)	Variant Better Than Baseline sents (words)	Variant W.E.R. on Subset	Baseline W.E.R. on Subset	W.E.R. After Combining
RASTA + phones, 17 frames	67 (269)	8.6%	37.2%	5.2%
RASTA + half-syllables, 17 frames	98 (366)	5.2%	36.3%	4.4%
modulation spectrogram + phones, 17 frames	94 (392)	6.9%	36.0%	4.4%
modulation spectrogram + half-syllables, 17 frames **focus**	103 (406)	5.9%	35.2%	4.3%

Table 6.19: On some sentences the experimental variant systems performed better than the baseline system, with the *clean* version of the Numbers development test set (1,206 sentences, 4,673 words). The number of words in these subsets of sentences, selected by an oracle, is shown in parentheses.

examining the idea of combining systems at the utterance level involved a "cheating" procedure for estimating an approximate upper bound on the accuracy achievable by combining the best output from two systems. This cheating procedure does not yield an actual upper bound for the combining procedure described in the next section, because the combining procedure uses more hypotheses per utterance.

In this cheating experiment the combined sentence output was created by taking the best scoring sentence from either of the systems in each pair, with knowledge of the true answers. Table 6.19 shows the number of sentences where the experimental system performed better than the baseline system. The table also gives word error rates for each system on this subset of sentences. These percentages illustrate that on a significant number of sentences the experimental systems achieve substantially greater accuracy despite a higher overall error rate. The table also shows that the system with the syllable-based elements (modulation spectrogram features and half-syllable recognition units) produced the largest number of better-than-baseline sentences. The RASTA-PLP system with half-syllable units was the runner-up. As seen in Table 6.20, with reverberant speech the system with modulation spectrogram features and phoneme-based units produced the largest number of better-than-baseline sentences. The focus system, with modulation spectrogram features and half-syllable recognition units, was not far behind.

These analyses and pilot experiments indicate that a careful combination of the outputs of two systems may afford considerable reduction in word error rate, to a level better than either of the constituent systems separately.

Variant (paired with Baseline)	Variant Better Than Baseline sents (words)	Variant W.E.R. on Subset	Baseline W.E.R. on Subset	W.E.R. After Combining
RASTA + phones, 17 frames	274 (1,123)	18.8%	51.9%	21.3%
RASTA + half-syllables, 17 frames	251 (968)	19.0%	53.3%	22.2%
modulation spectrogram + phones, 17 frames	377 (1,593)	15.2%	48.1%	18.0%
modulation spectrogram + half-syllables, 17 frames **focus**	345 (1,393)	15.8%	50.6%	18.9%

Table 6.20: On some sentences the experimental variant systems performed better than the baseline system, with the *reverberant* version of the Numbers development test set (1,206 sentences, 4,673 words). The number of words in these subsets of sentences, selected by an oracle, is shown in parentheses.

6.5.2 Combining

The combining procedure at the utterance level added the log likelihoods of the same sentence hypotheses from each of two systems at the end of the decoding process. This scheme was implemented using a sequence of three different decoders, YO [88] for its forced alignment capability, NOWAY [164, 163, 165] for its lattice generation function, and LATTICE2NBEST [166] for its lattice decoding ability.[6] A number of interfacing scripts glued the programs together, enabling state desynchronization to occur over the entire utterance.

The decoding sequence used can produce a somewhat different behavior from YO. The pruning of hypotheses is managed first by NOWAY, then by LATTICE2NBEST with a strict upper limit on the number of distinct hypotheses. Because of the dissimilar properties of the decoding process, these word error rates are not strictly comparable with other error rates reported thus far using only YO, though they do represent the best performance achieved with this approach.

The utterance combination procedure, illustrated in Figure 6.4, involved generating up to 150 best hypotheses from each system (i.e., the baseline system and the focus system). First, NOWAY was used to produce a lattice from each system, which was passed to LATTICE2NBEST to obtain the N-best hypotheses. The combining procedure merged the two hypothesis lists and rescored each utterance with forced alignment (via YO) using both recognition systems. Rescoring was necessary because the sets of utterances from the two systems were often different. The procedure added corresponding pairs of scores for each utterance in the merged N-best list and reordered the list of utterances according to this

[6] Related decoding technology is discussed in more detail in Section 3.4.

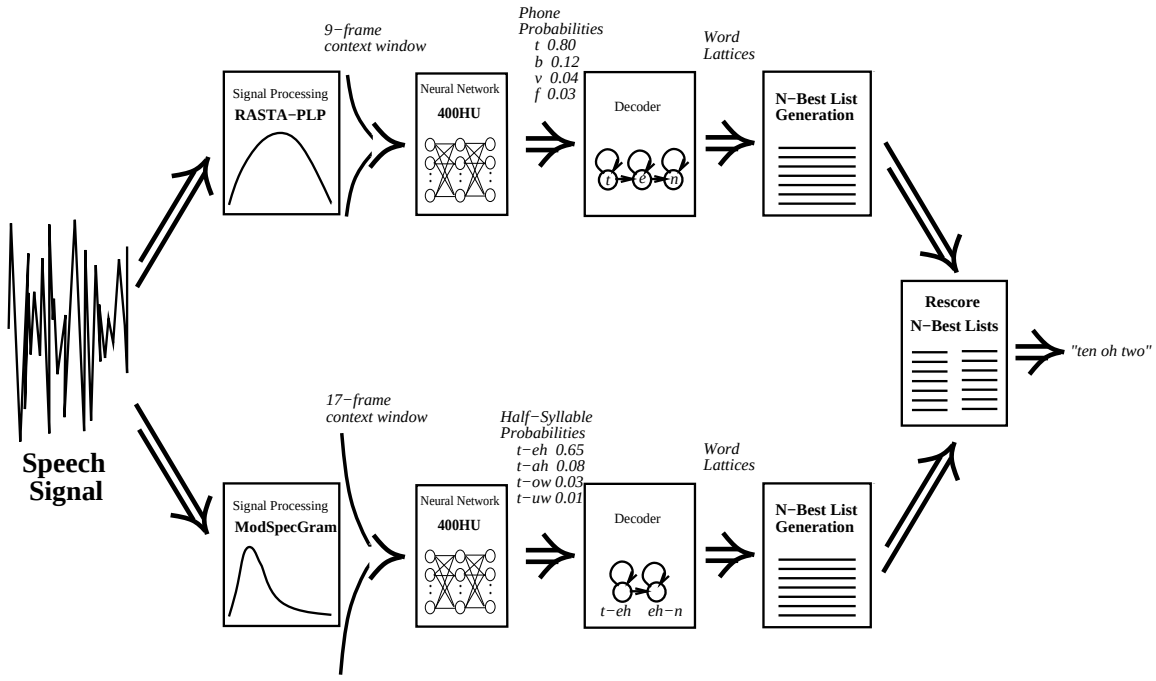

Figure 6.4: Combination of systems at the whole utterance level.

result. The top scoring (i.e., lowest cost) hypothesis of the combined list then emerged as the overall recognized sentence for the joined systems. The results of this procedure are shown in Tables 6.21 and 6.22.

The combining resulted in a substantial improvement in accuracy compared with the performance of the constituent systems with both clean and reverberant speech. The numbers show a 17.9% relative improvement in clean speech and a 30.0% relative improvement for reverberant speech. These results, however, were not quite as good as the combination scores for the syllable-level method, which restricted desynchronization to shorter intervals and allowed the two systems to interact at an earlier stage of processing.

6.5.3 Discussion

The utterance level analysis is consistent with the trends observed at the word level and at the frame and syllable levels.

Combining at the utterance level can be likened to two perceptual processes independently creating complete hypotheses for the spoken word sequences; one based on phones, the other based on syllables. A subsequent process then combines the complete set of hypotheses with an equal weighting and produces a unified output. The improvement in word-error rate can again be attributed to the successful fusion of two recognition systems with divergent behavior.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *half-syllables*, 17 frames **focus**	clean	6.8%	10.6%	5.6%
		reverb	29.3%	30.1%	20.0%

Table 6.21: Performance results (word error rate) scores of each system independently and after combining, at the utterance level on clean and reverberant versions of Numbers *development* test set.

Baseline	Variant	Condition	W. E. R. Baseline	W. E. R. Variant	W. E. R. Combined
RASTA + phones, 9 frames **baseline**	modulation spectrogram + *half-syllables*, 17 frames **focus**	clean	6.7%	10.0%	5.5%
		reverb	28.0%	30.1%	19.6%

Table 6.22: Performance results (word error rate) scores of each system independently and after combining, at the utterance level on clean and reverberant versions of Numbers *evaluation* test set.

Test Condition	Baseline phones	Frame-Level phones+phones	Syllable-Level phones+half-syllables	Utterance-Level phones+half-syllables
clean	6.7%	5.8%	5.1%	5.5%
reverb	28.0%	17.7%	16.7%	19.6%

Table 6.23: Performance results (word error rates) of baseline and combined systems for clean and reverberant versions of the Numbers evaluation test set.

6.6 Summary

In these experiments combining reasonably good recognition systems with low error correlation led to an improvement in recognition accuracy. The results are summarized in Table 6.23. The merging of the baseline system with the focus system at the syllable-level produced a 24% improvement in word error rate in the clean version (from 6.7% to 5.1%) and a 40% improvement in the reverberant case (from 28.0% to 16.7%). Combining these systems at other levels than the syllable-level also produced large increases in accuracy.

The analyses in this chapter indicate that these improvements are attributable to integrating the differences in recognition behavior between the baseline system and the focus system. The systems with syllable-based elements tended to make different errors from the phoneme-based baseline system. Counting the number of errors where a system variant and the baseline system recognized the same value yielded an error correlation measure, referred to as "Identical-Incorrect." This represents one method of coarsely quantifying the "sameness" of the errors between the two systems. This measure certainly does not reflect the full spectrum of behavioral differences between two systems. Its value lies in summarizing a considerable amount of information relevant to combination strategies in a few values. Tables 6.24 and 6.25 show the results of calculating this Identical-Incorrect value for the baseline system paired with the focus system as well as with each of the three supplemental variants for both clean and reverberant versions of the Numbers development test set. The percentages in these tables represent the fraction of total error tokens that fall under the Identical-Incorrect category. The columns corresponding to Tables 6.24 and 6.25 are highlighted in the main body of this chapter in boldface.

In each column from syllable level to utterance level, the tables show that the focus system has the smallest number of Identical-Incorrect errors.[7] These figures support the observation that the behavior of the focus system is largely complementary to that of the baseline system. Incorporating syllable-based elements appeared to promote divergence in recognition behavior and help create systems that can be combined towards an overall improvement in accuracy.

[7] For frame-level combination, the systems were constrained to be based on phoneme-oriented units only.

Variant (paired with Baseline)	Frame-Level	Syllable-Level	Word-Level	Utterance-Level
RASTA + phones, 17 frames	20.9%	42.7%	43.6%	43.4%
RASTA + half-syllables, 17 frames	N/A	23.3%	24.9%	20.0%
modulation spectrogram + phones, 17 frames	14.8%	20.5%	21.5%	19.4%
modulation spectrogram + half-syllables, 17 frames **focus**	N/A	13.4%	14.1%	10.0%

Table 6.24: The Identical-Incorrect values, as a percentage of total error analysis tokens, for each of the system variants paired with the baseline, at each of four stages. Reported for the *clean* speech version of the Numbers development test set.

Variant (paired with Baseline)	Frame-Level	Syllable-Level	Word-Level	Utterance-Level
RASTA + phones, 17 frames	38.2%	37.0%	38.9%	33.2%
RASTA + half-syllables, 17 frames	N/A	22.5%	24.9%	17.1%
modulation spectrogram + phones, 17 frames	18.6%	12.6%	13.6%	8.8%
modulation spectrogram + half-syllables, 17 frames **focus**	N/A	9.7%	11.0%	5.8%

Table 6.25: The Identical-Incorrect values, as a percentage of total error analysis tokens, for each of the system variants paired with the baseline, at each of four stages. Reported for the *reverberant* speech version of the Numbers development test set.

6.7 Conclusion

Combining experimental, syllable-based systems with the baseline system improved the recognition accuracy of Numbers over that of the individual systems alone, as can be seen in Table 6.23. These combined results are significantly improved over the baseline at the 0.05 significance level.[8] Analyses and comparisons of the systems individually and in pairs suggests that this benefit is due to the differences in recognition characteristics between the systems. When the error correlation between two systems is low, as quantified by the Identical-Incorrect measure, the systems appeared to complement each other, mitigating weaknesses and enhancing strengths.

The addition of syllable-based information helped to create systems with both reasonable recognition performance and disparate error characteristics by emphasizing different properties of the acoustic signal. The modulation spectrogram features, developed by Kingsbury and Greenberg, promoted divergence in errors. The half-syllable unit also added to the dissimilarity between systems, but not as much as did the features. A lower Identical-Incorrect value resulted between the baseline and the experimental alternatives with more syllable-based elements, shown in Tables 6.24 and 6.25. The focus system, with both modulation spectrogram features and the half-syllable recognition unit, was the most consistent across analysis levels in having the smallest proportion of errors identical to the baseline.

Combining this system with the baseline at the syllable level produced the overall best error rate for both clean and reverberant versions of the Numbers test sets. The improvement is also slightly greater than with the frame-level combination, a statistically significant effect ($p < 0.05$) for reverberant speech, with development test set data. With the reverberant version of the evaluation test set, the positive effect is not significant. The improvement observed with reverberant speech with this combining method is significantly larger than that for utterance-level combining. The major difference between this strategy and that at the shorter frame or longer utterance level was the added, heterogeneous structure produced by synchronization at the ends of syllables, which could vary considerably in length. The frame-level and utterance-level combination methods had more homogeneous, fixed interval synchronization properties. The error-rates for the frame-level combination scheme were almost as low as for the syllable-level, however, as shown in Table 6.23. The frame-level combination scheme had an implementation that was considerably simpler than the syllable-level approach.

Combining strategies that involve a common beginning for each syllable, such as combining at the frame and syllable level, allow for the possibility of using syllable onset information from Chapter 4 during the combining process. From the implementation point of view, the frame-level combination method can integrate this information most directly. At the syllable level, syllable onset information is less readily merged into the HMM recombination strategy[9] used. Incorporating syllable onsets would double the already enormous HMM models and neural network activation files required by the syllable-level combination implementation. For combining at the utterance level, where the two streams can become

[8] Significance testing used normal approximations to binomial distributions and used a Z-score to test whether the two distributions were significantly different.

[9] HMM recombination is described in more general terms in Section 3.5.

completely desynchronized, the syllable onset information must be incorporated into the two constituent recognition systems before the combining stage. Although not pursued here, using syllable onset information can make a further improvement in the overall accuracy of these systems.

This chapter explored combining the baseline system with the focus system at three levels: frames, syllables, and whole utterances. Combination methods varied from the very simple (frame level), to the somewhat more involved (syllable level), to the complex (utterance level). Each of the different combining levels has separate advantages and disadvantages, yet combining at any of the levels showed significant improvement over using a single system by itself. This suggests a possibly useful, more general hypothesis for combining systems, not limited to syllable-based investigations: given multiple recognition systems, each with reasonably good performance and low error correlation characteristics, merging the probabilities at some level can improve speech recognition performance over that of the individual, constituent systems.

Chapter 7

Discussion and Conclusion

Many of the basic brain mechanisms underlying human speech processing are still poorly understood. There is some evidence, however, that speech perception incorporates information related to the temporal properties of syllables. This observation suggested a strategy for automatic speech recognition, that of combining syllable-based information with a well-established phoneme-based speech recognizer. Combination and merging paradigms are not new; research along similar lines has been pursued at least since the early 1970s. The work described in this thesis has attempted to capitalize on the most recent research advances by using feature analysis methods newly developed by colleagues at ICSI and by pursuing combination strategies at several separate stages in the recognition process for greater overall performance. These experiments demonstrate the potential for improving speech recognition accuracy using systems and procedures that incorporate syllable-based information. These methods were effective for a modest-sized pilot task; work remains to develop efficient versions of these techniques for large vocabulary versions.

This chapter begins with a summary of the thesis. Section 7.2 discusses the implications of these results in detail and Section 7.3 lists the contributions made by this thesis to automatic speech recognition. Section 7.4 explains the future possibilities for this work, including the issue of large vocabulary tasks. The thesis concludes with some general reflections on the field of automatic speech recognition and its relation to this work.

7.1 Summary

This thesis began with a discussion of the role of the syllable in the identification and segmentation of speech. A short review described some of the syllable's function in lexical access in human perceptual systems. A literature search revealed that the syllabic properties of speech are highly controversial with, as yet, no definitive consensus. A study of syllable usage in conversational speech showed that a representative sample of casual speech (of considerable size) was relatively simple to describe with syllables and that the syllable may be an effective representational and organizational unit. This led to reconsidering the syllable for automatic speech recognition systems, which revealed both advantages and disadvantages for ASR.

System	Error Rate
no onset information (baseline)	9.1%
onset used with threshold	8.2%

Table 7.1: Performance results (word error rates) with and without acoustically-derived onsets.

The exposition continued with a discussion of the background of the ICSI speech recognition system and the Numbers task, the basis for all the experiments in this thesis.

Experiments incorporating syllable-based information into speech recognition began with integrating acoustically estimated syllabic onsets into a speech recognition system. The chosen methodology involved the design and implementation of a decoder with a separate syllabic level. "Cheating" experiments (using artificial onsets) showed that hints about the syllable segmentation of the speech input could substantially decrease the overall word error rate of the speech recognition system, even if the onsets were determined with only a modest degree of accuracy. Acoustically-estimated (non-cheating) onsets, based on acoustic features developed by Shire and Greenberg [184], were incorporated into the speech recognition system resulting in a 10% relative improvement in accuracy with clean speech (OGI Numbers), as shown in Table 7.1. A method for incorporating onsets without the use of a special decoder was later developed and shared with colleagues who applied these ideas to Broadcast News, a large vocabulary corpus, and achieved a similar improvement in accuracy [33].

These experiments indicated the potential of using syllable-based information at other levels. Investigating this involved the development of a focus experimental system with syllable-based, long-time-span elements at the levels of feature analysis, neural network output and recognition unit. The experiments also involved the development of three supplemental system variants, each with a different subset of syllable-based processing elements. The supplemental systems provided contrast and additional context for comparisons and analyses. The substitution of features or recognition units based on syllable-length, long time spans into the baseline system caused recognition performance to degrade moderately for clean speech and did not improve accuracy for reverberant speech. Using a longer context window helped improve recognition performance to some degree for the reverberant version of the Numbers corpus.

Inspection of the recognition output suggested that the experimental variants made different errors from the phone-oriented baseline system. That is, the variants produced recognition results with a low degree of error correlation with the baseline system, as ascertained by the error analysis described in Chapter 5. Pairing the baseline with the focus system, which incorporated the largest number of syllable-based elements, turned out to produce the lowest proportion of Identical-Incorrect errors (a measure of error correlation) over several levels of analysis, as shown in Table 7.2. The combining experiments involved merging systems at three stages of the decoding process: at the frame level (i.e., using the neural network outputs), at the syllable level and at the whole-utterance level. Combin-

Variant (paired with Baseline)	Frame-Level	Syllable-Level	Word-Level	Utterance-Level
RASTA + phones, 17 frames	38.2%	37.0%	38.9%	33.2%
RASTA + half-syllables, 17 frames	N/A	22.5%	24.9%	17.1%
modulation spectrogram + phones, 17 frames	18.6%	12.6%	13.6%	8.8%
modulation spectrogram + half-syllables, 17 frames **focus**	N/A	9.7%	11.0%	5.8%

Table 7.2: The proportion of Identical-Incorrect errors, as a percentage of total error analysis tokens, for each of the system variants paired with the baseline, at each of four stages. Reported for the *reverberant* version of the Numbers development test set.

Test Condition	Baseline phones	Frame-Level phones+phones	Syllable-Level phones+half-syllables	Utterance-Level phones+half-syllables
clean	6.7%	5.8%	5.1%	5.5%
reverb	28.0%	17.7%	16.7%	19.6%

Table 7.3: Performance results (word error rates) of Baseline and combined systems.

ing the baseline system with the appropriate syllable-based system at any of these levels produced significant reductions in error rates (shown in Table 7.3) of up to 24% relative reduction for clean speech[1] and 40% relative reduction for reverberant speech. This suggested that the approach was effective at improving the accuracy and robustness of the speech recognition system, especially with respect to reverberation in the speech signal.

7.2 Discussion

The research in this thesis was conducted along two major themes: 1) using syllable-based information in ASR and 2) using combination methods to incorporate additional information into a well-established, phoneme-based speech recognition system. The end result was an increase in recognition accuracy, beyond that achieved by the constituent systems singly. Attaining this research goal required that these two major points be developed in tandem. Experiments that considered only one of these two themes did not display the advantages

[1] A different baseline was used in the syllable onset experiments than in the combining of syllable-based systems experiments.

observed by integrating both. Studies with the Numbers corpus showed that the performance of the individual syllable-based systems represented a degradation compared to the baseline, while combining systems that were similar and that did not use much syllable information exhibited little of the marked improvements obtained by the best combinations.

7.2.1 Implications for Syllables in ASR

Researchers have generally assumed that there is a single basic unit of speech recognition. The arguments are often phrased in terms of "the syllable is right and the phoneme is wrong" or vice versa. These experiments and exploratory studies found that using syllable-based, long-time span information design elements in speech recognition elicited a kind of behavior different from phoneme-based, short-time span based systems. The longer-time span elements, however, caused a degradation in the representation of fine detail, probably due to the smearing of information over a longer time span. Hence, the best methodology was a combination of both systems. This simultaneously capitalized on long-time-span integration and short-time-span detail. The best result was obtained by the combination at the syllable-level. In terms of the literature discussion in Chapter 2, it appears that both syllables and phonemes are important units for automatic speech recognition.

Each of the recognition paradigms expressed in these experiments can be interpreted in terms of a dynamic association between particular units and hypothesized speech intervals. In the experiments with syllabic onsets, the speech interval was the length of a syllable and the associated units were phones and syllable onsets. Similarly, combining systems at the syllable-level can be interpreted as first hypothesizing syllable-length segments in the speech signal, then attaching phone and half-syllable units to these intervals. At the frame and utterance level, the experimental setup established the speech intervals *a priori* as the 25-ms frame and whole utterance, respectively. The experimental data gathered so far suggest that the dynamic segmentation of the speech signal at the syllabic level offers the greatest potential for significant gains in ASR performance.

Chapter 2 also related a number of advantages and disadvantages of using syllables in speech recognition. These experiments have touched upon a portion of the advantages cited in favor of syllables in ASR, namely those using syllable segmentation information and integrating speech information over syllable-length intervals. However, this thesis does not address several other syllable-related factors for ASR, such as the issue of using syllables to incorporate prosody. The systems developed also do not explore the potential savings in storage and execution time by sharing syllables, or of the potential improvement in search efficiency by exploiting the regular structure of syllables. The positive results reported in this thesis may be taken to indicate the latent benefit in incorporating additional properties of syllables.

In the experiments in this thesis, incorporating syllables has led to improvements in accuracy despite certain realized complications and other potential problems. The ambiguous nature of syllable boundaries is probably the most important factor limiting improvement due to using syllable onsets. However, Cook and Robinson were able to use the same syllable onset scheme to a similar, positive effect in a large vocabulary task, thus showing that the benefits are consistent and the method is scalable [34]. The ambiguity of syllable

boundaries did not directly affect the recognition performance of the systems with syllable-based elements at the signal processing, neural network context window and recognition unit level since these systems did not enforce rigorous temporal boundaries. Consistent, albeit phonologically approximate, syllabification of the words was sufficient for these experiments. Therefore, it may be possible to obtain improvements for other recognition tasks by using syllable-based information even without a perfect definition of the syllable or syllable boundary, and without waiting for the controversy surrounding the syllable to be resolved.

7.2.2 Implications for Combination in ASR

Increasing the number of parameters available to a recognition system without adding structure produced diminishing returns. The combination of syllable information with phone information enabled the more effective use of additional recognition system parameters. Long-time-span based information, most notably the modulation spectrogram features of Kingsbury and Greenberg, was key to developing a recognition system that produced significant improvements when combined with the baseline. The modulation spectrogram and half-syllable unit representations provided recognition results with errors different from those of the phone-based system.

Error analysis and the Identical-Incorrect metric broadly characterized the potential for improvement through combinations of multiple recognition systems. Table 7.2 shows the proportion of Identical-Incorrect errors committed by both the baseline and the experimental system variants. The number of such errors was lowest for the focus system, which had the most syllable-based design aspects. This suggests a paradigm of first developing systems that have reasonably good performance and low error correlations with one another, and then combining their outputs. This strategy may be particularly effective in improving robustness to surprises in the test set (i.e., where the test set has characteristics different from the training set).

Comparative error analysis is applicable in more general pattern classification tasks as well. Similar analyses can help coarsely quantify the differences in behavior between classifiers into a few meaningful numbers and thereby highlight possibilities for improving accuracy. The distribution of errors in the analysis categories can suggest the most promising combination strategy. For example, if a comparison between two systems showed that they produced entirely complementary errors (i.e., they never got the same word wrong), the desired combination strategy would be quite different than if the two systems always recognized the same words incorrectly, but recognized them in a disparate fashion.

7.3 Thesis Contributions

Humans can understand utterances from the Numbers corpus with near perfect accuracy in both clean and moderately reverberant conditions. Clearly, there is much work remaining in improving automatic speech recognition for even this simple task.

This thesis contributes to the advancement of computer science by presenting a viable

method for improving speech recognition by machines. The ASR community seems to be generally inclined against using the syllable for English due to unresolved linguistic issues and the considerable success of the phone. The experiments discussed show that improvements are possible using certain aspects of the syllable even in the absence of complete answers to the linguistic questions. The results of this work also contribute to the mounting evidence that combination methods have significant potential for improving speech recognition.

This research has incorporated several ideas derived from, or consistent with, theories of human speech perception, including the use of syllable onsets and syllable-length intervals and the combination of coarse processing elements. The experimental results underline the usefulness of clues from human audition for the interpretive understanding and development of automatic processing.

On a personal note, my work has been to develop, refine and extend the concepts and suggestions originally shared with me by Professors Morgan and Greenberg. I like to think that many of the ideas expressed in this thesis came from synergistic collaboration with my professors and my colleagues, rather than a one-way transferral. A major portion of my time was devoted to developing the infrastructure for experiments to demonstrate that syllable-based information can improve recognition accuracy. Speech recognition research is conducted largely through experiments; much of my effort has been invested in performing a large number of different trials. In this work I have greatly benefited from the atmosphere of mutual cooperation in ICSI's Realization Group; there is a great tradition of sharing ideas and implementations. Therefore, it may be instructive to attempt to list my specific contributions concretely.

In the beginning, I created the first scripts for the initial gathering of data about the use of syllables in conversational speech, before the study was later extended and expanded by others. For the work with syllable onsets, I designed and implemented a special purpose decoder, conducted numerous trials with incorporating Michael Shire's (ICSI) syllable onsets and designed alternatives to the special decoder for our distant colleagues to use. For the work with syllable-oriented recognition, I implemented the half-syllable unit representation and trained the syllable-based recognition systems. I conducted numerous experiments with these systems individually and in combination to examine their behavior. I derived and implemented error analysis techniques from the work of other researchers. Brian Kingsbury and Nikki Mirghafori (both of ICSI) helped with the implementation of the more successful combination methods, and Brian also contributed to running some of the later experiments.

This work entailed the implementation of considerable supporting software for incorporating syllable onsets, syllable-based recognition, combining at three different levels and automatic error analysis. These items can be used by others to pursue additional avenues in syllable-based recognition, or to analyze and combine two arbitrary systems. Some of this material is already being integrated into the research projects of colleagues at ICSI.

7.4 Future Extensions

Inevitably, there are parts of the story that remain incomplete.

7.4.1 Further Optimization

As mentioned in Chapter 5, the recognition systems that incorporate syllable-based elements have not been optimized as fully as the mature phoneme-based system. Experimental methods and time constraints limited the individual refinement of the systems. Further optimization possibilities include:

Using Improved Features

A revised version of the modulation spectrogram features is under development by Kingsbury and Greenberg [105]. The latest version of the features outperforms the older version used in this thesis, particularly for the reverberant test case. Using the refined version of the modulation spectrogram features may help to reduce the absolute error rate for the combined systems.

Using Improved Recognition Units

The half-syllable unit is a reasonable and effective starting point for representing syllables in speech recognition, but further study may reveal a more appropriate unit, such as whole syllables, or smaller parts of syllables, as have been used by others.[2] Further work along these lines may improve the performance of the syllable-based systems and therefore of the combined systems overall.

Selecting And Improving a Single Combination Strategy

Exploring combination strategies at different levels with the same systems has limited the degree to which the individual systems could be optimized. The latitude available for improving individual systems within a combination expands as the time between combination points increases. At the frame-level, refinements are restricted to those that do not cause mismatched phone recognition behavior. For direct comparison of error rates, the frame-level constraint limited the design of the system for combining at higher levels.

Selecting a single level of combination would allow extra optimization appropriate at that level. If the combination is to be performed at the frame level, system improvements would be limited to the areas of feature extraction and neural network architecture and training. The improvements should not alter the form of the output probabilities. The syllable level does not require common units since some amount of desynchronization between the recognition outputs of the two systems can be absorbed during the combination process. This means that training labels and recognition units can be independently optimized, so long as the recognition process can still synchronize at the endpoints of syllables. At the whole utterance level, the combining method imposes no optimization constraints. Each system can be optimized in an independent fashion since the recognition systems interact

[2] More information on previous use of different kinds of syllable-based information is described in Chapter 2.

only at the endpoints of an utterance. However, performance results on reverberant tests exhibited the least improvement of the three methods.

7.4.2 Scaling to Larger Vocabulary Tasks

Numbers is a practical and useful corpus for this experimental research, but it possesses a rather small lexicon, comprising only 32 words. The results derived from this small corpus may still extend to larger vocabulary tasks because of the general acoustic character of the corpus. Scaling up to larger tasks will entail addressing the following issues:

Dynamic Syllabification for Syllables Onsets

The extensibility of incorporating syllable-onsets is indicated by the work of colleagues with a 65,000-word vocabulary task [33]. Nonetheless, there are unanswered questions relating to the application of syllable-onsets to larger vocabularies. The pilot studies using onsets derived from knowledge of the correct answers (the "cheating" experiments) indicated a larger potential for improvement than has been realized. A likely impediment is the inadequacy of the current definition for syllable onset and the way it is currently incorporated on the basis of isolated words. Context-dependent syllable models may be useful in addressing this issue.

Syllable-level Combination with Large Vocabularies

For experiments with combining different systems, the best overall error rate was found with combinations at the syllable level. The particular method of combining at the syllable-level (HMM-recombination), however, requires additional implementation issues to be resolved. Even for the Numbers task, a combination of only two systems with limited desynchronization between streams required hundreds of probabilities per frame and word HMMs with thousands of states.

A method that avoids the creation of massive HMMs would ease the computational resource problem. Eric Fosler-Lussier (of ICSI) has been working on a two-level Viterbi decoder [54]. The two-level decoding algorithm [171, 159] was historically set aside in favor of the more efficient single-pass Viterbi decoding algorithm. The two-level approach, however, is more amenable to combination schemes. The likelihood for each word is calculated for every possible time alignment. Thus, combining scores from two systems at the ends of words is straightforward. Combination can be implemented without resorting to HMMs with thousands of states or other complex means, keeping the demand on computational resources at moderate levels. Combining at the ends of syllables within words would require an additional level of processing, because phone probability information would first have to be decoded into syllables, before the syllables could be decoded into words.

The method of combining recognition systems at the frame-level is fully extensible to larger tasks since the procedure merges probabilities at the phone/frame level. The error rates found by combining at the frame level were similar to the best error rates found at the syllable level and required much less implementation effort.

Combining at the utterance level does not involve such a close integration of the decoding procedure. Increasing the size of the vocabulary and the length of the utterances would require longer N-best lists to achieve good performance, and thereby lengthen the execution time of the complex rescoring process. This problem could be mitigated by using phrases or between-pause segments instead of entire utterances.

Number of Unique Syllable Units

Large vocabularies will incur a greater number of distinct syllable-based targets. As the number of different training targets increases, the number of training patterns for each target decreases, potentially reducing the accuracy of the probability estimation process. As the size of the task syllabary increases, issues such as similar sounding syllables and resyllabification phenomena arise. These may exert a much larger negative impact than in Numbers. These problems, however, mirror those of context-dependent units such as triphones. Strategies similar to ones developed for those units can be used to handle the increased complexity of a large number of syllables.

Efficiency and modularity might solve some of the scaling problems of syllable-based approaches. For example, instead of one massive neural network, using several smaller networks, perhaps arranged in a hierarchy of graduated generality, can perhaps be used to manage the complexity. Fritsch suggested and implemented this strategy for context-dependent acoustic models [58].

The Syllable-based Speech Processing Team of the 1997 LVCSR Workshop [67] developed a means of dealing with the problem of augmenting a syllabary for a large vocabulary: they used syllable models for the more frequently occurring words and handled the remainder with standard phoneme-based models. The team reported some success with this method. Such a strategy addresses the need for accommodating unusual syllable types, such as "scrounged" and "strength." Exotic syllable types can be described most straightforwardly in terms of phonemes.

7.4.3 Further Combining

This thesis described the use of syllabic information for ASR in two phases: the incorporation of syllable-onset estimates and the combining of syllable-based systems with phoneme-based systems. One direction for future work is the further combination of these two types of information. As mentioned previously, the syllable onset experiments can be related to the combining experiments at the syllable-level through an interpretation based on syllable-length intervals. Both methods can be said to hypothesize syllable-length intervals in speech and attach various features to the segments. Using syllable-onset information to constrain decoding in the component recognition systems of a combination may show a larger improvement than either method alone. Additional speech unit values could be attached to the same hypothesized, syllable-length interval. This further permutation requires closely matching the syllabary in each of the three constituent recognition systems. Since this question depends on the specific definition of the syllable, further research is necessary.

7.4.4 Parallel and Concurrent Computing

As noted in the introductory chapter, this work arose out of studying the syllable with a view to developing parallel decoding algorithms for a vector microprocessor. The syllable has several properties that are desirable for vector computing: 1) Syllable-based models may be conducive to removing conditional branches during execution and 2) Syllable-based models are a natural organizational unit for reducing redundant computation and defining the search space. Although the work in this thesis does not explore parallel computing further, some of the conclusions of this work are applicable to concurrent processing. Namely, combining information from multiple streams is an obviously concurrent operation. Eric Fosler-Lussier's two-level decoder [54] may map neatly onto a multiple processor machine, since the probabilities of different words (or syllables) are computed independently.

As mentioned in Chapter 3, some recent advances in speech recognition technology have been attributed to general improvements in hardware performance [32]. If this is the case, using parallel and concurrent machines should be highly advantageous to speech recognition research.

7.5 Reflections on the Future of ASR Research

The field of automatic speech recognition is entering a new stage of maturation. As a result, the research paradigm used is undergoing certain transitions. Integrating separate knowledge sources and merging systems that run in parallel will probably play a more substantial role in future investigative directions.

Not long ago, commercial ASR products were limited to fairly simple systems. More complex systems (e.g., large vocabulary, continuous speech) systems were confined to research laboratories. At the same time, the state of the art was basic enough that every researcher could create a personal speech recognition engine from scratch. Since then, research systems have increased dramatically in complexity and size. Without evidence using competitive systems, research results are often regarded as incomplete. Small research groups are currently encountering significant difficulties in developing their own recognition systems that are competitive with the offerings of larger, established players. Many small groups develop their systems with the help of HTK (Hidden Markov Model Tool Kit) [212], the CSLU Speech Toolkit [22], or STRUT (Speech Training and Recognition Unified Tool) [15], which provide many of the processing elements needed.

ASR has also grown into viable commercial products. As the market for speech recognition applications enlarges, consumers will begin to drive the industry. The product features desired by users, rather than basic science interests, will dictate the research agenda of many organizations [128]. For example, the commercial viability of using speech recognition for information retrieval has focused interest, and funding, on the aspects ASR appropriate for this task. Customers will more strongly influence the direction of research than will academia.

A similar evolutionary process occurred in the field of microprocessor design. Not very long ago, microprocessors had only tens of thousands of transistors and research groups

commonly designed and fabricated special purpose chips for local interests. Today, microprocessors are a consumer commodity; directions in new chip designs are driven by market forces. State of the art microprocessors are highly complex with millions of transistors and are continuously increasing in complexity. Researchers in academia very rarely create complete microprocessors from scratch because of the massive expenditure in resources necessary to achieve competitive performance. Instead, they focus on specific angles, such as low-power operation, and build simplified prototypes. Complete chips are most often left to large corporations to fully develop.

In ASR, academic recognition systems can still compete with industrial systems. University systems, such as Cambridge University's HTK and connectionist groups, and Carnegie Mellon University's group, are still among the front runners in organized evaluations [35]. Because state-of-the-art performance is considered critical, research directions have suffered from a certain degree of inertia. Innovative new systems are usually couched in small pilot studies which suffer in comparison to larger, more mature systems. Attempting radical departures from the established mode has become increasingly difficult [17]. There are initiatives aimed at combating this trend: some research organizations attempt to contribute by attaching their work to existing, state-of-the-art systems. Boston University, for example, entered the 1997 ARPA Switchboard Evaluation in collaboration with an industrial partner [149].

Collaborative efforts between research partners will figure more prominently in the future of automatic speech recognition. Joint efforts can be integrated in many different ways. Two common methods are illustrated in this thesis: 1) adding an auxiliary knowledge source to the main recognition system (the syllable onset work) and 2) computing information in parallel and merging the results (the combining of systems work). These kinds of methods have already given researchers an opportunity to take advantage of strengths and soften weaknesses within a flexible structure. Integrating information from various sources is probably the paradigm the human brain uses.

Interfacing among parts in collaborative work is a difficult engineering issue by itself, as shown by the efforts of the Verbmobil engineers. The Verbmobil project in Germany involved, at one point, 29 separate sites with 150 researchers and engineers [21]. The integration of the efforts was a large and time-consuming task, aside from the speech recognition aspects. The syllable can play a part in smoothing many kinds of interactions because of its function as a basic unit. The elemental role which the syllable is believed to play in many separate parts of the human auditory system can possibly be exploited to help with these engineering concerns.

This kind of collaborative research activity can be seen in the compiler community. Modern compilers have also become ponderously large, with many similar performance evaluation issues as speech recognition software. The National Compiler Infrastructure project attempts to address some of these problems: the project uses SUIF (Stanford University Intermediate Format) as a platform for supporting collaboration between compiler researchers [187]. SUIF aims towards a modular architecture that is easily extensible and maintainable. Some ASR research groups have already taken steps towards similar frameworks for speech recognition (for example, the public domain speech recognition technology effort headed by Joe Picone [152]).

Of course, not every research interest can fit into a combination or interface model. Some directions will necessitate the development of complete recognition systems from scratch. Extensive alliances also raise many logistical and political issues. Despite these drawbacks, collaboration through various forms of combining will probably become a more common occurrence in speech recognition research.

7.6 Conclusion

In this thesis I have shown that incorporating syllables into an established automatic speech recognition system can improve continuous speech recognition accuracy and robustness for a small vocabulary corpus. Syllable-oriented recognition extracted a different aspect of the speech signal from the phoneme-oriented recognition which led to greater overall performance when used together. Experiments with Numbers resulted in up to a 24% relative improvement for clean speech and up to a 40% relative gain for reverberant speech.

The improvement in recognition accuracy in the combined systems is attributed to the influence of syllable-based information in creating systems with strengths and weaknesses complementary to those of the baseline, phoneme-based system. In particular, modulation spectrogram features played a large role in creating systems with divergent errors. The use of the half-syllable unit further promoted the dissimilarity of errors, although to a lesser extent than the features. At each level of combination (frame, syllable and utterance), coarse error analyses showed that the system with the largest number of syllable-based design elements was also the system with the lowest error correlation with the baseline system.

The results of these experiments showed that the simplest method, frame-level combination, achieved a large proportion of the maximum gain observed.[3] Syllable-level combining achieved a slightly higher accuracy, but required additional complexity. The syllable-level and utterance-level combining methods admit considerably more possibilities for individualized optimization than does frame-level combination, because only similar outputs can be combined at the frame-level. Combining at the utterance level was the least effective for degraded speech, but the combination of systems across all levels resulted in performance improvements over the baseline.

Speech recognition has become a product in demand; there is considerable motivation to solve the problems that keep speech recognition applications from universal deployment. The work in this thesis uses the syllable unit and combination methods to take a small step towards that goal.

[3] Even the systems combined at the frame-level incorporated signal processing and neural network time spans that were roughly syllabic in length.

Appendix A

Recognition Units

This appendix lists, in table format, the phone set, half-syllable set and canonical pronunciations used in the recognition systems discussed in this thesis.

A.1 ICSI 56 Phoneme Set

ASR Phoneme Symbols[1]					
ICSI56set	IPA	Example	ICSI56set	IPA	Example
pcl	p^o	(p closure)	bcl	b^o	(b closure)
tcl	t^o	(t closure)	dcl	d^o	(d closure)
kcl	k^o	(k closure)	gcl	g^o	(g closure)
p	p	**p**ea	b	b	**b**ee
t	t	**t**ea	d	d	**d**ay
k	k	**k**ey	g	g	**g**ay
ch	$t\widehat{\int}$	**ch**oke	dx	ɾ	dir**t**y
f	f	**f**ish	jh	$d\widehat{ʒ}$	**j**oke
th	θ	**th**in	v	v	**v**ote
s	s	**s**ound	dh	ð	**th**en
sh	∫	**sh**out	z	z	**z**oo
m	m	**m**oon	zh	ʒ	a**z**ure
em	m̩	botto**m**	n	n	**n**oon
ng	ŋ	si**ng**	en	n̩	butto**n**
nx	ɾ̃	wi**nn**er	el	l̩	bott**le**
l	l	**l**ike	r	r	**r**ight
w	w	**w**ire	y	j	**y**es
hh	h	**h**ay	hv	ɦ	a**h**ead
er	ɝ	b**ir**d	axr	ɚ	butt**er**
iy	i	b**ee**t	ih	ɪ	b**i**t
ey	e	b**ai**t	eh	ɛ	b**e**t
ae	æ	b**a**t	aa	ɑ	f**a**ther
ao	ɔ	b**ough**t	ah	ʌ	b**u**t
ow	o	b**oa**t	uh	ʊ	b**oo**k
uw	u	b**oo**t	ix	ɨ	deb**i**t
aw	$ɑ^w$	ab**ou**t	ay	$ɑ^y$	b**i**te
oy	$ɔ^y$	b**oy**	ax	ə	**a**bout
h#		(silence)			

[1] Table derived from table for TIMIT, from Eric Fosler-Lussier, originally from Charles Wooters.

A.2 Half-Syllable Units

Half-Syllable Units[2]			
First Halves		Second Halves	
s-eh	v-ih	eh	ih-n-tcl
t-iy	v-ax	iy-n	ax-n
n-ah	n-tcl-t-iy	eh-v	iy
t-ih	v-ah	ah-n-tcl	ah
f-ih	ey	ih-f-tcl	ih-n
eh	w-ah	ey-tcl	ax-n-tcl
th-er	tcl-th-er	er-tcl	ah-n
s-ah	v-eh	eh-n	eh-v-n
ih	d-iy	ih	ey-dcl
s-ih	s-tcl-t-iy	ih-kcl	ey
d-tcl-t-iy	hh-ah	eh-dcl-d	ih-kcl-k
d-r-eh	hv-ah	ih-dcl-d	ih-dcl-t
d-r-ih	d-er	er-dcl-d	ah-n-dcl
d-eh	d-ih	eh-dcl	er-dcl
d-ow	r-eh	ow	ih-dcl
tcl-t-w-ae	t-w-eh	eh-l-f	ae-l
tcl-t-w-ow	tcl-t-w-ah	ow-v	ah-l-v
tcl-t-w-eh	n-ay	eh-l-v	ay
n-iy	n-dcl-d-iy	ay-n	ih-n-dcl
v-uh	n-d-iy	uh-n-dcl	eh-n-tcl
n-ih	ah	er	ax
z-ih	tcl-t-w-ax	ey-tcl-t	ey-t
th-ih	r-iy	ih-kcl-k-s	ay-n-tcl
s-tcl-t-ih	f-ow	ow-r	ay-tcl
t-eh	tcl-t-eh	ih-k	ao-r
r-ah	z-iy	ow-r-dcl	ow-r-tcl
t-w-ah	iy	ao-tcl	ao
r-ow	th-r-t-iy	ow-r-dcl	ow-r-tcl
tcl-th-r-iy	th-r-iy	uw	
f-ao	tcl-t-uw		
l-eh	l-ah		
ax	ow		
f-ay	s-t-iy		

[2]Half-syllable units were derived from word pronunciations, where each word was partitioned into syllables automatically. Each syllable was divided at the middle of the nucleus.

Since the definition of the syllable is poorly specified, any list of candidate syllables probably has some linguistic inconsistencies.

A.3 Numbers Pronunciations (*canonical* syllable based)

Canonical Syllable Pronunciations[3]	
oh	ow
zero	z-ih r-ow
one	w-ah-n
two	t-uw
three	th-r-iy
four	f-ao-r
five	f-ay-v
six	s-ih-k-s
seven	s-eh v-en
eight	ey-t
nine	n-ay-n
ten	t-eh-n
eleven	ix l-eh v-en
twelve	t-w-eh-l-v
thirteen	th-er t-iy-n
fourteen	f-ao-r t-iy-n
fifteen	f-ih-f t-iy-n
sixteen	s-ih-k-s t-iy-n
seventeen	s-eh v-en t-iy-n
eighteen	ey t-iy-n
nineteen	n-ay-n t-iy-n
twenty	t-w-eh-n-t-iy
thirty	th-er t-iy
forty	f-ao-r t-iy
fifty	f-ih-f t-iy
sixty	s-ih-k-s t-iy
seventy	s-eh v-en t-iy
eighty	ey t-iy
ninety	n-ay-n t-iy
hundred	h-ah-n d-r-ax-d
[uh]	ax
[um]	ax-m

[3]Canonical pronunciation and syllabification in ICSI56 phonetic orthography. Syllabification approximated from CELEX data.

Bibliography

[1] Takayuki Arai and Steven Greenberg. The temporal properties of spoken Japanese are similar to those of English. In *Eurospeech*, Rhodes, Greece, September 1997. ESCA.

[2] Aristotle. Categories. http://classics.mit.edu, 350 B.C.E. Translated by E. M. Edghill.

[3] W. C. Athas and C. L. Seitz. Multicomputers: Message-passing concurrent computers. *IEEE Computer*, 21(8):9–24, August 1988.

[4] Steve Austin, Richard Schwartz, and Paul Placeway. The forward-backward search algorithm. In *ICASSP*, volume 1, pages 697–700, Toronto, May 1991. IEEE.

[5] Carlos Avendano, Sangita Tibrewala, and Hynek Hermansky. Multiresolution channel normalization for ASR in reverberant environments. In *Eurospeech*, volume 3, pages 1107–1110, Rhodes, Greece, September 1997. ESCA.

[6] R. H. Baayen, R. Piepenbrock, and H. van Rijn. The CELEX lexical database. cdrom, 1993.

[7] L. Bahl, P. Cohen, A. Cole, F. Jelinek, B. Lewis, and R. Mercer. Further results on the recognition of a continuously read natural corpus. In *ICASSP*, volume 3, pages 872–875, Denver, Colorado, April 1980. IEEE.

[8] L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny. A fast match for continuous speech recognition using allophonic models. In *ICASSP*, volume 1, pages 17–20, San Francisco, California, March 1992. IEEE.

[9] L. Bahl, P. Gopalakrishan, D. Kanevsky, and D. Nahamoo. Matrix fast match: A fast method for identifying a short list of candidate words for decoding. In *ICASSP*, volume 1, pages 345–348, Glasgow, Scotland, May 1989. IEEE.

[10] L. Bahl, F. Jelinek, and R. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.

[11] James K. Baker. The DRAGON system- An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):24–29, February 1975.

[12] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, May 1967.

[13] Jon Atli Benediktsson, Johannes R. Sveinsson, Okan K. Ersoy, and Phillip H. Swain. Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, 8(1):54–64, January 1997.

[14] Christopher M. Bishop. *Neural Networks for Pattern Recognition*, chapter 6.6.1, pages 226–228. Clarendon Press, New York, New York, 1995. Interpretation of hidden units.

[15] Jean-Marc Boite. Speech training and recognition unified tool, 1998. More information can be found at http://tcts.fpms.ac.be/speech/strut.html.

[16] Antonio Bonafonte, Rafael Estany, and Eugenio Vives. Study of subword units for Spanish speech recognition. In *Eurospeech*, volume 3, pages 1607–1610, Madrid, Spain, September 1995. ESCA.

[17] Hervé Bourlard. Towards increasing speech recognition error rates. In *Eurospeech*, volume 2, pages 883–893, Madrid, Spain, September 1995. ESCA.

[18] Hervé Bourlard, Bart D'hoore, and Jean-Marc Boite. Optimizing recognition and rejection performance in wordspotting systems. In *ICASSP*, volume 1, pages 373–376, Adelaide, South Australia, April 1994. IEEE.

[19] Hervé Bourlard and Stéphane Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *ICSLP*, volume 1, pages 426–429, Philadephia, Pennsylvania, October 1996.

[20] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Press, 1994.

[21] Thomas Bub, Wolfgang Wahlster, and Alex Waibel. Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In *ICASSP*, volume 1, pages 71–74, Munich, Germany, April 1997. IEEE.

[22] CSLU speech toolkit, 1998. More information can be found at http://cse.ogi.edu/CSLU/toolkit/toolkit.html.

[23] Lin Lawrance Chase. Blame assignment for errors made by large vocabulary speech recognizers. In *Eurospeech*, volume 3, pages 1563–1566, Rhodes, Greece, September 1997. ESCA.

[24] Lin Lawrance Chase. *Error-Responsive Feedback Mechanisms for Speech Recognizers*. PhD thesis, Carnegie Mellon University, The Robotics Institute, Pittsburgh, Pennsylvania, April 1997.

[25] Colin Cherry and Roger Wiley. Speech communication in very noisy environments. *Nature*, 214:1164, June 1967.

[26] Kenneth W. Church. Phonological parsing and lexical retrieval. In Uli H. Frauenfelder and Lorraine Komisarjevsky Tyler, editors, *Spoken Word Recognition*, Cognition Special Issues, chapter 3, pages 53–69. MIT Press, 1987.

[27] Kenneth W. Church and William A. Gale. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech and Language*, (5):19–54, 1991.

[28] John Clark and Colin Yallop. *Phonetics and Phonology*, chapter 5, pages 124–127, 287. Basil Blackwell, Ltd., Cambridge, Massachusetts, 1990.

[29] George N. Clements and Samuel Jay Keyser. *CV Phonology: A Generative Theory of the Syllable*. The MIT Press, Cambridge, Massachusetts, 1983.

[30] R. A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. In *Eurospeech*, volume 1, pages 821–824, September 1995.

[31] Ron Cole, Lynette Hirschman, Les Atlas, Mary Beckman, Alan Biermann, Marcia Bush, Mark Clements, Jordan Cohen, Oscar Garcia, Brian Hanson, Hynek Hermansky, Steve Levinson, Kathy McKeown, Nelson Morgan, David G. Novick, Mari Ostendorf, Sharon Oviatt, Patti Price, Harvey Silverman, Judy Spitz, Alex Waibel, Clifford Weinstein, Steve Zahorian, and Victor Zue. The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21, January 1995.

[32] Richard Comerford, John Makhoul, and Richard Schwartz. The voice of the computer is heard in the land and it listens too! *IEEE Spectrum*, 34(12):39–47, December 1997.

[33] G. Cook and T. Robinson. Transcribing Broadcast News with the 1997 Abbot system. In *ICASSP*, Seattle, Washington, April 1998. IEEE.

[34] G. D. Cook, D. J. Kershaw, J. D. M. Christie, and A. J. Robinson. Transcription of Broadcast Television and Radio News: the 1996 Abbot system. In *DARPA Speech Recognition Workshop*, Westfields Internatinal Conference Center, Chantilly, Virginia, February 1997. DARPA.

[35] Coordinated by National Institute of Standards and Technology. *Conversational Speech Recognition Workshop DARPA Hub-5e Evaluation*, Baltimore, Maryland, May 1997.

[36] M. Cravero, R. Pieraccini, and F. Raineri. Definition and evaluation of phonetic units for speech recognition by hidden Markov models. In *ICASSP*, volume 3, pages 2235–2238, Tokyo, Japan, April 1986. IEEE.

[37] Anne Cutler, Sally Butterfield, and John N. Williams. The perceptual integrity of syllabic onsets. *Journal of Memory and Language*, 26:406–418, 1987.

[38] Anne Cutler, Jacques Mehler, Dennis Norris, and Juan Segui. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25:385–400, 1986.

[39] Walter Daeleman and Antal van den Bosch. Generalization performance of back-propagation learning on a syllabification task. In M.F.J. Drossaers and A Nijholt, editors, *Proceedings of TWLT3: Connectionism and Natural Language Processing*, pages 27–37, University of Twente, 1992.

[40] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, August 1980.

[41] Renato De Mori and Michael Galler. The use of syllable phonotactics for word hypothesization. In *ICASSP*, volume 2, pages 877–880, Atlanta, Georgia, May 1996. IEEE.

[42] Renato De Mori and Giovanna Giordano. A parser for segmenting continuous speech into pseudo-syllabic nuclei. In *ICASSP*, volume 3, pages 876–879, Denver, Colorado, April 1980. IEEE.

[43] John R. Deller, Jr., John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, New York, 1993.

[44] N. Rex Dixon and Harvey F. Silverman. The 1976 modular acoustic processor. *IEEE Transactions of Acoustics, Speech and Signal Processing*, ASSP-25(5):367–379, October 1977.

[45] Stéphane Dupont, Hervé Bourlard, and Christophe Ris. Using multiple time scales in a multi-stream speech recognition system. In *Eurospeech*, pages 3–6, Rhodes, Greece, October 1997.

[46] Harold T. Edwards. *Applied Phonetics: The Sounds of American English*. Singular Publishing Group, Inc., San Diego, California, 1992.

[47] Dan Ellis. SYLLIFY. Inhouse software at ICSI, 1996. Tcl/TK interface for Fisher's TSYLB2 program.

[48] Lee D. Erman and Victor R. Lesser. The Hearsay-II speech understanding system: A tutorial. In W. A. Lea, editor, *Trends in Speech Recognition*, chapter 16, pages 361–381. Speech Science Publications, Apple Valley, MN, 1980. Reprinted in (Waibel and Lee, 1990).

[49] Kevin R. Farrell, Ravi P. Ramachandran, and Richard J. Mammone. An analysis of data fusion methods for speaker verification. In *ICASSP*, Seattle, Washington, April 1998. IEEE.

[50] Michael Finke and Alex Waibel. Flexible transcription alignment. In *ASRU*, pages 34–40, Santa Barbara, CA, December 1997. IEEE.

[51] Bill Fisher. TSYLB2. Source code available through FTP from NIST, 1995.

[52] Eric Fosler. Automatic learning of a model for word pronunciations: Status report. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*. NIST, May 13–15 1997.

[53] Eric Fosler. Evidence for syntactic and semantic repair effects in auditory processing. Linguistics 220 Class Project, May 1997.

[54] Eric Fosler-Lussier. TWO-LEVEL DECODER. Inhouse software at ICSI, 1998. An implementation of the two-level decoding algorithm with the capability to combine intermediate scores.

[55] Uli H. Frauenfelder. The interface between acoustic-phonetic and lexical processing. In M. E. H. Schouten, editor, *The Auditory Processing of Speech– From Sounds to Words*, number 10 in Speech Research, pages 325–338. Mouton de Gruyter, New York, 1992.

[56] Norman R. French, Charles W. Carter, Jr., and Walter Koenig, Jr. The words and sounds of telephone conversations. *The Bell System Technical Journal*, IX:290–325, April 1930.

[57] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996.

[58] Jürgen Fritsch. ACID/HNN: A framework for heirarchical connectionist acoustic modeling. In Sadaoki Furui, B.-H. Juang, and Wu Chou, editors, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 164–171, Santa Barbara, California, December 1997. IEEE.

[59] Osamu Fujimura. Syllable as a unit of speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):82–87, February 1975.

[60] Osamu Fujimura. Syllables as concatenated demisyllables and affixes. *Journal of the Acoustical Society of America*, 59(Suppl. 1):S55, Spring 1976.

[61] Osamu Fujimura. Demisyllables as sets of features: comments on Clements's paper. In John Kingston and Mary E. Beckman, editors, *Between the grammar and physics of speech*, number 1 in Papers in laboratory phonology, chapter 18, pages 334–340. Cambridge University Press, Cambridge, UK, 1990.

[62] Osamu Fujimura. Syllable timing computation in the C/D model. In *ICSLP*, pages 519–522, Yokohama, Japan, September 1994.

[63] Osamu Fujimura. Prosodic organization of speech based on syllables: The C/D model. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 3, pages 10–17, Stockholm, Sweden, August 1995.

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

[65] Sadaoki Furui and Chin-Hui Lee. Robust speech recognition– An overview. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, page 93, Snowbird, Utah, December 1995. IEEE.

[66] M. J. F. Gales and S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *ICASSP*, volume 1, pages 233–236, San Francisco, California, March 1992. IEEE.

[67] Aravind Ganapathiraju, Vaibhava Goel, Joseph Picone, Andres Corrada, George Doddington, Katrin Kirchhoff, Mark Ordowski, and Barbara Wheatley. Syllable– a promising recognition unit for LVCSR. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, December 1997. IEEE.

[68] Jean-Luc Gauvain. A syllable-based isolated word recognition experiment. In *ICASSP*, volume 1, pages 57–60, Tokyo, Japan, April 1986. IEEE.

[69] Dan Gildea and Eric Fosler-Lussier. Numbers lexicon. Inhouse lexicon specification for the Numbers corpus., 1996.

[70] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP*, volume 1, pages 517–520, San Francisco, California, March 1992. IEEE.

[71] David Graff. The 1996 Broadcast News Speech and Language-Model Corpus. In *DARPA Speech Recognition Workshop*, Westfields International Conference Center, Chantilly, Virginia, February 1997. DARPA.

[72] P. D. Green, L. A. Boucher, N. R. Kew, and A. J. H. Simons. The SYLK project final report. By private communication with P. Green, 1993.

[73] P. D. Green, N. R. Kew, and D. A. Miller. Speech representations in the SYLK recognition project. In M. P. Cooke, S. W. Beet, and M. D. Crawford, editors, *Visual Representation of Speech Signals*, chapter 26, pages 265–272. John Wiley, 1993.

[74] Steven Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Workshop on the Auditory Basis of Speech Perception*, pages 1–8, Keele, United Kingdom, July 1996. ESCA.

[75] Steven Greenberg. On the origins of speech intelligibility. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-a-Mousson, France, April 1997. ESCA.

[76] Steven Greenberg. The Switchboard transcription project. In Frederick Jelinek, editor, *1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, number 24 in Research Notes, Baltimore, Maryland, April 1997. Center for Language and Speech Processing, Johns Hopkins University.

[77] Steven Greenberg. Personal communication., 1998.

[78] Steven Greenberg. Speaking in shorthand– a syllable-centric perspective for understanding pronunciation variation. In *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kekrade, Netherlands, May 1998. ESCA.

[79] Steven Greenberg, Joy Hollenback, and Dan Ellis. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *ICSLP*, volume Supplement, pages S24–S27, Philadelphia, Pennsylvania, October 1996.

[80] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, volume 3, pages 1647–1650, Munich, Germany, April 1997. IEEE.

[81] François Grosjean and James Paul Gee. Prosodic structure and spoken word recognition. In Uli H. Frauenfelder and Lorraine Komisarjevsky Tyler, editors, *Spoken Word Recognition*, Cognition Special Issue, pages 135–155. MIT Press, Cambridge, Massachusetts, 1987.

[82] Michael Hammond. Syllable parsing in English and French. Available through http://aruba.ccit.arizona.edu/ hammond, May 1995.

[83] Alfred Hauenstein. The syllable re-revisited. Technical Report TR-96-035, ICSI, August 1996.

[84] Alfred Hauenstein. Using syllables in a hybrid HMM-ANN recognition system. In *Eurospeech*, volume 3, pages 1203–1206, Rhodes, Greece, September 1997. ESCA.

[85] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[86] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[87] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, January 1994.

[88] Mike Hochberg. YO. Software package, WERNICKE distribution., August 1993. Viterbi decoder in use at ICSI.

[89] Zhihong Hu, Johan Schalkwyk, Etienne Barnard, and Ronald Cole. Speech recognition using syllable-like units. In *ICSLP*, volume 2, pages 1117–1120, Philadephia, Pennsylvania, October 1996.

[90] X. D. Huang and M. A. Jack. Semi-continuous hidden Markov models for speech signals. *Computer Speech and Language*, 3(3):239–252, July 1989. Reprinted in (Waibel and Lee, 1990).

[91] Melvyn J. Hunt, Matthew Lennig, and Paul Mermelstein. Use of dynamic programming in a syllable-based continuous speech recognition system. In David Sankoff and Joseph B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, chapter 5, pages 163–188. Addison-Wesley Publishing Company, Inc., Reading Massachusetts, 1983.

[92] M.J. Hunt, M. Lennig, and P. Mermelstein. Experiments in syllable-based recognition of continuous speech. In *ICASSP*, volume 3, pages 880–883, Denver, Colorado, April 1980. IEEE.

[93] F. Jelinek. Fast sequential decoding algorithm using a stack. *IBM J. Res. Develop.*, 13:675–685, November 1969.

[94] F. Jelinek. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, chapter 8, pages 450–506. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

[95] F. Jelinek, L. R. Bahl, and R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21:250–256, May 1975.

[96] James J. Jenkins, Winifred Strange, and Salvatore Miranda. Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America*, 95(2):1030–1043, February 1994.

[97] John T. Jensen. *English Phonology*, volume 99 of *Series IV- Current Issues in Linguistic Theory*, chapter 3, pages 47–76. John Benjamins Publishing Company, Philadelphia, 1993.

[98] M. Jones and P.C. Woodland. Modelling syllable characteristics to improve a large vocabulary continuous speech recogniser. In *ICSLP*, volume 4, pages 2171–2174, Yokohama, Japan, September 1994.

[99] Rhys James Jones, Simon Downey, and John S. Mason. Continuous speech recognition using syllables. In *EuroSpeech*, volume 3, pages 1171–1174, Rhodes, Greece, September 1997. ESCA.

[100] Dan Jurafsky and Nikki Mirghafori. ICSI speech recognition system. Inhouse document at ICSI, 1995.

[101] Daniel Kahn. *Syllable-based Generalizations in English Phonology*. Outstanding Dissertations in Linguistics. Garland Publishing, New York, 1980.

[102] P. Kenny, R. Hollan, V. Gupta, M Lennig, P Mermelstein, and D. O'Shaughnessy. A^*-admissible heuristics for rapid lexical access. In *ICASSP*, volume 1, pages 689–692, Toronto, Canada, May 1991. IEEE.

[103] Michael Kenstowicz and Charles Kisseberth. *Generative Phonology*. Harcourt Brace Jovanovich, Orlando, 1979.

[104] Brian E. D. Kingsbury. Personal communication, March 1998. Modulation spectrogram processing diagram.

[105] Brian E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*. PhD thesis, UC Berkeley, 1998. To be published.

[106] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP*, volume 2, pages 1259–1262, Munich, Germany, April 1997. IEEE.

[107] Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 1998. In press.

[108] Katrin Kirchhoff. Phonologically structured HMMs for speech recognition. In *Second Meeting of the ACL SIG in Computational Phonology*, pages 45–49, Santa Cruz, California, June 1996. ACL.

[109] Katrin Kirchhoff. Syllable-level desynchronisation of phonetic features for speech recognition. In *ICSLP*, volume 4, pages 2274–2276, Philadephia, Pennsylvania, October 1996.

[110] Katrin Kirchhoff. Statistical analysis of the VERBMOBIL corpus. Unpublished Memo, April 1997.

[111] Dennis H. Klatt. Review of the ARPA speech understanding project. *The Journal of the Acoustical Society of America*, 62(6):1324–1366, December 1977. Reprinted in (Waibel and Lee, 1990).

[112] H. Klemm, F. Class, and U. Kilian. Word- and phrase spotting with syllable-based garbage modelling. In *Eurospeech*, volume 3, pages 2157–2160, Madrid, Spain, September 1995. ESCA.

[113] Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, New York, third edition, 1993.

[114] Wayne A. Lea, Mark F. Medress, and Toby E. Skinner. A prosodically guided speech understanding strategy. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):30–38, February 1975.

[115] Kai-Fu Lee, Hsiao-Wuen Hon, and Raj Reddy. A overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(1):35–45, January 1990.

[116] Lin-Shan Lee. Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine*, pages 63–100, July 1997.

[117] Victor R. Lesser, Richard D. Fennell, Lee D. Erman, and D. Raj Reddy. Organization of the Hearsay II speech understanding system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):11–23, February 1975.

[118] Franklin Mark Liang. *Word Hy-phen-a-tion By Com-pu-ter*. PhD thesis, Stanford University, June 1983.

[119] Sung-Chien Lin, Lee-Feng Chien, Keh-Jiann Chen, and Lin-Shan Lee. A syllable-based very-large-vocabulary voice retrieval system for Chinese databases with textual attributes. In *Eurospeech*, volume 1, pages 203–206, Madrid, Spain, September 1995. ESCA.

[120] Richard Lippmann. Speech perception by humans and machines. In *Workshop on the Auditory Basis of Speech Perception*, pages 309–316, Keele, United Kingdom, July 1996. ESCA.

[121] E. Lleida, J.B. Mariño, J. Salavedra, and A. Bonafonte. Syllabic fillers for Spanish HMM keyword spotting. In *ICSLP*, volume 1, pages 5–8, Banff, Alberta, Canada, October 1992.

[122] B. T. Lowerre and D. R. Reddy. The HARPY speech understanding system. In W. A. Lea, editor, *Trends in Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1980.

[123] Dominic W. Massaro. Preperceptual auditory images. *Journal of Experimental Psychology*, 85(3):411–417, 1970.

[124] Dominic W. Massaro. Preperceptual images, processing time and perceptual units in auditory perception. *Psychological Review*, 79(2):124–145, 1972.

[125] Dominic W. Massaro. Perceptual units in speech recognition. *Journal of Experimental Psychology*, 102(2):199–208, 1974.

[126] Shoichi Matsunaga, Takeshi Matsumura, and Harald Singer. Continuous speech recognition using non-uniform unit based acoustic and language models. In *Eurospeech*, volume 3, pages 1619–1622, Madrid, Spain, September 1995. ESCA.

[127] Jacques Mehler, Jean Yves Dommergues, and Uli Frauenfelder. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20:298–305, 1981.

[128] William S. Meisel. State of the art: Applications. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 29–44, Snowbird, Utah, December 1995. IEEE.

[129] Paul Mermelstein. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am*, 58(4):880–883, October 1975.

[130] Joanne L. Miller and Peter D. Eimas. Observations on speech perception, its development, and the search for a mechanism. In Judith C. Goodman and Howard C. Nusbaum, editors, *The Development of Speech Perception:The Transition from Speech Sounds to Spoken Words*, chapter 2, pages 37–55. The MIT Press, Cambridge, Massachusetts, 1994.

[131] N. Morgan, H. Hermansky, and H.G. Hirsch. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In *ICASSP*, volume 1, pages 83–86, Minneapolis, Minnesota, April 1993. IEEE.

[132] Nelson Morgan. *Room Acoustics Simulation with Discrete-Time Hardware*. PhD thesis, UC Berkeley, 1980.

[133] Nelson Morgan and Hervé Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.

[134] Nelson Morgan, Hervé Bourlard, Steve Greenberg, and Hynek Hermansky. Stochastic perceptual auditory-event-based models for speech recognition. In *ICSLP*, pages 1943–1946, Yokohama, Japan, September 1994.

[135] Nelson Morgan and Hynek Hermansky. RASTA extensions: Robustness to additive and convolutional noise. In *Proceedings of the Workshop on Speech Processing in Adverse Conditions*, Cannes, France, November 1992.

[136] Nelson Morgan, Su-Lin Wu, and Hervé Bourlard. Digit recognition with stochastic perceptual speech models. In *Eurospeech*, Madrid, Spain, September 1995.

[137] Hy Murveit, John Butzberger, Vassilios Digalakis, and Mitch Weintraub. Large-vocabulary dictation using SRI's *Decipher*^{TM} speech recognition system: Progressive search techniques. In *ICASSP*, volume 2, pages 319–322, Minneapolis, Minnesota, April 1993. IEEE.

[138] H. Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(2):263–271, April 1984.

[139] H. Ney and X. Aubert. A word graph algorithm for large vocabulary, continuous speech recognition. In *ICSLP*, pages 1355–1358, Yokohama, Japan, September 1994.

[140] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10,000-word continuous speech recognition. In *ICASSP*, volume 1, pages 9–12, San Francisco, California, March 1992. IEEE.

[141] NIST. Continuous speech recognition corpus, September 1993. National Institute of Standards and Technology Speech.

[142] NIST. SCLITE version 1.3. Distributed by NIST, March 1996. Scores speech recognition system output.

[143] Dennis Norris and Anne Cutler. The relative accessibility of phonemes and syllables. *Perception and Psychophysics*, 43(6):541–550, 1988.

[144] Lynne C. Nygaard and David B. Pisoni. Speech perception: New directions in research and theory. In Joanne L. Miller and Peter D. Eimas, editors, *Speech, Language and Communication*, volume 11 of *Handbook of Perception and Cognition*, chapter 3, pages 63–96. Academic Press, San Diego, California, 2 edition, 1995.

[145] J.J. Odell, V. Valtchev, P. C. Woodland, and S.J. Young. A one pass decoder design for large vocabulary recognition. In *Proceedings ARPA Human Language Technology Workshop*, pages 405–410. ARPA, Morgan Kaufmann, March 1994.

[146] Martin Oerder and Hermann Ney. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP*, volume 2, pages 119–122, Minneapolis, Minnesota, April 1993. IEEE.

[147] Ralph N. Ohde and Donald J. Sharf. *Phonetic Analysis of Normal and Abnormal Speech*. MacMillan Publishing Company, New York, 1992.

[148] Douglas O'Shaughnessy. *Speech Communication*, chapter 5, pages 164–203. Addison-Wesley Publishing Company, Reading, Massachusetts, 1987.

[149] Mari Ostendorf, R. Bates, J. Hancock, R. Iyer, A. Kannan, I. Shaik, and M. Siu. The Boston University LVSCR benchmark system. In *Conversational Speech Recognition Workshop DARPA Hub-5E Evaluation*. NIST, May 1997.

[150] Douglas B. Paul. Algorithms for an optimal A^* search and linearizing the search in the stack decoder. In *ICASSP*, volume 1, pages 693–696, Toronto, Canada, May 1991. IEEE.

[151] Barbara Peskin, Larry Gillick, Natalie Liberman, Mike Newman, Paul van Mulbregt, and Steven Wegmann. Progress in recognizing conversational telephone speech. In *ICASSP*, volume 3, pages 1811–1814, Munich, Germany, April 1997. IEEE.

[152] Joe Picone. Public domain speech recognition technology. Personal communication., 1998. Newly established project.

[153] D. B. Pisoni, T. D. Carrell, and S. J. Gans. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, 34(4):314–322, 1983.

[154] B. Plannerer and B. Ruske. Recognition of demisyllable based units using semicontinuous hidden Markov models. In *ICASSP*, pages I581–I584, San Francisco, California, March 1992.

[155] B. Plannerer and B. Ruske. A continuous speech recognition system using phonotactic constraints. In *Eurospeech*, pages 859–862, Berlin, Germany, September 1993.

[156] Patti Price, William M. Fisher, Jared Bernstein, and David S. Pallett. The DARPA 1000-word Resource Management database for continuous speech recognition. In *ICASSP*, volume 1, pages 651–654, New York, New York, April 1988. IEEE.

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

[158] L. Rabiner. Applications of speech recognition in the area of telecommunications. In *ASRU*, pages 501–510, Santa Barbara, CA, December 1997. IEEE.

[159] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*, chapter 7.3, pages 395–400. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[160] L. Rabiner, B.-H. Juang, S. Levinson, and M. Sondhi. Recognition of isolated digits using hidden Markov Models with continuous mixture densities. *AT&T Technical Journal*, 64(6):1211–1234, July-August 1985.

[161] WORDSCORE. Inhouse software at ICSI, February 1997. Scores speech recognition system output.

[162] W. Reichl and G. Ruske. Syllable segmentation of continuous speech with artificial neural networks. In *Eurospeech*, pages 1771–1774, Berlin, Germany, September 1993.

[163] Steve Renals and Mike Hochberg. Decoder technology for connectionist large vocabulary speech recognition. Technical Report CUED/F-INFENG/TR.186, Cambridge University Engineering Department, September 1995.

[164] Steve Renals and Mike Hochberg. Efficient search using posterior phone probability estimates. In *ICASSP*, volume 1, pages 596–603, Detroit, Michigan, May 1995. IEEE.

[165] Steve Renals and Mike Hochberg. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *ICASSP*, volume 1, pages 149–152, Atlanta, Georgia, May 1996. IEEE.

[166] Tony Robinson. LATTICE2NBEST. Part of SLIB package, 1997.

[167] Aaron E. Rosenberg, Lawrence R. Rabiner, Jay G. Wilpon, and Daniel Kahn. Demisyllable-based isolated word recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(3):713–726, June 1983.

[168] Paul Rozin, Susan Poritsky, and Raina Sotsky. American children with reading problems can easily learn to read English represented by Chinese characters. *Science*, 171(3977):1264–1267, March 1971.

[169] G. Ruske, B. Plannerer, and T. Schultz. Stochastic modeling of syllable-based units for continuous speech recognition. In *ICSLP*, pages 1503–1506, Banff, Canada, October 1992.

[170] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*, chapter 3–4. Prentice Hall, Englewood Cliffs, NJ, 1994.

[171] H. Sakoe. Two-level DP-matching– a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(6):588–595, December 1979.

[172] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):43–49, February 1978.

[173] H. B. Savin and T. G. Bever. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, 9:295–302, 1970.

[174] Florien Schiel. Verbmobil concurrent models. Private communication., 1997.

[175] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul. Improved hidden Markov modeling of phonemes for continuous speech recognition. In *ICASSP*, volume 3, page 35.6, San Diego, California, March 1984. IEEE.

[176] Richard Schwartz and Steve Austin. A comparison of several approximate algorithms. In *ICSLP*, volume 1, pages 701–704, Toronto, Canada, May 1991. IEEE.

[177] Richard Schwartz and Yen-Lu Chow. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses. In *ICASSP*, volume 1, pages 81–83, Albuquerque, New Mexico, April 1990. IEEE.

[178] Richard Schwartz, Jack Klovstad, John Makhoul, and John Sorensen. A preliminary design of a phonetic vocoder based on a diphone model. In *ICASSP*, volume 1, pages 32–35, Denver, Colorado, April 1980. IEEE.

[179] Holger Schwenk. Using boosting to improve a hybrid HMM/neural network speech recognizer. Oral presentation at "Machines that Learn" workshop, Snowbird Utah., April 1998.

[180] J. Segui, U. Frauenfelder, and J. Mehler. Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, 72:471–477, 1981.

[181] Juan Segui. The syllable: A basic perceptual unit in speech processing. In Herman Bouma and Don G. Bouwhius, editors, *Attention and Performance X: Control of Language Processes, Proceedings of the Tenth International Sumposium on Attention and Performance*, pages 165–181, Hillsdale, New Jersey, 1984. Lawrence Erlbaum Associates.

[182] Juan Segui, Emmanuel Dupoux, and Jacques Mehler. The role of the syllable in speech segmentation, phoneme identification and lexical access. In Gerry Altmann, editor, *Cognitive Models of Speech Processing*, chapter 12, pages 263–280. MIT Press, 1990.

[183] William Shakespeare. The Tempest. Online version. From the Complete Works of William Shakespeare. http://the-tech.mit.edu/Shakespeare/works.html.

[184] Michael L. Shire. Syllable onset detection from acoustics. Master's thesis, UC Berkeley, 1997.

[185] Frank K. Soong and Eng-Fong Huang. A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition. In *ICASSP*, volume 1, pages 705–708, Toronto, Canada, May 1991. IEEE.

[186] Switchboard corpus: Recorded telephone conversations. Produced by NIST, sponsored by DARPA, October 1992.

[187] SUIF compiler system, 1998. More information can be found at http://www-suif.stanford.edu.

[188] David L. Thomson. Ten case studies of the effect of field conditions on speech recognition errors. In *ASRU*, pages 511–518, Santa Barbara, CA, December 1997. IEEE.

[189] Neil Todd. Towards a theory of the principal monaural pathway: Pitch, time and auditory grouping. In Willam Ainsworth and Steven Greenberg, editors, *Workshop on the Auditory Basis of Speech Perception*, pages 216–221, Keele University, United Kingdom, July 1996. ESCA.

[190] Neil Todd and Christopher Lee. A sensory-motor theory of speech perception: Implications for learning, representation and recognition. In Steven Greenberg and William Ainsworth, editors, *Listening to Speech: An Auditory Perspective*. Oxford University Press, 1998. To be published.

[191] Mikio Tohyama. Response statistics of rooms. In Malcolm J. Crocker, editor, *Encyclopedia of Acoustics*, volume 2, chapter 77, pages 913–923. John Wiley and Sons, Inc., New York, New York, 1997.

[192] Rebecca Treiman and Andrea Zukowski. Toward an understanding of English syllabification. *Journal of Memory and Language*, 29(1):66–85, February 1990.

[193] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *ICASSP*, volume 2, pages 845–848, Albuquerque, New Mexico, April 1990. IEEE.

[194] A. P. Varga and R. K. Moore. Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition. In *Eurospeech*, volume 3, pages 1175–1178, Genova, Italy, September 1991. ESCA.

[195] Klara Vicsi and Attila Vig. Text independent neural network/rule based hybrid continuous speech recognition. In *Eurospeech*, volume 3, pages 2201–2204, Madrid, Spain, September 1995. ESCA.

[196] Alex Waibel. *Prosody and Speech Recognition*. Research Notes in Artificial Intelligence. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1988.

[197] Richard M. Warren. Identification times for phonemic components of graded complexity for spelling of speech. *Perception and Psychoacoustics*, 9(4):345–349, 1971.

[198] Richard M. Warren. Perceptual processing of speech and other perceptual patterns: Some similarities and differences. In Steven Greenberg and William Ainsworth, editors, *Listening to Speech: An Auditory Perspective*. Oxford University Press, 1998. To appear.

[199] Richard M. Warren, James A. Bashford, and Daniel A. Gardner. Tweaking the lexicon: Organization of vowel sequences into words. *Perception and Psychophysics*, 47(5):423–432, 1990.

[200] Richard M. Warren, Eric W. Healy, and Magdalene H. Chalikia. The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *Journal of the Acoustical Society of America*, 100(4):2452–2461, October 1996.

[201] John Wawrzynek, Krste Asanović, Brian E. D. Kingsbury, James Beck, David Johnson, and Nelson Morgan. Spert-II: A vector microprocessor system. *IEEE Computer*, 29(3):79–86, 1996.

[202] Robert Weide. The Carnegie Mellon Pronouncing Dictionary v0.4. Carnegie Mellon University, 1996.

[203] Walter Weigel. Continuous speech recognition with vowel-context-independent hidden Markov models for demisyllables. In *ICSLP*, volume 2, pages 701–704, Kobe, Japan, November 1990.

[204] Gethin Williams and Steve Renals. Confidence measures for hybrid HMM/ANN speech recognition. In *Eurospeech*, volume 4, pages 1955–1958, Rhodes, Greece, October 1997. ESCA.

[205] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young. The development of the 1994 HTK large vocabulary speech recognition system. In *Spoken Language Systems Technology Workshop*, Austin, Texas, January 1995. ARPA.

[206] Kevin Woods, W. Philip Kegelmeyer Jr., and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 19(4):405–410, April 1997.

[207] Charles Clayton Wooters. *Lexical Modelling in a Speaker Independent Speech Understanding System.* PhD thesis, UC Berkeley, November 1993. ICSI Technical Report TR-93-068.

[208] Su-Lin Wu. Properties of stochastic perceptual auditory-event-based models for automatic speech recognition. Master's thesis, UC Berkeley, May 1995. Also appears as ICSI Technical Report TR-95-023.

[209] Su-Lin Wu, Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *ICASSP*, Seattle, Washington, April 1998. IEEE.

[210] Su-Lin Wu, Michael L. Shire, Steven Greenberg, and Nelson Morgan. Integrating syllable boundary information into speech recognition. In *ICASSP*, volume 1, Munich, Germany, April 1997. IEEE.

[211] Steve Young. Large vocabulary continuous speech recognition: A review. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 3–28, Snowbird, Utah, December 1995. IEEE.

[212] Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *The HTK Book for HTK Version 2.1.* Cambridge University, 1997.