INTERNATIONAL COMPUTER SCIENCE INSTITUTE

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

A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition

Shuangyu Chang

TR-02-007

September 2002

Abstract

Current-generation automatic speech recognition (ASR) systems assume that words are readily decomposable into constituent phonetic components ("phonemes"). A detailed linguistic dissection of state-of-the-art speech recognition systems indicates that the conventional phonemic "beads-on-a-string" approach is of limited utility, particularly with respect to informal, conversational material. The study shows that there is a significant gap between the observed data and the pronunciation models of current ASR systems. It also shows that many important factors affecting recognition performance are not modeled explicitly in these systems.

Motivated by these findings, this dissertation analyzes spontaneous speech with respect to three important, but often neglected, components of speech (at least with respect to English ASR). These components are articulatory-acoustic features (AFs), the syllable and stress accent. Analysis results provide evidence for an alternative approach of speech modeling, one in which the syllable assumes preeminent status and is melded to the lower as well as the higher tiers of linguistic representation through the incorporation of prosodic information such as stress accent. Using concrete examples and statistics from spontaneous speech material it is shown that there exists a systematic relationship between the realization of AFs and stress accent in conjunction with syllable position. This relationship can be used to provide an accurate and parsimonious characterization of pronunciation variation in spontaneous speech. An approach to automatically extract AFs from the acoustic signal is also developed, as is a system for the automatic stress-accent labeling of spontaneous speech.

Based on the results of these studies a syllable-centric, multi-tier model of speech recognition is proposed. The model explicitly relates AFs, phonetic segments and syllable constituents to a framework for lexical representation, and incorporates stress-accent information into recognition. A test-bed implementation of the model is developed using a fuzzy-based approach for combining evidence from various AF sources and a pronunciation-variation modeling technique using AF-variation statistics extracted from data. Experiments on a limited-vocabulary speech recognition task using both automatically derived and fabricated data demonstrate the advantage of incorporating AF and stress-accent modeling within the syllable-centric, multi-tier framework, particularly with respect to pronunciation variation in spontaneous speech.

This technical report is a reprint of the dissertation of Shuangyu Chang filed with the University of California, Berkeley in Fall 2002.

Committee in charge:

Professor Nelson Morgan, Cochair
Dr. Lokendra Shastri, Cochair
Dr. Steven Greenberg
Professor Edwin R. Lewis
Professor David L. Wessel
Professor Lotfi A. Zadeh

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank Dr. Steven Greenberg, who has provided me the most guidance and support during the past few years. His keen insights about human speech perception have greatly helped me shape my own view on speech and language processing. His out-of-the-box thinking and meticulousness towards scientific work have always been a source of inspiration for my study and research. I feel very fortunate and grateful to have Steve as a mentor, a colleague and a friend.

I would also like to express deep gratitude to the co-chairs of my dissertation committee, Dr. Lokendra Shastri and Professor Nelson Morgan. Lokendra has not only taught me much about neural computation, knowledge representation, memory and learning, but also brought me to speech research when he first offered me a graduate student researcher position almost five years ago. Morgan has always been a generous source of support and trust. I developed a significant interest in speech recognition from taking his Audio Signal Processing course. With his broad experience and deep knowledge, Morgan has offered me many excellent advices on research, helping to keep me away from wishful thinking. Steve, Lokendra and Morgan all spent an enormous amount of time helping me with every detail of this dissertation, which could not have been possible without them.

I am very grateful to other members of my dissertation committee, Professor Ted Lewis, Professor David Wessel and Professor Lotfi Zadeh, for taking time to read this dissertation and provide stimulating discussions. In particular, I have always been inspired by Professor Zadeh's courage, perseverance and passion for original and unconventional scientific discovery.

It has been a great pleasure being a member of the Speech Group at ICSI. In particular, I appreciate the help of, and interactions with, many past and present group members, including (but not exclusively) Barry Chen, Daniel Gildea, David Gelbart, Adam Janin, Eric Fosler-Lussier, Brian Kingsbury, Jeff Bilmes, Stephane Dupont, Katrin Kirchoff, Mike Shire, Chuck Wooters, Barbara Peskin, Liz Shriberg, Andreas Stolcke, Takayuki Arai and Dan Ellis. Special thanks go to several ICSI colleagues, Nancy Chang, Srini Narayanan and Carter Wendelken for many years of wonderful interactions. I have benefited tremendously from collaborations with Rosaria Silipo, Joy Hollenback, Leah Hitchcock, Hannah Carvey and Mirjam Wester on several research projects. Thanks also go to Jack Godfrey of NSA and George Doddington of NIST for their support on the Phoneval project.

Research would have been much more difficult without the help of the computer system administrators, David Johnson, Markham Dickey and Jane Edwards. Thanks to Lila Finhill, Maria Quintana and the rest of the ICSI administrative staff for making ICSI a great place to work, and to Kathryn Crabtree and Peggy Lau for helping me deal with the bureaucracy of the University (and for bearing with me for having an unusually large dissertation committee).

I am deeply indebted to my parents for their love, encouragement and sacrifice over the many years. They must be very happy to know that I have finally finished school! I also thank my sister, Yiqiao, and her husband, Hanli, for always being there for me.

Finally and above all, my wife, Jiangxin Wang, deserves my deepest appreciation for her support and affection. She gives meaning to everything I do. This work is for her.

To Jiangxin.

Chapter 1

Introduction

The goal of this thesis is to identify significant elements and structure of spoken language beyond the conventional phone-based model of speech, particularly for building accurate and efficient models for automatic speech recognition, through detailed statistical analysis and motivated experiments on spontaneous spoken American English discourse.

1.1 The Conventional Model of Speech Recognition

The ease of conveying a rich body of information has made speech the most natural form of human communication. More recently, wide-spread use of computers has created a great demand for powerful and efficient communication methods between humans and machines, far beyond what conventional input-output devices, such as the keyboard, mouse, text and graphics display, are able to provide. Naturally, this has led to enormous expectations on using speech for human-machine communication [101][21][28]. Automatic speech recognition (ASR) by machine is one of the essential problems that have to be solved to enable such voice-based interaction.

The task of ASR (or automatic speech transcription) is to obtain a sequence of words corresponding to the information contained in the acoustic signal associated with speech. Most of the current-generation ASR systems adopt a statistical pattern recognition approach [107][8][66][146]. In such an approach the ASR problem is formulated as:

$$M^* = argmax_M P(M|X) \tag{1.1}$$

where M is any possible word string and X some representation of the acoustic input. Thus, the system seeks the word string that has the highest probability given the acoustic input. By applying Bayes' rule, Equation 1.1 can be written as:

$$M^* = argmax_M \frac{P(X|M)P(M)}{P(X)} \tag{1.2}$$

$$= argmax_M P(X|M)P(M) \tag{1.3}$$

The denominator $P(X)$ of Equation 1.3 does not depend on M and is therefore omitted during recognition (as in Equation 1.3). The ASR problem is thus decomposed into model-

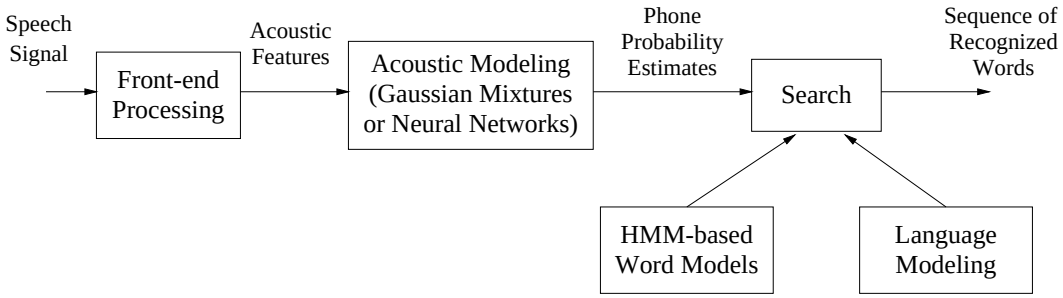

Figure 1.1: Major components of a typical current-generation ASR system.

ing two probability distributions, the acoustic likelihood $P(X|M)$ and the prior probability of the word string $P(M)$. The following paragraphs briefly describe how this pattern recognition approach is implemented in current-generation ASR systems. Major components of a typical ASR system are shown in Figure 1.1.

In a conventional ASR system, the raw acoustic signal is first processed into spectral-like features such as Mel-Frequency Cepstral Coefficients (MFCC) [22] or Perceptual Linear Prediction (PLP) [61] features, derived from the short-term spectrum of speech; the techniques are often inspired by certain properties of human auditory processing. The design goal of this front-end processing is to obtain the essential features for recognition while suppressing the irrelevant information in the acoustic signal and reducing the information rate (hence reducing the computational load associated with subsequent recognition stages). The spectral-like feature output of the front-end processing corresponds to the X in $P(X|M)$.

The spectral-like features extracted in the front-end processing are used by the acoustic model for probability estimation of specific sub-word units. The phoneme (or the phone) is the predominant sub-word unit for most of the ASR systems today (at least for English and other Indo-European languages), and both context-independent and context-dependent (e.g. the triphone) models are common. The traditional Hidden Markov Model (HMM) based systems use mixture-of-Gaussian models [107] to estimate the acoustic output distribution of each phone state. As an alternative, the hybrid HMM/ANN-based systems use artificial neural networks (ANN) [8][90] to estimate the posterior probabilities of phone states, which are then converted to scaled likelihoods with an application of Bayes' rule.

In both types of systems, HMM models are used to combine phone state probability estimates with pronunciation models contained in the lexicon (also known as the dictionary) to evaluate the acoustic likelihood, $P(X|M)$. Pronunciation models of words are generally specified as a sequence of phones. If the lexicon contains more than one pronunciation variant per word for some or all of the words, it is often referred to as a multiple-pronunciation lexicon.

The prior probability of the word string $P(M)$ is estimated by the language model. Most of the current-generation ASR systems adopt an N-gram language model, which specifies the probability of the current word given the previous $N-1$ words in the word

string. The language model often requires a large amount of text material to accurately estimate the N-gram probabilities. Of course, it is generally not possible to consider every possible word string M. Practical ASR systems require elaborate algorithms to search through the hypothesis space for the optimal word string, and efficient pruning techniques are vital for the success of the search.

ASR research in the past few decades has made significant progress and the recognition performance can be quite satisfactory on a number of tasks with certain constraints. For example, ASR systems perform adequately in limited domains with a limited vocabulary, and where speech is spoken in a formal mode or recorded under pristine acoustic conditions. However, spontaneously spoken, conversational speech as well as speech in adverse acoustic environments remain very difficult for machines. For example, state-of-the-art ASR systems yield ca. 20-30% word-error rate on large-vocabulary, conversational material spoken over the telephone, making such systems impractical in many commonly encountered environments.

In contrast, the ability of humans to understand speech is much less hindered by such factors as differences in speaking style (spontaneous or scripted speech), variability of pronunciation, background noise, reverberation and large vocabulary size. This level of capability and robustness of human speech perception may be due to the brain's processing of speech from several complementary perspectives across different time constants and structural units. Ultimately, recognition may be the result of the convergence of evidence from heterogeneous information sources. Although our understanding of exactly how humans process speech is quite limited, some knowledge and insights into human speech recognition is likely to be very helpful in building superior ASR systems. For example, comparison between ASR system output and manual annotation by human experts on conversational speech material may provide useful information, such as where the automatic system has not done well.

As described previously, one of the prevailing ideas underlying conventional ASR systems is the phone-based model of speech using either linguistically defined phonemic units or automatically derived phone-like subword units (e.g. [80][123]). Such a model assumes that spoken words are merely a linear sequence of discrete phonetic (or phone-like) segments, strung together like beads on a string (where in this instance the string is time). Successful application of this model relies on the assumption of decomposability of lexical entries into such phonetic segments, as well as on the ability of such phonetic segments to capture the invariance of speech across time and speakers with the aid of the HMM-based statistical modeling. This assumption may be quite reasonable for carefully enunciated speech material, but for spontaneous, conversational speech, the situation is never so simple. The key problem may be the enormous amount of pronunciation variation observed in such circumstances.

In [100] Ostendorf argues against using the phoneme as the basic subword unit in speech recognition, and cites several studies that point to acoustic variability as being a key problem faced by ASR systems dealing with spontaneous speech [139][87]. She also cites studies that showed very limited gain of using phone-level pronunciation variation modeling [110], and proposes several factors that may underlie this observation, all of which call for the incorporation of some linguistic units other than the phones.

Coarticulation, the overlapping of adjacent articulations, is a commonly observed phenomenon in natural speech [78]. For example, the different allophones of /k/ in *key* and *caw* have different places of articulation because of the differences in the following vowels. This coarticulation effect may even transcend the segmental boundary to modify the articulation of a non-adjacent segment (consider the different degrees of lip-rounding in the allophones of /k/ in *clean* and *clue*). The context-dependent phone models (e.g. the triphones) may partially capture the coarticulation effect but in a cumbersome manner. This approach gives rise to a large number of different models and requires a substantial amount of training data. Alternatively, since, very often, only a small number of articulatory feature dimensions are involved in coarticulation, the coarticulation effect may be captured more accurately and more succinctly by considering a more granular representation of the phonetic detail such as the articulatory-acoustic features (cf. Chapter 3 for a more detailed description). This has been at least part of the motivation for using parallel and overlapping features to model speech by Deng and colleagues [26][25][129] and for the hidden articulator modeling by Richardson et al. [109]. In addition, incorporating combinations of articulatory-acoustic features may help capture novel sound patterns that deviate from any existing phone models.

If speech were truly a sequence of phonemic beads on a string, exchanging beads of the same phoneme at different positions along the string should have little effect on intelligibility. However, this is not likely to be the case and the realization of a phonetic segment is strongly dependent on its position within the speech utterance, in particular, its position within the syllable [51][37]. For example, a phonetic segment at the syllable onset position may be substantially different from that at the syllable coda position (cf. Chapter 4 for further discussion). Because of the complex syllable structure of languages such English, even context-dependent phone models are not likely to entirely capture this syllable-level variation. As described in Chapter 4, being an articulatorily coherent unit of speech, the syllable exhibits greater stability in speech than the phonetic segment, and an explicit modeling at the syllable level may be very helpful for speech recognition.

Furthermore, when contextual and prosodic information such as the stress accent is incorporated in addition to syllable information, more systematic patterns of pronunciation variation can be obtained for spontaneous speech (cf. Chapter 5 for further discussion). Accurate modeling of such suprasegmental-level information requires explicit modeling of speech beyond the conventional phone-based approach.

It should be noted that explicit modeling of the syllable does not necessarily imply that *phonemic* beads on a string are replaced with *syllabic* beads on a string. The realization of the syllable is also highly dependent on the contextual and prosodic information. For example, the realization of the junctural element between two syllables depends on the stress-accent levels. Moreover, the manifestation of articulatory-acoustic features may also transcend syllable boundaries as it does segmental boundaries.

The discussion above argues for an effort to seek alternative representation of speech beyond the conventional phonemic-beads-on-a-string model, and suggests that various linguistic levels, both above and below the phonetic tier of speech, should be examined. It is very likely that an optimal approach is one that incorporates information from several distinct linguistic levels within an organized structure. One of the key considerations is the

ability to capture pronunciation variation phenomena of spontaneous speech.

1.2 Finding Alternatives

The discussion in the previous section suggests a potential mismatch between the assumptions made by the conventional phone-based model of speech and the reality of speech recognition, especially for spontaneous, natural speech. This section outlines an approach for seeking alternative models to address this gap between models and reality, with a special focus on pronunciation variation in spontaneous speech. An important aspect of this approach is to rely on statistical analysis of manually annotated transcription of conversational material, especially at the exploratory analysis and diagnostic stages.

A convenient starting point for seeking alternatives to the conventional model of speech is to analyze the performance of current-generation ASR systems based on the conventional phone-based model and to ascertain the significant factors underlying the recognition errors made by these systems. Without access to the detailed components of various systems, useful insights can still be gained by careful statistical analysis of the system outputs at various linguistic levels with respect to the manual annotation of a common set of spontaneous speech material.

The next step is to focus more closely on certain linguistic factors most germane to recognition from linguistic dissection of conventional ASR systems. For each of the factors of interest the advantages, as well as the limitations of being incorporated into speech recognition, are to be examined in detail from both the representational and computational perspectives. Furthermore, special attention should be paid to how various linguistic factors interact with each other in a systematic fashion, and how a structure consisting of these factors can be used to account for complex phenomena of pronunciation variation in spontaneous speech.

Based on the identified elements and structure that are most relevant for speech recognition, as well as their interaction patterns, an alternative framework can be proposed in place of the conventional model. The framework should take full advantage of significant elements and structure of speech, especially for addressing the deficiency of the conventional model in capturing the pronunciation variation associated with spontaneous speech. However, the framework should not make commitment to the specific computational techniques used, which is an engineering detail more relevant within the context of a particular implementation of the model. The framework should also leave room for improvements and augmentation to address issues not explicitly considered in the conventional model such as higher-level linguistic processing.

A test-bed implementation is to be developed using the proposed framework targeting a constrained task. The objective of building the test-bed implementation is to perform controlled experiments for testing certain hypotheses made on the basis of the framework within a transparent framework, using both automatically derived and fabricated data. The test-bed implementation is not intended to serve as a full-scale, powerful recognition system. However, its limitations and any simplifying assumptions it makes need to be explicitly delineated. Experimental results are analyzed in detail, and form the basis

for future improvements and for identifying promising directions of future research. Such a cycle of exploratory analysis, model improvement and experimental analysis may be repeated until a satisfactory alternative to the conventional phone-based model of speech is obtained. A fully functional recognition system using a unified computational framework should be implemented only after such a model has been obtained.

This thesis takes only the first few steps along the approach outlined above. Nevertheless it is hoped that this will contribute to a better understanding of the speech recognition problem and help identify promising future directions of research. The following section describes the organization of the rest of the thesis.

1.3 Thesis Outline

Chapter 2 presents a detailed statistical analysis of recognition outputs from a number of state-of-the-art, large-vocabulary, speech recognition systems on a spontaneous American English discourse task, with respect to dozens of linguistic parameters. The study identifies a number of significant factors in recognition errors and points to the limited utility of the phonemic-beads-on-a-string model used in current-generation ASR systems. Moreover, it provides a motivation for seeking alternative representations of speech, which is the focus of the remainder of the thesis.

Based on the findings from the linguistic dissection of the LVCSR systems, Chapters 3-5 describe our analysis and experiments with respect to three very important, but often neglected, components of spoken language (at least with respect to English ASR systems) – articulatory-acoustic features (AFs), syllable structure and stress accent. In particular, through concrete examples and statistics, the description illustrates how complex phenomena in spontaneous speech, such as pronunciation variation, can be captured in parsimonious fashion through systematic interaction of AFs with syllable information and stress accent. Chapter 3 describes an approach for automatically extracting AFs from the speech signal, which has many advantages for incorporating such features into recognition systems. Chapter 4 presents evidence to support a syllable-centric approach for speech processing. Chapter 5 provides a detailed account of stress accent in spontaneous American English discourse, including the development of an automatic stress-accent labeling system.

In Chapter 6, a syllable-centric, multi-tier model of speech, incorporating AFs and stress accent, is proposed as an alternative to the conventional phonetic-segment-based model of speech. A test-bed implementation of the multi-tier model is described and details of each component are provided. A fuzzy-based approach for combining evidence from various articulatory-acoustic feature sources is presented, together with a pronunciation-variation modeling technique using AF variation statistics extracted from the data. Chapter 7 describes controlled experiments performed on a constrained task (OGI Numbers95 [12]) along with detailed analysis of experimental results. Finally, Chapter 8 concludes the thesis with further discussion and proposed directions for future work.

Chapter 2

Linguistic Dissection of LVCSR Systems

The past few decades have witnessed a tremendous improvement in automatic speech recognition (ASR). Many systems are able to achieve very good performance on constrained tasks (i.e., either the domain is limited or the speech is confined to a single speaker). However, automatic recognition of spontaneous speech of unlimited domain still remains a very challenging task. For example, state-of-the-art ASR systems obtained ca. 20-30% word-error rate on the Switchboard corpus [42] (casual telephone dialogues) in recent evaluations [98]. This performance is still far from what is required for routine automatic transcription of spontaneous material.

While the need for significant improvement is obvious, the increasingly complex and sophisticated architecture of current-generation ASR systems make it very difficult to understand why they do not work as well as they should, although such knowledge would be very beneficial to advancing the technology. Without the ability to access (and understand) detailed components of a system it is difficult to provide intelligent diagnostics. However, the main framework and principles of speech modeling adopted by many ASR systems today have largely converged to phone-based statistical modeling using Hidden Markov Models. In conjunction with well-defined and carefully annotated evaluation material, this similarity among various systems allows detailed and informative analysis of the functional architecture of various ASR systems simply from the outputs at various linguistic levels.

In 2000 and 2001 we were given the opportunity to perform such linguistic dissection [59][57] of several large-vocabulary continuous speech recognition (LVCSR) systems on the Switchboard corpus, for which a substantial amount of material has been phonetically labeled and segmented by linguistically trained individuals [49].

The goal of the diagnostic evaluation is to identify significant factors that affect recognition performance and to provide a basis for alternative and novel approaches to speech modeling superior to the current ones. This chapter describes the diagnostic evaluation of Switchboard-corpus LVCSR systems in detail (with additional information on the evaluation procedure in Appendix A). It shows, through statistical analysis, the beneficial effect of accurate specification of phonetic features, syllable-level parameters, stress-

accent and other prosodic features on word recognition. This will also provide a motivation for developing a syllable-based, multi-tier model of speech processing, which incorporates articulatory-acoustic feature and stress-accent information.

2.1 Background Information

Over the past few years the National Institute of Standards and Technology (NIST) has sponsored annual competitive evaluation of LVCSR systems for spontaneous American English. The Switchboard corpus [42] (in tandem with the Call Home and Broadcast News corpora) has been used to assess the state of automatic speech recognition (ASR) performance. Switchboard is unique among the large-vocabulary corpora in that a substantial amount of material has been phonetically labeled and segmented by linguistically trained individuals from the Switchboard Transcription Project (STP) [49], and thus provides a crucial set of "reference" materials with which to assess and evaluate the phonetic and lexical classification capabilities of current-generation ASR systems.

In both 2000 and 2001, we performed detailed diagnostic evaluation of several LVCSR systems during the annual evaluations. The diagnostic evaluation materials used in 2000 and 2001 had similar characteristics but were distinct in certain respects. This section provides general information on the corpus materials as well as a brief description of the ASR systems, upon which the evaluation was performed. Additional information pertaining to phone mappings and the detailed evaluation procedure is presented in Appendix A.

2.1.1 Corpus Material

Year-2000 Material

The Switchboard corpus contains informal dialogues recorded over the telephone. The Year-2000 diagnostic evaluation was performed on a fifty-four-minute, phonetically annotated subset of the Switchboard corpus, distinct from the materials used in the competitive evaluation. All of this material had previously been manually transcribed at the phonetic-segment level and manually segmented at either the phonetic-segment or the syllabic level. The syllable-segmented material was subsequently segmented at the phonetic-segment level by an automatic procedure (very similar to that described in Chapter 3 and also in [14]) trained on 72-minutes of manually segmented Switchboard material. This automatic segmentation was manually verified. In addition, this material has also been manually labeled at the stress-accent level where each syllable was marked with respect to stress accent.

The diagnostic material was carefully chosen to cover a broad range of speaker characteristics. It contains 581 different speakers, a relatively equal balance of female and male speakers, a broad distribution of utterance durations, coverage of all seven U.S. dialect regions in Switchboard corpus, a wide range of discussion topics and variability in subjective recognition difficulty (from very easy to very hard).

Year-2001 Material

The Year-2001 material has many characteristics in common with the Year-2000 material. It was manually transcribed at the phonetic-segment level and segmented at the syllabic level, and was subsequently segmented at the phonetic-segment level by the same automatic procedure as used in year 2000 (and manually verified). The stress-accent labels were initially produced by an automatic stress-accent labeler (see Chapter 5 and [58]) and subsequently verified by a linguistically trained individual.

The main difference between the two data sets is that the Year-2001 material was a subset of the competitive evaluation corpus; moreover, it contains a relatively even distribution of data derived from three recording conditions: one cellular and two land-line conditions. This material has 21 separate conversations (42 separate speakers) and a total of 74 minutes of spoken language material (including filled pauses, junctures, etc.), divided into 917 separate utterances.

2.1.2 Participating Systems

Eight separate sites participated in the Year-2000 evaluation - AT&T, BBN, Cambridge University (CU), Dragon Systems (DRAG), Johns Hopkins University (JHU), Mississippi State University (MSU), SRI International and the University of Washington (UW). Each of the eight sites provided word and phonetic-segment output on the fifty-four-minute material from the recognition system used for the competitive evaluation portion of Switchboard corpus. Six of the eight sites also provided word and phone-level output of forced-alignments (constrained recognition with the knowledge of the word sequence) associated with the same material.

In year 2001, seven sites participated in the evaluation - AT&T, BBN, IBM, JHU, MSU, Philips (PHI) and SRI. Six of the seven sites provided unconstrained recognition output at the word and phonetic-segment levels and all seven sites provided forced-alignment output at the word and phone levels[1]. It should be noted that the systems participating in these evaluations were optimized for word recognition rather than for phone recognition *per se*. The phone-level output was generally a by-product of the word-recognition process, although various systems differed in how the phone-level output was extracted.

While systems from different sites have many different functions and features, some fundamental characteristics of speech modeling are shared by the various systems. Almost all systems have triphone- (or quinta-phone-) based gender-dependent acoustic models, trained with maximum likelihood (ML) or maximum mutual information (MMI) criteria [107]. The lexicon often contains multiple pronunciations (usually with different probabilities). Many systems have multiple passes for decoding, where the later passes often employ increasingly higher-order language models for rescoring word lattices generated at earlier passes. Some form of adaptation (such as that based on maximum-likelihood, linear regression (MLLR) [82]) is also common in many of the systems.

[1]Due to certain format inconsistencies, some of the data from a few sites were excluded from the analysis.

2.2 Analysis Results

Detailed statistical analyses were performed on the extracted data (cf. Appendix A for a detailed description of the evaluation procedure) and the highlights of the analysis results are described in this section. Although some of the material from each year's evaluation are unique, many statistical patterns are shared in common by the two sets of materials.

2.2.1 Word and Phone Error Patterns

Word- and phone-level error patterns were computed for both the constrained (forced-alignment based on word-level transcripts) and unconstrained recognition material. For the Year-2001 evaluation in particular we have added a phone-mapping procedure to allow for certain phones commonly confused among human transcribers to be scored as "correct" even though they would otherwise be scored as "wrong." We call this specific mapping the transcription-compensated (TC) form, in contrast to the uncompensated (TU) form where only common phone ambiguities were allowed. Appendix A.1 provides a detailed description of the difference between TC and TU phone mappings. Also for the Year-2001 evaluation the time-mediated scoring at both the word and phone levels include two different levels of tolerances: a strict time-mediation that heavily penalizes time-mismatches between reference segments and system output segments, and a lenient time-mediation that deweights this time-mismatch penalty. Refer to Appendix A.2.2 for a more detailed description of the difference between strict and lenient time-mediation. The Year-2000 evaluation is thus equivalent to using TU phone mappings and strict time-mediation.

Word-error rates for the ASR systems range between 27 and 43% for the Year-2000 material (cf. Figure 2.1, upper panel). For the Year-2001 material word-error rates range between 33 and 49% with strict time-mediation and between 31 and 44% with lenient time-mediation (cf. Figure 2.1, lower panel). The phone recognition error rate is relatively high for the forced-alignment output (35-49% for the Year-2000 evaluation; 40-51% for the Year-2001 evaluation, with strict time-mediation and TU phone mappings) only slightly lower than the phone error for unconstrained recognition (39-55% for the Year-2000 evaluation; 43-55% for the Year-2001 evaluation, with strict time-mediation and TU phone mappings).

Figure 2.2 shows the break-down of phone errors for the Year-2000 constrained (upper panel) and unconstrained recognition (lower panel) with respect to the types of error. Substitution is the primary form of phone recognition error in both constrained and unconstrained recognition. In the constrained recognition, substitution and deletion rates are relatively even across different sites, while insertion rate varies more across sites. The relatively high insertion error rate in the constrained recognition suggests an inflexibility of the ASR systems' pronunciation models in dealing with non-canonical pronunciation variants such as phone deletions that occur frequently in conversational speech. In the unconstrained recognition, insertion rate and deletion rate roughly trade off each other such that the sum of insertion and deletion rates is roughly even across sites; there is a great variability in substitution rate across sites. The Year-2001 data exhibit a similar trend in the break-down of phone-error types.

Figure 2.3 shows the phone-error rates for the Year-2001 constrained (upper panel)

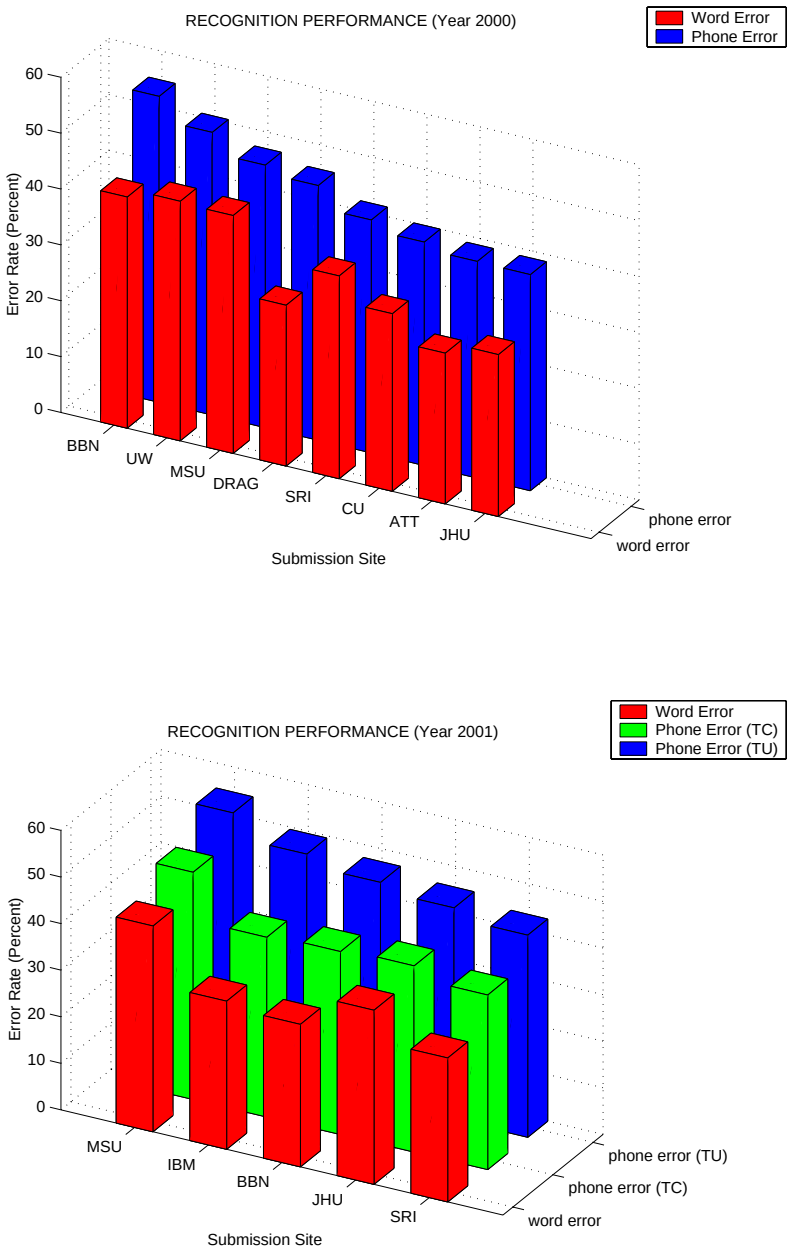

Figure 2.1: Word and phone-error rates for unconstrained recognition (Upper: the Year-2000 data; Lower: the Year-2001 data). For the Year-2001 data, TC and TU refer to transcription-compensated and -uncompensated phone mappings, respectively. Both used strict time-mediation, as for the Year-2000 data. The correlation coefficient between phone- and word-error rates is 0.78 for the Year-2000 data, and 0.93 (TC) and 0.72 (TU) for the Year-2001 data.

and unconstrained recognition (lower panel) using four different scoring conditions. The consistent, significant decrease in phone-error rates from using strict time-mediation to using lenient time-mediation suggests that the temporal alignment of the phone segments generated by the ASR systems are often inaccurate (with respect to manual segmentation). For the Year-2000 material, the phonetic-segment boundaries generated by the constrained recognition systems differ by an average of 32 ms (40% of the mean phone duration) from the hand-labeled material. The decrease in phone-error rates from using TU phone mappings to using TC phone mappings suggests that the phone-confusion patterns of the ASR systems' output share certain characteristics in common with the phone-confusion patterns among human transcribers.

Figure 2.1 (upper panel for the Year-2000 material and lower panel for the Year-2001 material) illustrates the relationship between phone- and word-error magnitude across submission sites. For the Year-2001 material the correlation coefficient (r) between phone- and word-error magnitude is 0.93 for strict time-mediation and TC phone mappings, suggesting a significant dependence of word-recognition performance on the accuracy of recognition at the phonetic-segment level. With strict time-mediation and TU phone mappings, this correlation coefficient is $r = 0.72$. The difference in the correlation coefficients of phone- and word-error magnitude between using TC and TU phone mappings suggests that certain phone-confusion patterns have been captured through acoustic modeling of the ASR systems, partially compensating for the inflexibility of the pronunciation models. For the Year-2000 material (using strict time-mediation, TU phone mappings), the correlation coefficient between phone- and word-error magnitude is $r = 0.78$. Such results suggest that word recognition may heavily depend on the accuracy of recognition at the phonetic-segment level. Thus, improving acoustic modeling that enhances phonetic recognition performance is likely to help improve word recognition. However, it should also be noted that the LVCSR systems were not optimized for phone recognition but rather for word recognition. Moreover, the phone-level output was largely constrained by the pronunciation models associated with the recognized words (cf. Section 2.2.6 for analysis on pronunciation variation and word errors).

It is of interest to ascertain whether the number of phonetic segments in a word bears any relation to the pattern of phone errors in both *correctly* and *incorrectly* recognized words (cf. Figure 2.4). Interestingly, the tolerance for phonetic-segment errors in correctly recognized words is not linearly related to the length of the word. The tolerance for error (ca. 1-1.5 phones) is roughly constant for word lengths of four phones or less. This pattern is observed regardless of the form of error. The relatively low tolerance for phone misclassification (except for words of very short length) implies that the pronunciation and language models possess only a limited capacity to compensate for errors at the phonetic-segment level. The tolerance for phone-deletion errors in correctly recognized words is particularly low, suggesting that the ASR systems are more sensitive to phone deletions than phone substitutions and insertions.

In contrast, the average number of phones misclassified in incorrectly recognized words does increase in quasi-linear fashion as a function of word length (with the possible exception of insertions), a pattern consistent with the importance of phonetic classification for accurate word recognition.

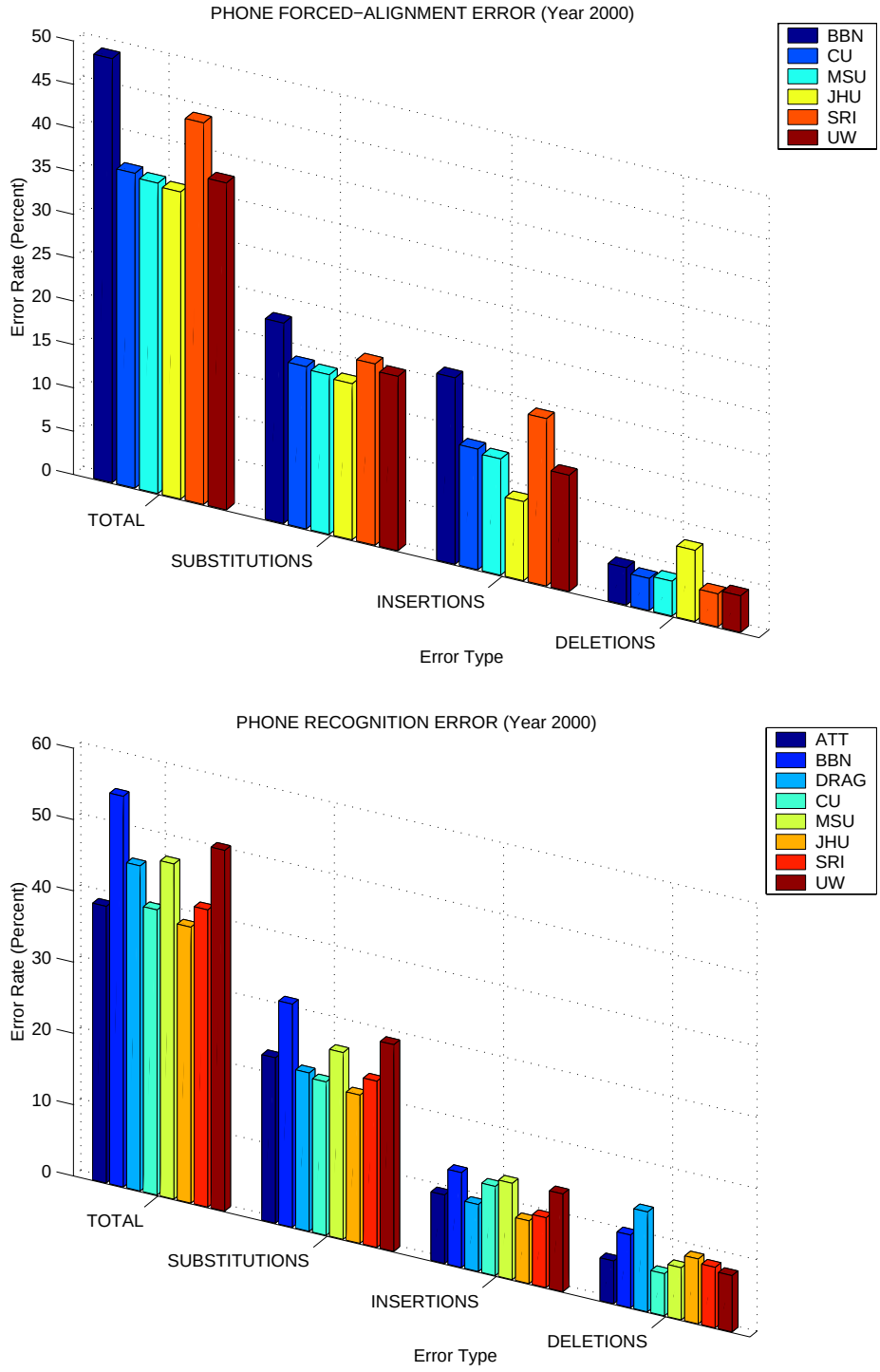

Figure 2.2: Phone-error breakdown of constrained (upper panel) and unconstrained recognition (lower panel) output for the Year-2000 data. In both cases substitutions are the primary form of phone recognition error.

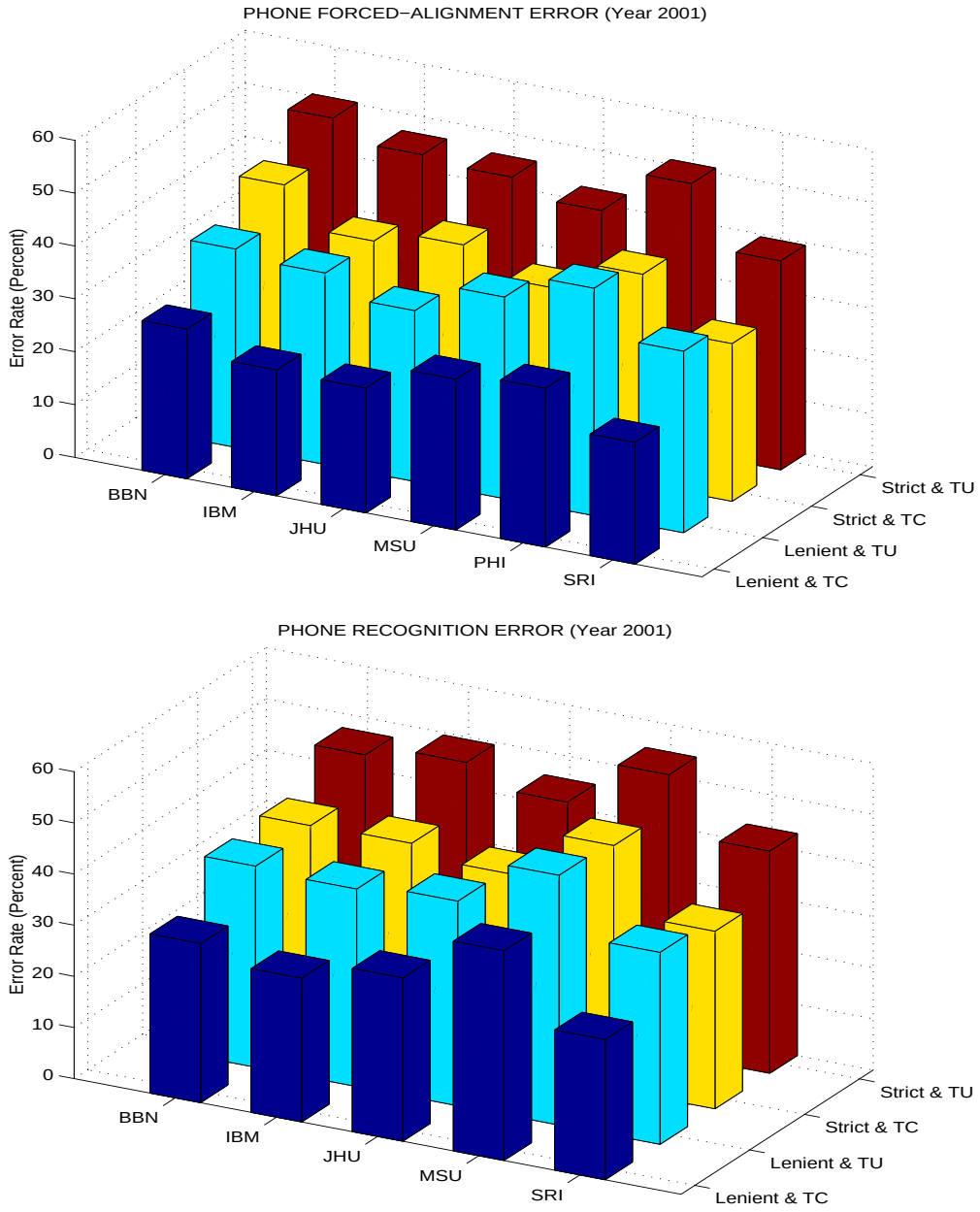

Figure 2.3: Phone errors of constrained (upper panel) and unconstrained recognition (lower panel) of the Year-2001 data under four different scoring conditions. "Strict" and "Lenient" refer to strict- and lenient-time-mediation, respectively; "TC" and "TU" refer to transcription-compensated and -uncompensated phone mappings, respectively.

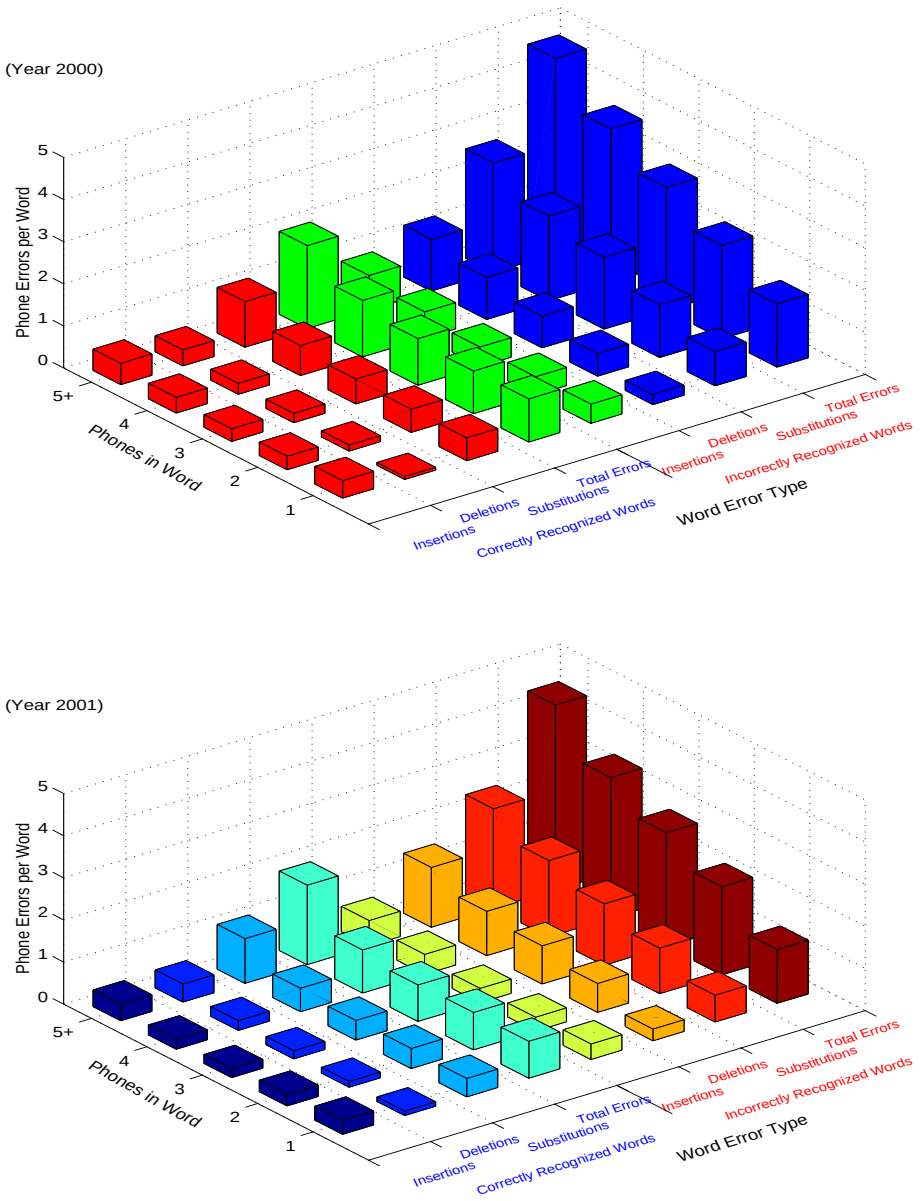

Figure 2.4: The number of phone errors as a function of word length (with respect to the number of phone segments) for both correctly and incorrectly recognized words. The upper panel is the Year-2000 data and the lower panel the Year-2001 data. The tolerance for phonetic-segment errors in correctly recognized words are roughly constant for word lengths of four phones or less. In contrast, the average number of phone errors in incorrectly recognized words is quasi-linearly related to word length.

2.2.2 Syllable Structure and Syllable Position

The syllable is a fundamentally important unit of the English language and the words in English are organized into distinct forms of syllabic structures (consonant and vowel sequences, cf. Chapter 4 for a more detailed description of syllables in spoken English). Therefore, it is of interest to examine recognition error patterns according to their lexically based syllable structure. Figure 2.5 (upper panel) shows word-error rate as a function of syllable structure for the Year-2000 material. The highest error rates are associated with vowel-initial syllables. This may be explained by the observation that consonantal onsets are often much more stable (with less deviation from canonical pronunciation) than nuclei and codas (cf. Section 4.3). Thus, syllable forms lacking such stable onsets (i.e., vowel-initial syllables) are likely to possess few stable cues for word recognition. For the same reason, consonant-initial syllable forms (especially those with consonant-cluster onsets) tend to exhibit a relatively lower word-error rate, as observed in Figure 2.5. Polysyllabic words tend to exhibit lower word-error rates than monosyllabic words, and this effect is particularly pronounced with respect to deletions. While there are roughly comparable substitution rates across syllable forms, polysyllabic words tend to have much lower deletion rates than monosyllabic words. This may be due to merely greater number of phonetic cues contained in polysyllabic words than in monosyllabic words for lexical access; it may also be due to the greater likelihood of containing stress-accented syllables in polysyllabic words, which tend to have greater information content than monosyllabic words.

A similar relationship between word error and syllable structure is observed in the Year-2001 material (cf. Figure 2.5, lower panel), with the exception that the "CVCV" form has an unusually high substitution error rate albeit being polysyllabic and consonant-initial. Detailed study of the evaluation material suggests that the elevated error rate associated with "CVCV" forms in the year-2001 data is due to the inconsistent representation of word-compounds such as "gonna" ("going to") and "wanna" ("want to") in the reference transcript and the recognition lexicon.

Like word errors, error patterns at the phonetic-segment level are also intimately linked to syllable structure, as well as to position within the syllable. Figure 2.6 shows phone classification accuracy at different positions in several common syllable structures, from the Year-2001 forced-alignment submissions. Onset consonants in "CVC" syllables (cf. Figure 2.6, upper panel) tend to be highly concordant with the manual annotation, while coda consonants are somewhat less concordant. Poor concordance for certain segments, such as [zh] and [nx] in both onsets and codas, and [dh] in codas, may be due to the low frequency of their occurrences. Phonetic segments in consonantal cluster onsets and codas (cf. Figure 2.6, middle panel) exhibit a relatively lower degree of concordance than those associated with simple onsets and codas. Certain segments in consonantal clusters, such as [g] and [p], are much less concordant in codas than in onsets, which is not observed in "CVC" syllables. Diphthongs and tense, low monophthongs (cf. Figure 2.6, lower panel) tend to be concordant with the manual transcript, while lax monophthongs are generally less concordant. Diphthongs in "CV" syllables (open syllables) tend to have higher relative frequencies (except [ay]) than in "CVC" syllables (closed syllables). Diphthongs also exhibit relatively comparable degree of concordance in the two syllable forms, while monophthongs

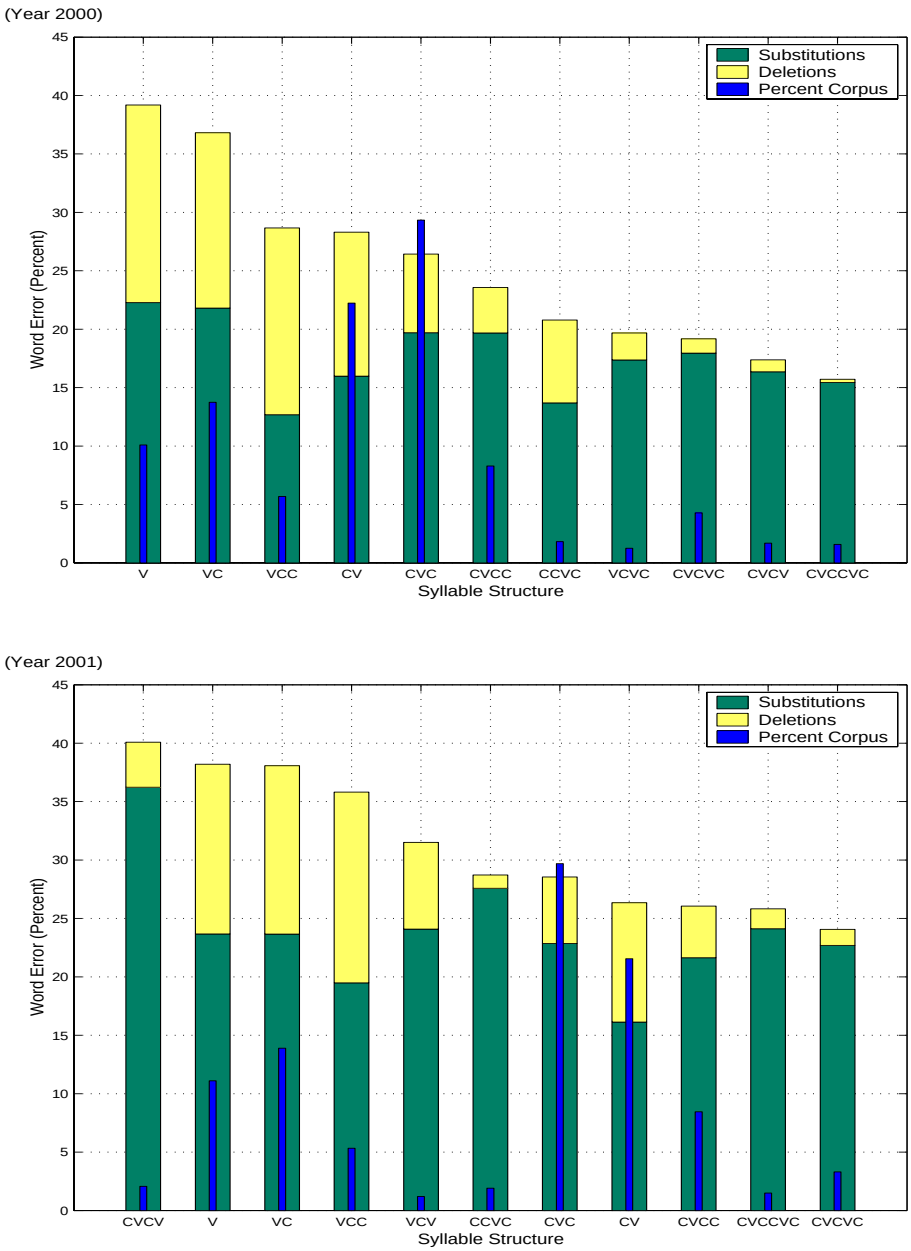

Figure 2.5: Word error (substitution and deletion) as a function of syllable structure and the proportion of the corpus associated with each type of syllable structure. The upper panel shows the Year-2000 data and the lower panel the Year-2001 data. Only the most frequent 11 syllable structures are graphed for each year's data. Vowel-initial forms tend to exhibit higher word-error rates than consonant-initial forms; monosyllabic forms generally exhibit more errors than polysyllabic forms.

are generally less concordant in "CV" syllables than in "CVC" syllables. This suggests that diphthongs in open syllables may be playing a role similar to that of the sum of the nucleus and coda in closed syllables.

2.2.3 Articulatory-acoustic Features and Syllable Position

Phonetic segments can be decomposed into more elementary constituents based on their articulatory bases, such as place (e.g., labial, labio-dental, alveolar, velar), manner (e.g., stop, fricative, affricate, nasal, liquid, glide, vocalic), voicing and lip-rounding. Two additional dimensions were also used in the current analyses - front-back articulation (for vowels only) and the general distinction between consonantal and vocalic segmental forms.

Figure 2.7 show the AF-error patterns (the Year-2000 material) partitioned according to lexical syllable structure and whether a word was correctly or incorrectly recognized, for onset, nucleus and coda positions of a syllable. Similar AF-error patterns across syllable positions are exhibited by the Year-2001 data (omitted here).

The articulatory features associated with consonantal onsets (Figure 2.7, upper panel) exhibit a relatively low tolerance for error. Moreover, the error rate is four to five times greater for AFs in misclassified words relative to correctly recognized lexical items. Place and manner features are particularly prone to error in misclassified words, suggesting that these onset-consonant AFs are particularly important for correctly distinguishing among words. The AF error patterns for consonantal codas (Figure 2.7, middle panel) are similar to those associated with consonantal onsets, except that there is a higher (ca. 50%) tolerance for error among the former for manner, place, voicing and lip-rounding features. In particular, syllable codas exhibit a much higher tolerance to voicing errors than syllable onsets. Vowel/consonant errors are rare in both onsets and codas, even for incorrectly recognized words, implying that manner feature errors in both onsets and codas are generally consonant-class confusions. There is relatively small variation in AF errors across syllable forms, except relatively fewer AF errors in polysyllabic words than monosyllabic words in onsets.

The AF error patterns associated with vocalic nuclei (Figure 2.7, lower panel) display a pattern different from those associated with onsets and codas. There is a much higher tolerance of error for classification of AFs associated with correctly recognized words, which is particularly marked for place and front-back features. Moreover, there is a considerably higher degree of AF classification error among the nuclei compared to onsets and codas, particularly among the place and front-back dimensions. Such data imply that classification of vocalic nuclei is considerably less precise than for the onsets and codas. However, unlike onsets and codas, nuclei in correctly recognized words have a much lower level of vowel/consonant errors (which is equivalent to manner errors for nuclei) than nuclei in incorrectly recognized words, suggesting the importance of correct detection of syllable nuclei. There is a greater variation in AF errors across syllable forms in nuclei than in onsets and codas, and the variation pattern is different among AF dimensions. For example, "V" syllables exhibit a high degree of tolerance for front-back error but a relatively low level of tolerance for vocalic height (place) error.

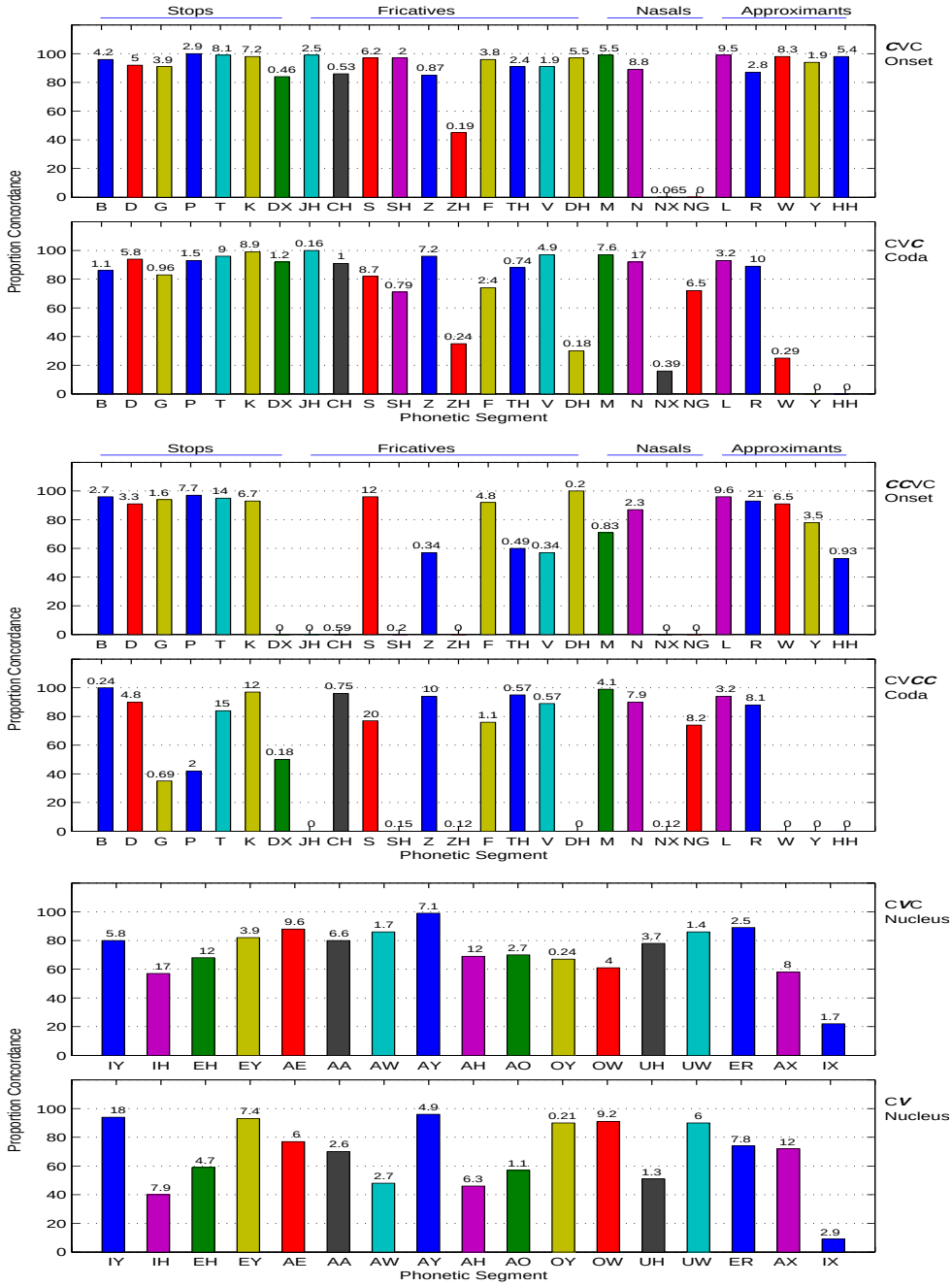

Figure 2.6: Phone recognition accuracy for onset, nucleus and coda positions for the most common syllable structures in the Year-2001 forced-alignment output. Upper panel: onset and coda in CVC syllables; middle panel: consonant cluster onset and coda (in CCVC and CVCC syllables); lower panel: the nucleus in CVC and CV syllables. The number on top of each bar is the frequency of occurrence (in percentage) associated with each segment in the specified position and syllable form.

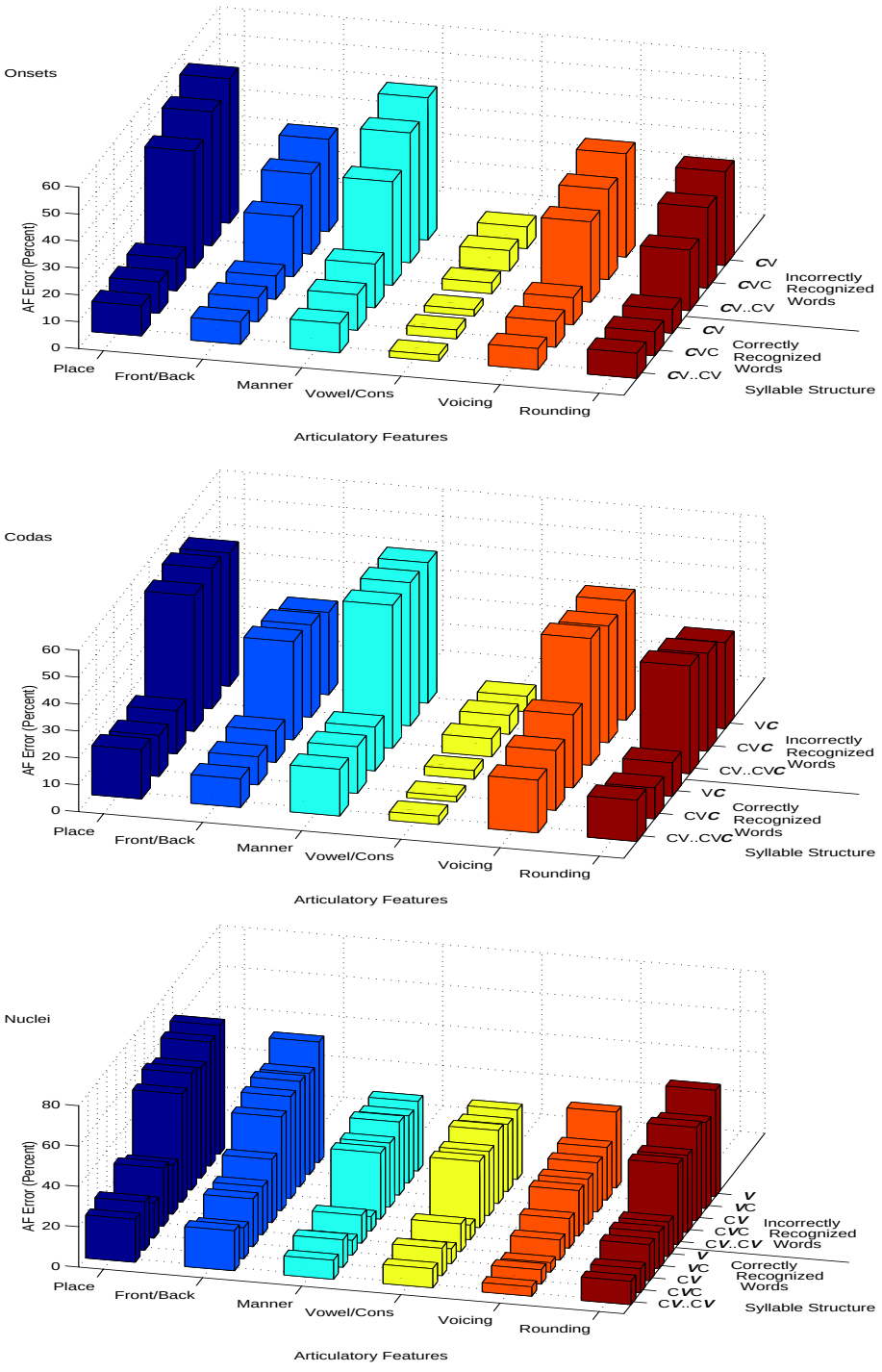

Figure 2.7: The average error in classification of articulatory features associated with each phonetic segment for consonantal onsets (upper panel), codas (middle panel) and vocalic nuclei (lower panel). "CV..CV" and "CV..CVC" indicate a polysyllabic word.

2.2.4 Prosodic Stress Accent and Word Errors

English is often characterized as a stress-accent language as it utilizes a combination of acoustic cues, including loudness, duration and pitch variation, to emphasize certain syllables over others [5][83]. Besides providing informational focus in natural speech, prosodic stress accent also affects the pronunciation of phonetic elements [37][51].

The Year-2000 diagnostic evaluation material contains stress-accent markings manually labeled by two linguistically trained individuals; this material was used to ascertain the relation between error rate and stress-accent level. As illustrated in Figure 2.8 (upper panel), there is a ca. 50% higher probability of a recognition error when a word is entirely unaccented. The relation between stress accent and word-error rate is particularly apparent for deletions and is manifest across all ASR systems (Figure 2.8, lower panel). This effect suggests that it may be helpful to model stress accent explicitly in ASR systems. Figure 2.9 shows corresponding statistics computed on the Year-2001 evaluation material where the stress-accent labels were derived using an automatic stress-accent labeling system (described in Section 5.3) and manually verified by a linguistically trained individual. The Year-2001 material exhibits a relationship between word-error rate and stress accent nearly identical to that exhibited by the Year-2000 material. Stress accent will be discussed in more detail in Chapter 5.

2.2.5 Speaking Rate and Word Errors

ASR systems generally have more difficulty recognizing speech that is of particularly fast [89][91] or slow [91] tempo. A variety of methods have been proposed for automatically estimating speaking rate from the acoustic signal as a means of adapting recognition algorithms to the speaker's tempo [89][91][132].

The speaking rate of each utterance in the diagnostic material was measured using two different metrics. The first, MRATE [92], derives its estimate of speaking rate by combining several acoustic-based measures including a multi-band correlation function of the signal energy, using a peak-picking routine and also a full-band version of this routine, as well as the spectral moment for a full-band energy envelope. The MRATE is roughly correlated with transcribed syllable rate although it tends to underestimate the rate for fast speech [92][36]. The second metric used is based directly on the number of syllables spoken per second and is derived from the transcription material.

Figure 2.10 illustrates the relation between MRATE and word-error rate for the Year-2000 material. Word error does not change very much as a function of MRATE. In many instances the highest error rates are associated with the middle of the MRATE range, while the flanks of the range often exhibit a slightly lower proportion of word errors.

It was found in several studies that the linguistic measure of speaking rate often has a high correlation with word-error rate of ASR systems [120][89]. Figure 2.11 illustrates the relation between word-error rate and syllables per second (derived from the manual transcript). In contrast to MRATE, this linguistic metric exhibits a much higher correlation between abnormal speech tempo and ASR performance. Utterances slower than 3 syllables/sec or faster than 6 syllables/sec exhibit 50% more word-recognition errors than

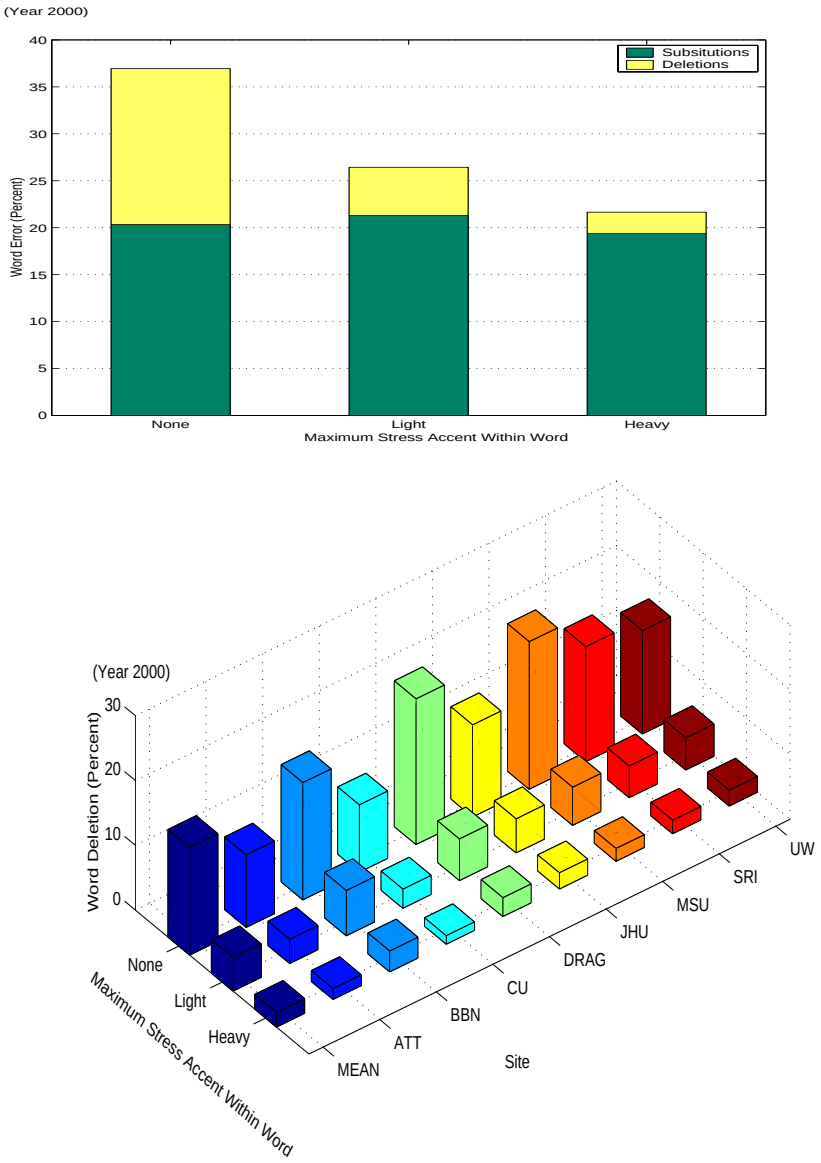

Figure 2.8: Upper panel: The average word error (substitution and deletion) as a function of the maximum stress-accent level associated with a word from the Year-2000 data, averaged across eight sites. Lower panel: the average number of word deletions as a function of the maximum stress-accent level. A maximum stress-accent level of "0" indicates that the word was completely unaccented; "1" indicates that at least one syllable in the word was fully accented; an intermediate level of stress accent is associated with a value of "0.5."

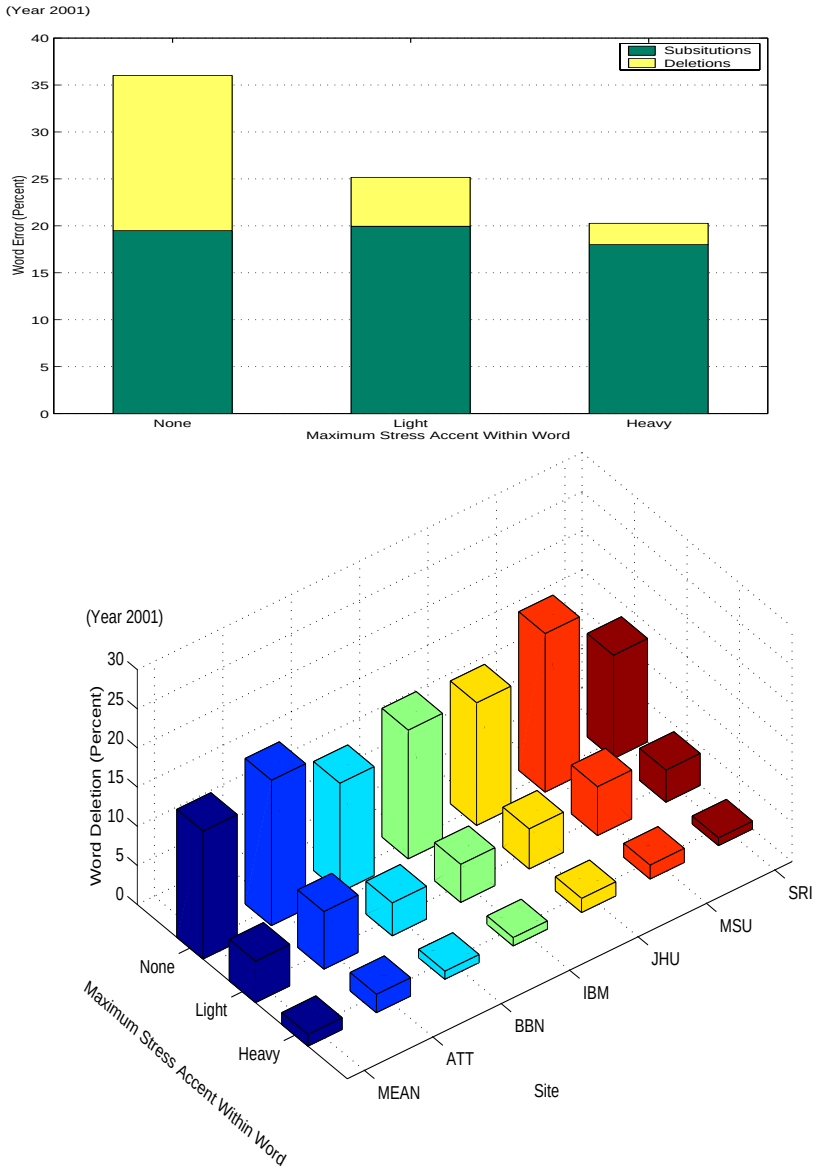

Figure 2.9: Upper panel: The average word error (substitution and deletion) as a function of the maximum stress-accent level associated with a word from the Year-2001 data, averaged across eight sites. Lower panel: the average number of word deletions as a function of the maximum stress-accent level. Stress-accent magnitudes are as described in Figure 2.8.

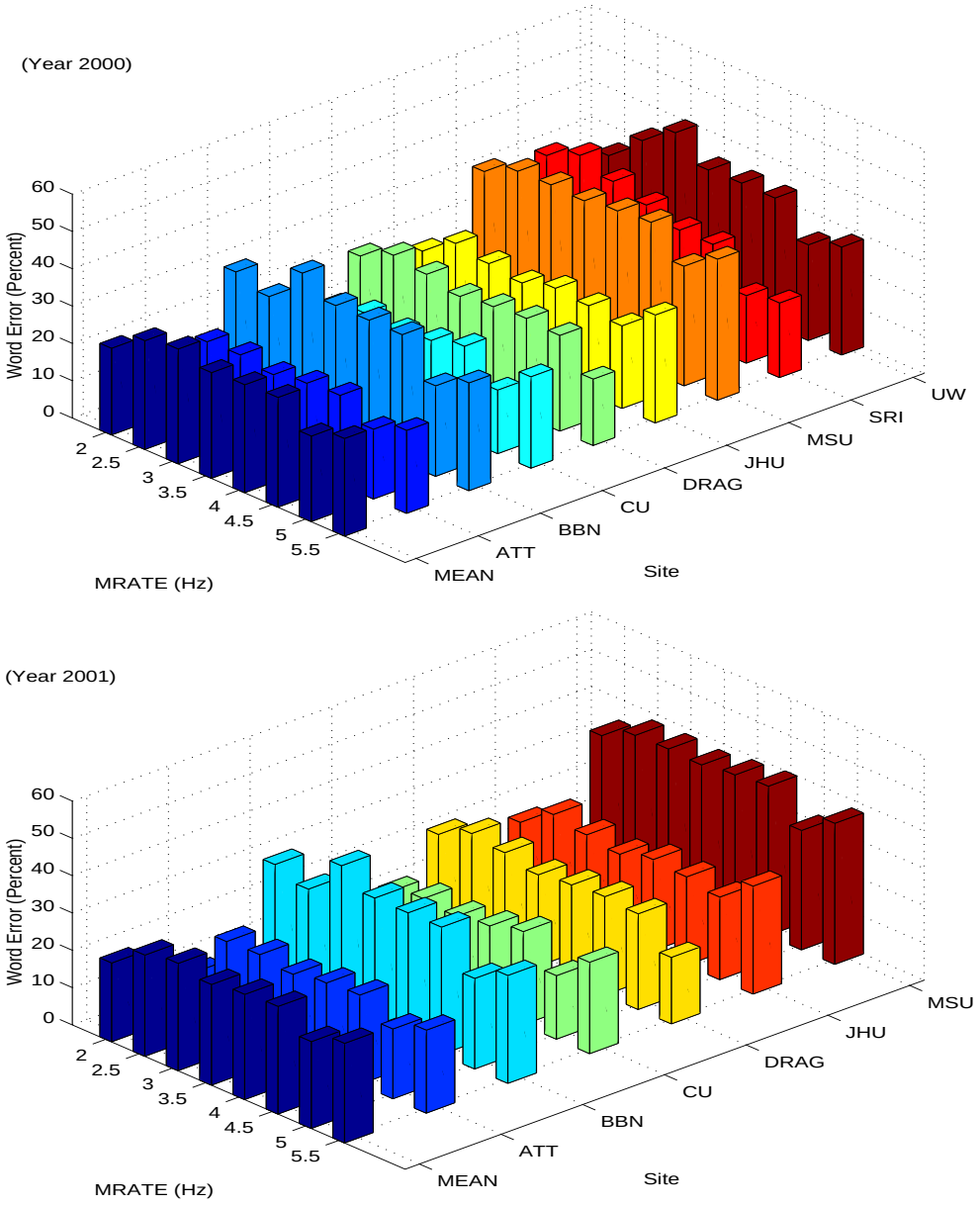

Figure 2.10: The relationship between word-error rate for each site (as well as the mean) and an acoustic measure of speaking rate (MRATE). Upper panel: the Year-2000 data; lower-panel: the Year-2001 data.

their counterparts in the core of the normal speaking range. Similar patterns were observed on the the Year-2001 data (cf. Figures 2.10 and 2.11). Such data imply that algorithms based on some form of linguistic segmentation related to the syllable are more likely to be a better predictor of word-error rate of ASR systems than those based purely on acoustic properties of the speech signal.

In this study, speaking rate was measured at the utterance level. However, the utterance may not be the smallest unit over which the speaking rate is constant. There are often significant variations in speaking rate over the duration of a speech utterance [149]. Thus, it may be more appropriate to measure speaking rate over a shorter time interval than the utterance. There is a systematic relationship between speaking rate and stress accent. Fast speech tend to have greater proportion of syllables unaccented than slow speech. Thus, a localized measure of speaking rate may be closely related to the proportion of accented (or unaccented) syllables within a small group of syllables.

2.2.6 Pronunciation Variation and Word Errors

All ASR systems participating in the evaluation contain multiple pronunciation modeling (such that each word form could potentially be associated with several different phonetic representations). However, there appear to be a far greater number of pronunciation variants in the transcription material than observed in the ASR system outputs (for both unconstrained recognition and forced-alignment). For example, the word "time" has at least nine different pronunciations according to the manual transcript ([t ay m],[t ax m],[t ah m],[t aa mx],[t aa m],[t ay n],[t aa ay m],[t aw m],[t ay]), while only one pronunciation of the word ([t ay m]) was found in the output from most of the systems. For another example, the word "from" has twelve different pronunciations according to the manual transcript ([f ax m],[f r ax mx],[f em],[f er m],[f r em],[f ah m],[f r ax m],[f r ah m],[f eh m],[f ax],[f r ah],[th eh l m]), while three pronunciations of the word ([f er m],[f r ax m],[f r ah m]) were found in the system outputs. For an example of even more complex pronunciation variation patterns, see Table 4.1 in Chapter 4 for the 63 different pronunciation variants of the word "that" found in the Year-2001 material. If the system outputs truly reflect the range of pronunciation variations in the current generation ASR systems' lexicon [2], it seems to be extremely inadequate in dealing with the highly non-canonical pronunciation of conversational speech.

Being aware of the large disparity between the number of pronunciation variants in the transcript and in the system outputs, it is of interest to ascertain the impact of pronunciation variation on word error. Figure 2.12 illustrates the relationship between the average number of pronunciation variants and the word-correct rate (one minus the sum of substitution rate and deletion rate) across six sites for the frequently occurring words (at least ten occurrences)[3] in the Year-2001 material. The correlation coefficient is quite

[2] It should noted that most of the ASR systems use context-dependent phone (e.g. triphone) models, but the phone-level output was converted to context-dependent phones prior to submission. Thus, the evaluation was performed only using context-independent phones, which may underestimate the variability captured by the ASR systems. Nevertheless, majority of the analyses and conclusions remain valid.

[3] Infrequently occurring words tend to have only one (or very few) pronunciation.

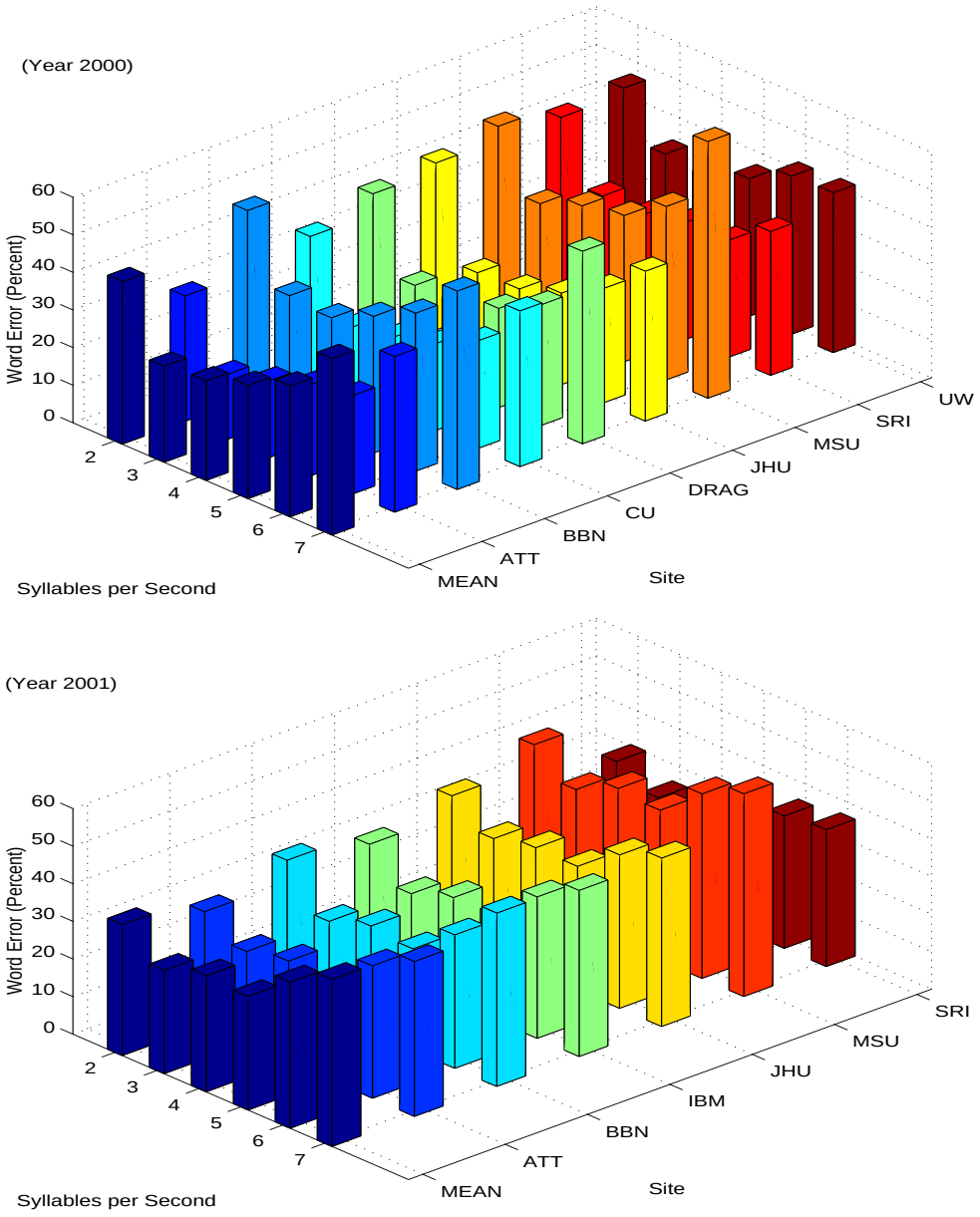

Figure 2.11: The relationship between word-error rate for each site (as well as the mean) and a linguistic measure of speaking rate (syllables per second). Upper panel: the Year-2000 data; lower-panel: the Year-2001 data. Note the "U" shape in word-error rate as a function of speaking rate for each site (and the mean), indicating that very slow and very fast speech tends to have more word errors than speech spoken at a normal tempo.

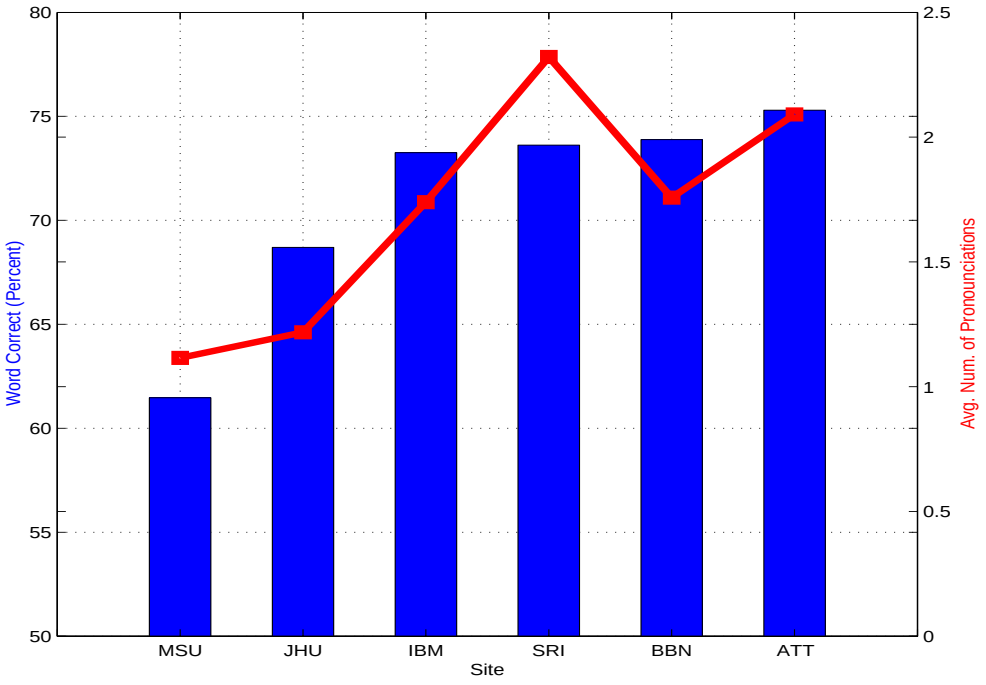

Figure 2.12: The relationship between word-correct rate (one minus the sum of substitution and deletion rates) and the average number of pronunciation variants per word (for words with at least ten occurrences) found in the system outputs for the Year-2001 material. The correlation coefficient (r) is 0.84, suggesting that more sophisticated pronunciation modeling is likely to yield higher word recognition performance.

high, 0.84, suggesting that more sophisticated pronunciation modeling is likely to yield better word recognition performance. This phenomenon of pronunciation variation has great impact on ASR performance, particularly for spontaneous speech, and has been a focus of much research (e.g [36][68][141]). We will return to this issue in later chapters viewed from the perspective of the syllable, articulatory-acoustic features and stress accent.

2.3 Summary

This chapter has described the linguistic dissection of several state-of-the-art LVCSR systems using the Switchboard corpus. System outputs at both the word and phonetic-segment levels were compared to manually annotated transcripts at the word, syllable, phonetic-segment and prosodic stress-accent levels. Detailed statistical analysis of recognition error patterns was performed with respect to dozens of linguistic and acoustic parameters. Additional information on file format conversion, phone mapping and scoring procedure is given in Appendix A.

Some of the major findings from the linguistic dissection are as follows:

- There exists a high correlation between word and phone-error rates, suggesting that performance at the word level largely depends on phone-level classification and that improving acoustic modeling is likely to yield better word recognition performance.
- For correctly recognized words the tolerance for phone errors is roughly constant for words with four phones or less; for incorrectly recognized words, the number of phone errors increases in quasi-linear fashion as a function of word length.
- Syllable structure is a significant factor in determining word-error rate, especially word deletion. Words of vowel-initial syllable forms exhibit a much higher word-error rate than words of consonant-initial forms; monosyllabic words generally exhibit higher word-error rates than polysyllabic words.
- The levels of tolerance of articulatory-feature errors differ depending on both the position within the syllable and the particular AF dimension. Manner and place of articulation of onsets and codas are particularly important for correctly distinguishing among words. Classification of vocalic nuclei is considerably less precise than for the onsets and codas.
- Entirely unaccented words tend to have a much higher error rate (especially deletion rate) than words having at least some stress accent.
- Speaking rate (in terms of syllables per second) is a factor of word recognition error. Very slow and very fast speech tend to have much higher word-error rate than speech at normal tempo.
- The number of pronunciation variants per word is usually much smaller in the evaluation system outputs than in the reference transcripts, and a high correlation was observed between the word-correct rate and the average number of pronunciation variants per word across recognition sites.

Results of the linguistic dissection of LVCSR systems suggests there is a significant gap between models and the observed data. Pronunciation models are inadequate to capture the pronunciation variation phenomena of spontaneous speech. Useful prosodic features, such as stress accent and speaking rate, are rarely taken into account explicitly in conventional models. ASR systems have not made sufficient use of important information contained in linguistic levels other than the phonetic tier, such as the syllable and the articulatory features. These findings motivate the development of an alternative model of speech recognition in the remainder of the thesis.

Chapter 3

Articulatory-acoustic Features

The linguistic dissection results described in the previous chapter highlighted a number of important factors affecting recognition performance of many ASR systems. In particular, it emphasized the relationship between word-recognition errors and errors made at the phonetic-segment level. When phonetic segments are decomposed into more granular articulatory-acoustic features (AFs), a number of interesting patterns emerge. The tolerance of AF errors depends largely on the specific feature dimension and position within a syllable. For example, manner- and place-of-articulation dimensions are particularly prone to error in incorrectly recognized words; nuclei exhibit a greater tolerance for AF errors than onsets and codas in correctly recognized words. Overall, AF errors found in correctly recognized words are about one third that of incorrectly recognized words, suggesting a relatively small amount of AF deviations are tolerated without a significant impact on word recognition. In Chapters 4 and 5, additional analysis of AF deviations between canonical and realized forms will be discussed that reveal a systematic pattern of AF deviations as a function of syllable position and stress-accent level.

Together, this evidence suggests that it may be beneficial to incorporate information at the AF level in speech modeling. This chapter describes our approach to the automatic extraction of AFs and provides evidence in support of using AFs as fundamental building blocks of speech models, especially for spontaneous speech. But first, a brief description of AFs is provided and previous work by other researchers in the field is described.

3.1 Background and Previous Work

Speech is a communication process between a speaker and a listener. A speech sound is generated by specific articulatory movement of the speaker, travels through a medium such as air, and reaches the listener's ear. It is thus not unreasonable to describe speech sounds by the articulatory configuration associated with them. However, the ultimate goal of the recognition process at the listener's end is to deduce the speaker's intended meaning carried in the acoustic signal rather than to produce a completely faithful characterization of the articulatory configuration of the speaker. Therefore, it may be advantageous to use an abstract representation that captures the essential aspects of articulation, but at

the same time, possesses specific correlates in the acoustic signal. Such a representation of speech can be made in terms of articulatory-acoustic features (AFs) (interested readers may refer to [84][78][126] for more information on the theory of articulatory phonetics).

One very commonly used AF is *voicing*, which describes the state of the glottis as to whether the vocal folds are vibrating during articulation. This feature usually assumes a value of *voiced* or *voiceless* but may also be used to characterize *breathy voice* (*murmur*) and *creaky voice* (*larygenalized*). *Manner*-of-articulation characterizes the type of articulatory closure and degree of obstruction of the airstream associated with the articulators. Commonly encountered manner-of-articulation classes in English include *vocalic*, *nasal*, *stop* (*plosive*), *fricative* (including *affricate*), *flap* and *approximant*. A separate AF dimension pertains to the locus of maximum articulatory constriction – *place* of articulation, which may assume a number of different values ranging from *(bi)labial* to *glottal*. The possible values of *place* of articulation that a sound can assume depends on its *manner* of articulation. For example in American English, a *fricative* may have a place value of *labio-dental* (e.g. [f]), *inter-dental* (e.g. [th]), *alveolar* (e.g. [s]) or *palatal* (e.g. [sh]), while a *stop* may have *bilabial* (e.g. [p]), *alveolar* (e.g. [t]) or *velar* (e.g. [k]) place of constriction. For the *vocalic* sounds the *horizontal place* dimension may be categorized into *front*, *central* and *back*, and it is closely related to the difference between the second and the first formant frequencies ($f_2 - f_1$). Together, these three AF dimensions distinguish the majority of phonetic segments in American English. Several other AF dimensions are also useful: vocalic *height* describes the height of the tongue body during vocalic production and is closely related to the first-formant frequency f_1; *lip-rounding* describes whether the lips are rounded (or not) and is reflected in f_1 and f_2; vocalic *tenseness* distinguishes between *lax* (e.g. [ih],[eh],[uh],[ax],[ix] in English) and *tense* vowels (e.g. the diphthongs plus [ao],[aa],[ae]); and spectrally *dynamic* distinguishes between monophthongs (with a relatively stable spectrum) and diphthongs (with a more dynamic spectrum).

There has been an increasing set of attempts to use articulatory-based features in speech recognition in recent years. Several studies have physically measured articulatory configuration data in conjunction with simultaneously recorded acoustic data. These data may be measured by an x-ray microbeam [140], laryngograph [9] or electro-magnetic articulograph [38]. Some of these data have been used to develop algorithms for direct inversion from acoustics to articulatory configurations [102][147], with some limited success. However, such physically measured articulatory data are generally unavailable for most commonly used speech corpora. Some researchers have used manually labeled phonetic transcripts for some corpora and the canonical mapping from phonetic-segments to predefined articulatory features, and used statistical machine learning techniques (such as neural networks [76][73] and Bayesian networks [113]) to train classifiers of articulatory features. The approach we take for extracting AFs from acoustic input in this work also falls into this category. Others have developed constrained systems (e.g. linear dynamical systems or HMMs) with articulatory states as latent variables (often to replace the conventional phonetic states). Such systems do not necessarily require articulatory data during training, but the resulting articulatory trajectories are often compared to the physically measured articulatory configurations for evaluation. For example, Hogden et al. [64] developed a maximum-likelihood training algorithm for an articulatory-constrained continuity mapping of acoustics. Richard-

son et al. [109] introduced a hidden-articulatory Markov model for speech recognition where the hidden states denote articulatory configurations. Several researchers have adopted dynamic Bayesian networks to capture the relationship between acoustics and articulatory features [151][125]. Others have attempted to recover articulatory trajectories with linear dynamical systems [38] and articulatory-constrained HMMs [111]. In a series of developments over the past decade, Deng and colleagues [24][23][27][25][129] have introduced elaborate HMM systems of overlapping articulatory features incorporating phonological rules in the design process. In their most recent development [129], an overlapping-feature-based phonological model that represents long-span contextual dependencies and high-level linguistic constraints showed significant improvements over the conventional triphone-based models on the TIMIT corpus [79].

Several advantages of incorporating articulatory features have been noted in previous studies:

- AFs are more flexible than traditional phonetic segments in building accurate pronunciation models capable of capturing non-canonical realizations (especially useful in modeling spontaneous speech);
- as the basic building block of the phonetic tier of speech, AFs can be combined in many different ways to specify speech sounds found in a wide variety of languages, and are thus more likely to be cross-linguistically adaptable;
- classification in terms of broadly based AFs is likely to achieve better performance than phonetic segments, and is likely to be more robust to variations in speaking style and under adverse acoustic-interference conditions;
- different AF dimensions may contribute differentially to recognition and may benefit from being treated separately.

The following sections describe our approach to automatic extraction of AFs from acoustic inputs. Our results are consistent with the advantages of using AFs enumerated above. In conjunction with the description of syllable-level processing and stress-accent patterns in subsequent chapters, we also argue that AFs are closely tied to syllable structure and vary more systematically as a function of syllable position and stress-accent patterns than phonetic segments.

3.2 Automatic Extraction of Articulatory-acoustic Features

This section describes the system that we have developed for automatic extraction of articulatory-acoustic features from speech input [14], performance evaluation of the system and extension to automatic phonetic labeling.

3.2.1 System Description

The system described in this section contains two stages – front-end processing and classification of AFs using neural networks. An overview of the two processes is illustrated

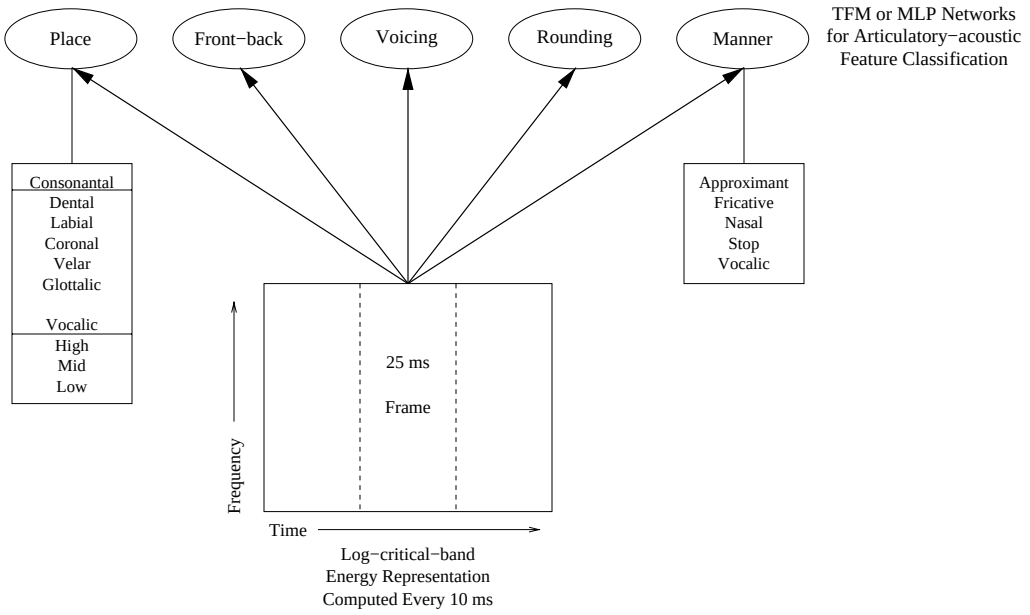

Figure 3.1: Illustration of the neural-network-based AF classification system. Each oval represents a Temporal Flow Model (TFM) or Multi-Layer Perceptron (MLP) network for recognition of an AF dimension. See text for detail.

in Figure 3.1.

Pre-Processing

The speech signal is converted into a spectro-temporal representation (log-compressed critical-band energy features) in the following manner. First, a power spectrum is computed every 10 ms (over a 25-ms window, referred to as a frame) and this spectrum partitioned into critical-band-like channels between 0.3 and 3.4 kHz using Bark-scale trapezoidal filters similar to those used in the PLP preprocessing [61]. The power spectrum is logarithmically compressed in order to preserve the general shape of the spectrum distributed across frequency and time (an example of which is illustrated in Figure 3.3 for the manner-of-articulation features, *vocalic* and *fricative*).

Neural Networks for AF Classification

An array of independent neural networks classify each 25-ms frame along the AF dimensions of interest using the log-compressed, critical-band energy features. Each neural network has a number of output nodes that correspond to the possible feature values of the particular AF dimension plus a separate class for "silence." Two types of neural networks have been used in our experiments for AF classification: temporal-flow model (TFM) networks [137] and multi-layer perceptrons (MLP) [6].

A typical MLP network used in our system possesses a single hidden layer of nodes (often referred to as a two-layer network because of its two layers of active links). The MLPs are fully connected between the input nodes and the hidden nodes, as well as between the hidden and output nodes. Additional bias (threshold) links with constant input values are connected to the hidden and output nodes. The hidden nodes use logistic activation functions and the output nodes have softmax activation functions [6].

A TFM network supports arbitrary link connectivity across multiple layers of nodes, admits feed-forward as well as recurrent links, and allows variable propagation delays to be associated with links (cf. Figures 3.2). The recurrent links in TFM networks provide an effective means of smoothing and differentiating signals as well as detecting the onset (and measuring the duration). Using multiple links with variable delays allows a network to maintain an explicit context over a specified window of time and thereby makes it capable of performing spatiotemporal feature detection and pattern matching. Recurrent links, used in tandem with variable propagation delays, provide a powerful mechanism for simulating certain properties (such as shortterm memory, integration and context sensitivity) essential for processing time-varying signals such as speech. In the past TFM networks have been successfully applied to a wide variety of pattern-classification tasks including phoneme classification [138][135], optical character recognition [35] and syllable segmentation [117].

The architecture of the TFM networks used for classification of articulatory acoustic features was manually tuned using a three-dimensional representation of the log-power-spectrum distributed across frequency and time that incorporates both the mean and variance of the energy distribution associated with multiple (typically, hundreds or thousands of) instances of a specific phonetic feature or segment derived from the phonetically annotated, OGI Stories-TS corpus [11]. Each phonetic-segment class was mapped to an array of AFs, and this map was used to construct the spectro-temporal profile (STeP) for a given feature class. For example, the STeP for the manner feature, *vocalic* (cf. Figure 3.3, upper-panel), was derived from an average over all instances of vocalic segments in the corpus. The STeP extends 500 ms into the past, as well as 500 ms into the future relative to the reference frame (time 0), thereby spanning an interval of 1 second, similar to that used in TRAPs [62]. This extended window of time is designed to accommodate co-articulatory context effects. The frequency dimension is partitioned into critical-band-like, quarter-octave channels. The variance associated with each component of the STeP is color-coded and identifies those regions which most clearly exemplify the energy-modulation patterns across time and frequency associated with the feature of interest (cf. Figure 3.3) and can be used to adjust the network connectivity in appropriate fashion.

The training targets for both the MLP and TFM networks were derived from manually labeled phonetic transcripts via a canonical mapping from phones to AFs (e.g. Tables 3.1 and 3.4). For each 25-ms frame the target for the output node corresponding to the desired AF value was assigned a "1" and the remaining classes "0." Both types of networks use a minimum cross-entropy error function [108][6] during training, which in conjunction with the "1/0" target assignment, trains the network outputs to approximate posterior probabilities of each target class given the input. The inputs to the MLP networks at each frame contains not only the pre-processed, spectro-temporal features of the current

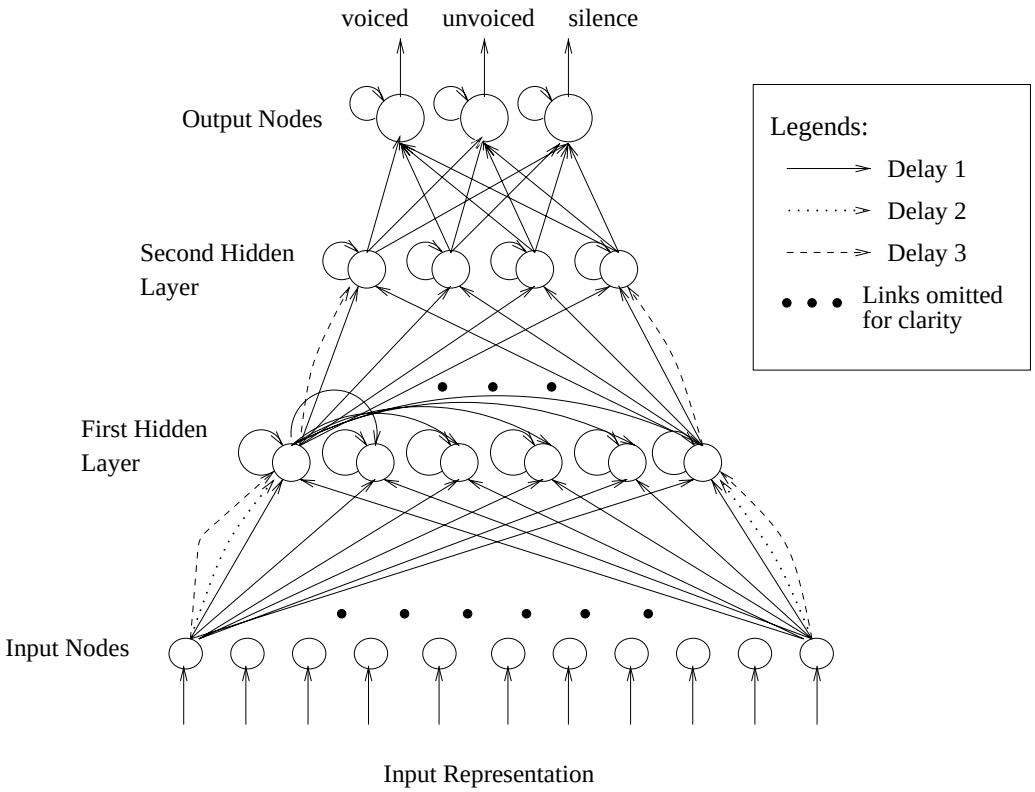

Figure 3.2: A typical example of a Temporal Flow Model (TFM) network for the voicing classification. Actual number of layers, number of nodes and link connectivity may differ depending on the specific classification task. TFM networks support arbitrary connectivity across layers, provide for feed-forward, as well as recurrent links, and allow variable propagation delays across links.

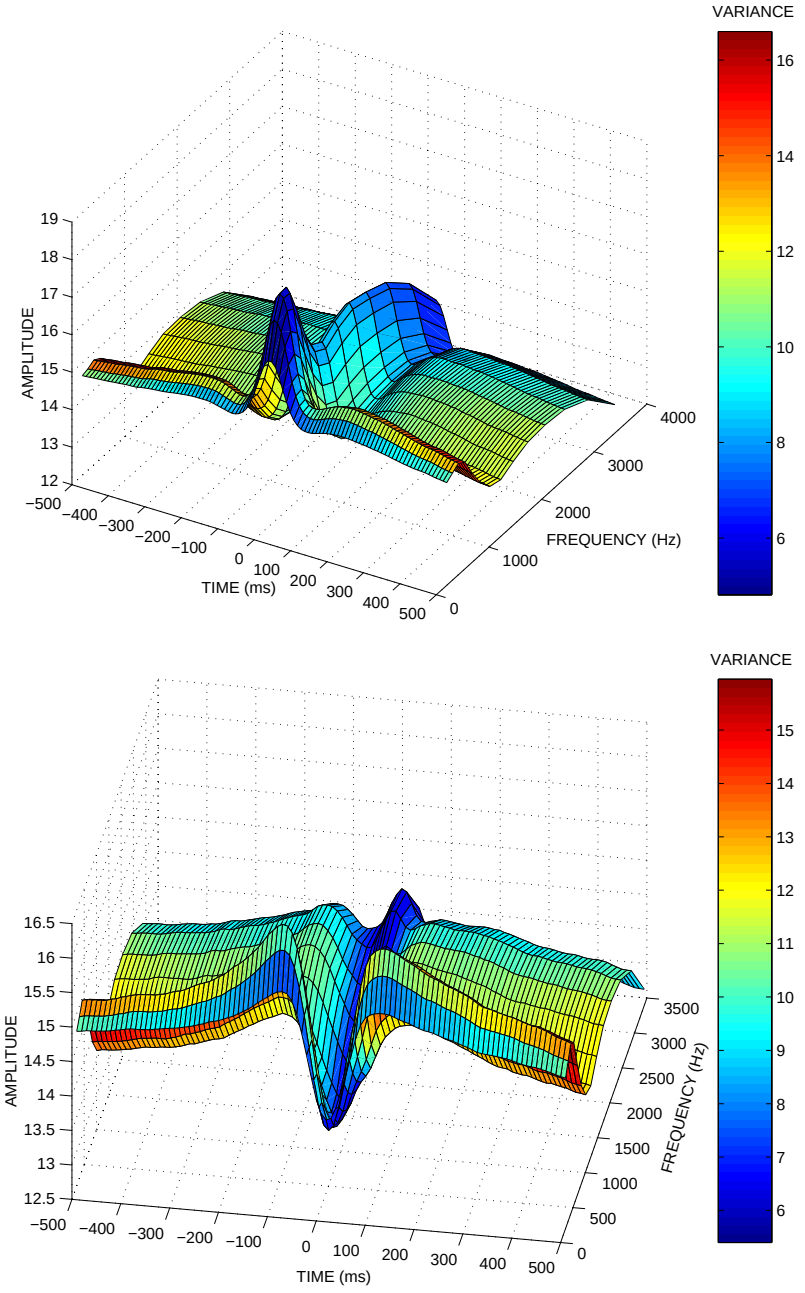

Figure 3.3: Spectro-temporal profiles (STePs) of the manner features, *vocalic* (upper-panel) and *fricative* (lower-panel), computed from the OGI Stories-TS corpus [11]. Each STeP represents the mean (by amplitude) and variance (by color-coding) of the energy distribution associated with multiple (typically, hundreds or thousands of) instances of a specific phonetic feature or segment. The frequency dimension is partitioned into critical-band-like, quarter-octave channels.

frame but also that of several adjacent frames (preceding and following the current one) to simulate a temporal context [90]. In contrast, the TFM networks only require the input features from the current frame since their time-delay and recurrent links implicitly provide a context of variable duration over time.

The MLP networks were trained with a standard, online, back-propagation algorithm [112] adapted to speech processing [90]. The TFM networks were trained with a back-propagation-through-time (BPTT) [136] algorithm coupled with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [6][136], a second-order, gradient-based optimization algorithm. The MLP and TFM networks require comparable computation for each epoch through training data (for networks with the same number of active parameters). However, the TFM network requires far more epochs to converge, a likely consequence of the diminishing gradients propagated through time [150]. As a consequence it often takes much longer to train a TFM network than to train an MLP network even when the latter contains a greater number of active parameters. Therefore, despite of the modeling advantages offered by TFM networks, most of our experiments were conducted using MLP networks because of its training efficiency.

3.2.2 Evaluation

Initial experiments to evaluate the AF classification were performed on the Numbers95 corpus [12], comprising spontaneous speech material collected and phonetically annotated (i.e., labeled and segmented) at the Oregon Graduate Institute. This corpus contains the numerical portion (mostly street addresses, phone numbers and zip codes) of thousands of telephone dialogues and contains a lexicon of 32 words and an inventory of roughly 30 phonetic segments. The speakers in the corpus were of both genders and represent a wide range of dialect regions and age groups. The AF classifiers were trained on ca. 2.5 hours of material with a separate 15-minute cross-validation set. The AF targets were derived from the manually labeled phonetic transcription with a fixed phone-to-AF mapping, as shown in Table 3.1. Testing and evaluation of the system was performed on an independent set of ca. one hour's duration.

The accuracy of the TFM-based AF classification ranges between 79% (place of articulation) and 91% (voicing) (cf. Table 3.2). The MLP-based AF classification (with a nine-frame context) achieved slightly lower accuracies (cf. Table 3.2), while using almost an order-of-magnitude more adjustable, active parameters than the TFM counterparts.

3.2.3 Extension to Automatic Phonetic Labeling

It is also possible to perform phonetic-segment classification from the AF classification results by using another neural network that maps the AF probabilities obtained at the output of the TFM or MLP networks onto phonetic-segment labels, similar to the approach used by Kirchoff [76]. In our experiments, the phone classification was carried out by an MLP network with a single hidden layer of between 200 (for a MLP-based AF classification) and 400 (for TFM-based AF classification) units to maintain a relative balance between the total numbers of free parameters in the two systems. A context window

Phone	Voicing	Manner	Place	Front-Back	Rounding
d	voice+	stop	coronal	nil	nil
t	voice-	stop	coronal	nil	nil
k	voice-	stop	velar	nil	nil
s	voice-	fricative	coronal	nil	nil
z	voice+	fricative	coronal	nil	nil
f	voice-	fricative	labial	nil	nil
th	voice-	fricative	dental	nil	nil
v	voice+	fricative	labial	nil	nil
hh	voice-	fricative	glottal	nil	nil
n	voice+	nasal	coronal	nil	nil
l	voice+	approximant	coronal	nil	nil
r	voice+	approximant	rhotic	nil	nil
w	voice+	approximant	labial	nil	round+
y	voice+	approximant	high	nil	nil
hv	voice+	approximant	glottal	nil	nil
iy	voice+	vocalic	high	front	round-
ih	voice+	vocalic	high	front	round-
eh	voice+	vocalic	mid	front	round-
ey	voice+	vocalic	mid	front	round-
ae	voice+	vocalic	low	front	round-
aa	voice+	vocalic	low	back	round-
aw	voice+	vocalic	low	back	round+
ay	voice+	vocalic	low	front	round-
ah	voice+	vocalic	mid	back	round-
ao	voice+	vocalic	low	back	round-
ow	voice+	vocalic	mid	back	round+
uw	voice+	vocalic	high	back	round+
er	voice+	vocalic	rhotic	nil	round-
ax	voice+	vocalic	mid	back	round-
h#	silence	silence	silence	silence	silence

Table 3.1: Phone-to-AF mapping for the AF classification experiments on the Numbers95 corpus. The mappings were adapted from Kirchhoff [76].

Network	Front-back	Lip-rounding	Manner	Place	Voicing
TFM	83.4	85.6	84.4	78.8	91.1
MLP	82.6	83.4	82.6	75.0	89.8

Table 3.2: Frame-level TFM- and MLP-based AF classification accuracy (percentage) on the Numbers95 corpus development test set.

of 9 frames (105 ms) was used by the MLP network. The output of this MLP contains a vector of phone-posterior-probability estimates for each frame and was evaluated for its accuracy with respect to manually labeled phonetic transcripts. The TFM/MLP system achieved a frame-level phone accuracy of 79.4% and the MLP/MLP system had 78.1%. For comparison, we also computed phonetic-segment classification accuracy obtained by a direct mapping from log-compressed, critical-band energy features to phonetic-segment classes using an MLP network without the intermediate stage of AF classification; the result was 73.4% when the number of adjustable parameters used was similar to the total number of parameters in the TFM/MLP system.

This matrix of phonetic-posterior-probabilities over time can be further converted into a linear sequence of phone labels and segmentation boundaries via a decoder. A hidden-Markov-model (HMM) was applied to impose a minimum-length constraint on the duration associated with each phonetic-segment (based on segmental statistics of the training data), and a Viterbi-like decoder with a phone-bigram model (derived from the training data) was used to compute the sequence of phonetic segments over the entire length of the utterance. This bipartite phone-decoding process is analogous to that used for decoding word sequences in ASR systems. However, in the present application the "lexical" units are phones, rather than words, and the "words" contain clusters of articulatory features rather than phones. A 19.3% phone-error rate (8.1% substitution, 6.4% deletion and 4.9% insertion) was obtained with this approach. Figure 3.4 illustrates the sample output of this automatic phonetic labeling system (ALPS) on one of the test-set utterances.

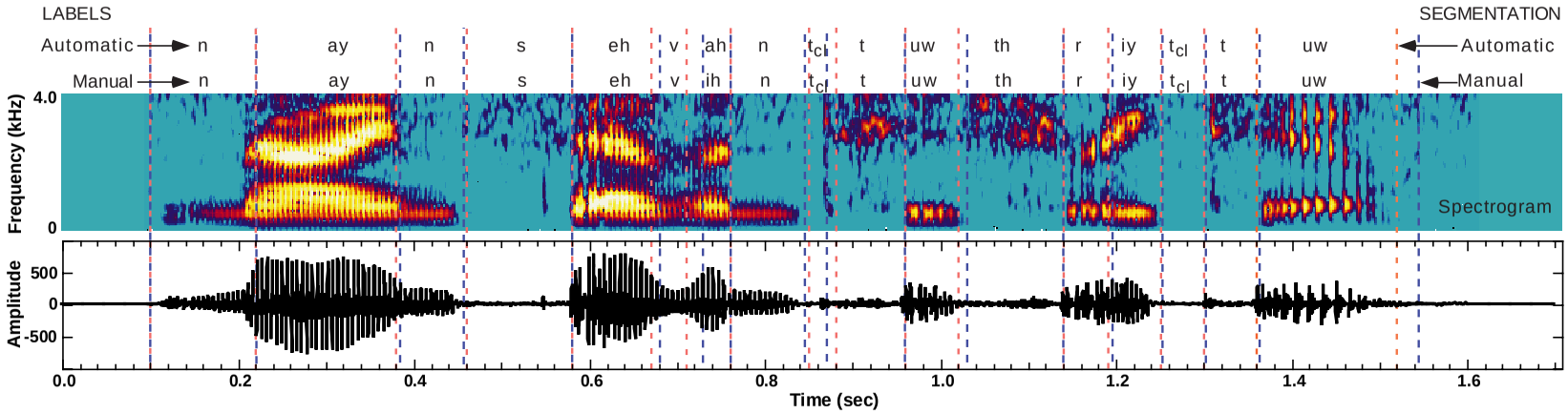

Figure 3.4: The labels and segmentation generated by the automatic phonetic transcription system for the utterance "Nine, seven, two, three, two" are compared to those produced manually. The top row shows the phone sequence produced by the automatic system. The tier directly below is the phone sequence produced by a human transcriber. The spectrographic representation and waveform of the speech signal are shown below the phone sequences as a means of evaluating the quality of the phonetic segmentation. The manual segmentation is marked in purple, while the automatic segmentation is illustrated in orange. From [14].

Frame Tolerance	Hits	False Alarms
± 1 (10 ms)	38.4	58.5
± 2 (20 ms)	76.0	20.9
± 3 (30 ms)	83.7	13.2

Table 3.3: Accuracy of phonetic segmentation as a function of the temporal tolerance window and partitioned into error type (hits/false alarms).

It is of interest to ascertain the frame location of phonetic-segment classification errors as a means of gaining insight into the origins of mislabeling for this material. Specifically, it is important to know whether the classification errors are randomly distributed across frames or are concentrated close to the segment boundaries. The data illustrated in Figure 3.5 indicate that a disproportionate number of errors are concentrated near the phonetic-segment boundaries in regions inherently difficult to classify accurately as a consequence of the transitional nature of phonetic information in such locations. Nearly a third of the phone classification errors are associated with boundary frames associated with just 17% of the utterance duration. The accuracy of phone classification is only 61% in the boundary frames, but rises to 80% or higher for frames located in the central region of the phonetic segment.

The accuracy of phonetic segmentation can be evaluated by computing the proportion of times that a phonetic segment onset is correctly identified (hits) by the ALPS system relative to the instances where the phone onset (as marked by a human transcriber) is located at a different frame (false alarms). The data in Table 3.3 indicate that the ALPS system matches the segmentation of human transcribers precisely in ca. 40% of the instances. However, automatic segmentation comes much closer to approximating human performance when a tolerance level of more than a single frame is allowed (76-84% concordance with manual segmentation). The average deviation between the manual and automatic segmentation is 11 ms, an interval that is ca. 10% of the average phone duration in the Numbers95 corpus.

3.3 Manner-specific Training and the "Elitist" Approach

In subsequent experiments we have applied our approach to AF classification on a separate American English corpus with a much larger vocabulary and a more balanced phonetic inventory than that of the Numbers95 corpus. This section describes these experiments and, in particular, an "elitist" approach to delineate regions of speech with high confidence in AF classification is described; a manner-specific training scheme for enhancing classification of place and other AF dimensions [13][17] is also presented.

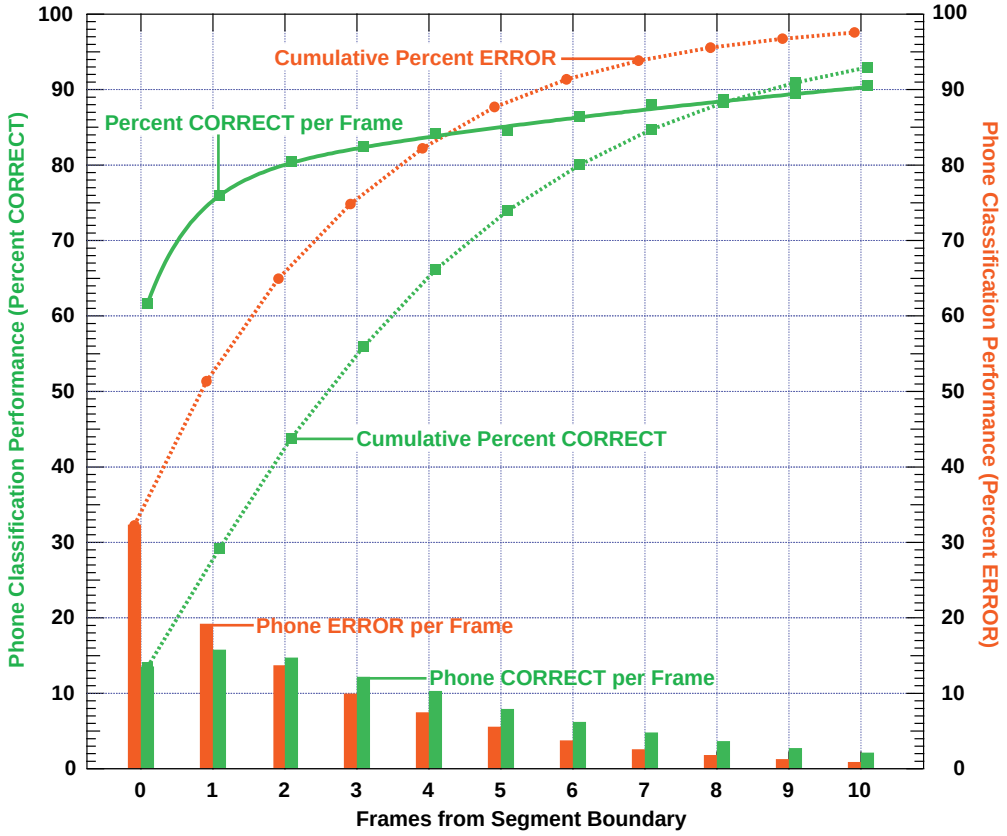

Figure 3.5: Phonetic-segment classification performance as a function of frame (10 ms) distance from the manually defined phonetic-segment boundary. Contribution of each frame to the total number of correct (green) and incorrect (orange) phonetic segments classified by the automatic system is indicated by the bars. The cumulative performance over frames is indicated (dotted lines), as is the percent correct phone classification for each frame (green squares, with a double-exponential, solid-line fit to the data). From [14].

3.3.1 AF Classification on the NTIMIT Corpus

The AF classification approach described in the previous section was applied to a subset of the NTIMIT corpus with 3300 sentences (comprising 164 minutes of speech) for training and 393 sentences (19.5 minutes) for testing. NTIMIT ([65]) is a spectrally restricted variant of the TIMIT corpus (8-kHz bandwidth; cf. [79]), that has been passed through a phone network (between 0.3 and 3.4 kHz), providing an appropriate set of materials with which to develop a phonetic annotation system destined for telephony-based applications. The corpus contains a quasi-phonetically balanced set of sentences read by native speakers (of both genders) of American English, whose pronunciation patterns span a wide range of dialectal variation. The phonetic inventory of the NTIMIT corpus is listed in Table 3.4, along with the articulatory-feature equivalents for each segment. The phonetic annotation (both labeling and segmentation) associated with the NTIMIT corpus was manually annotated at MIT by trained linguistic personnel on the spectrally unfiltered version of the TIMIT corpus.

For the NTIMIT experiments we have used an MLP-based AF classification system similar to that used with the Numbers95 corpus except that: the AF dimensions being classified and the corresponding phone-to-AF mappings were different (cf. Table 3.4); the input features to the MLP networks also included the deltas (first-derivatives) pertaining to the spectro-temporal contour over both time and frequency dimensions.

Table 3.5 shows the overall frame-level AF classification accuracies for seven different AF dimensions obtained using the NTIMIT corpus. Within each AF dimension not all feature values achieved the same classification accuracy (cf. Table 3.6 for a confusion matrix for the place of articulation features) and this variability in performance reflects to a certain degree the amount of training material available for each feature.

3.3.2 An "Elitist" Approach

The previous section discussed an observation made on the Numbers95 phonetic-segment classification output – a disproportionate number of errors are concentrated near the phonetic-segment boundaries (cf. Figure 3.5). A similar analysis was performed on the AF classification output from the NTIMIT corpus. With respect to feature classification, just as in phonetic classification, not all frames are created equal. For the manner-of-articulation features, the 20% frames that are closest to the segmental borders have an average frame accuracy of 73%, while the 20% of frames closest to the segmental centers have an average frame accuracy of 90%. This "centrist" bias in feature classification is paralleled by a concomitant rise in the "confidence" with which MLPs classify AFs. This is similar to the high correlation between the posterior probability estimates and phone-classification accuracy observed in connectionist speech recognition [8]. Figure 3.6 illustrates this phenomenon by displaying the average frame accuracy of manner-of-articulation classification and the average maximum MLP output as a function of frame position within a segment for all frames, as well as for vocalic and consonantal frames analyzed separately.

This observation suggests that we may use the maximum MLP output at each frame as an objective metric with which to select frames most "worthy" of manner des-

CON	*Manner*	*Place*	*Voi*	*Sta*	*APPR*	*Height*	*Place*	*Voi*	*Sta*
[p]	Stop	Bilabial	-	-	[w]*	High	Back	+	-
[b]	Stop	Bilabial	+	-	[y]	High	Front	+	-
[t]	Stop	Alveolar	-	-	[l]	Mid	Central	+	-
[d]	Stop	Alveolar	+	-	[el]	Mid	Central	+	-
[k]	Stop	Velar	-	-	[r]	Mid	Rhotic	+	-
[g]	Stop	Velar	+	-	[er]	Mid	Rhotic	+	-
[ch]	Fric	Alveolar	-	-	[axr]	Mid	Rhotic	+	-
[jh]	Fric	Alveolar	+	-	[hv]	Mid	Central	+	-
[f]	Fric	Lab-den	-	+					
[v]	Fric	Lab-den	+	+					
[th]	Fric	Dental	-	+	*VOW*	*Height*	*Place*	*Ten*	*Sta*
[dh]	Fric	Dental	+	-	[ix]	High	Front	-	+
[s]	Fric	Pre-alv	-	+	[ih]	High	Front	-	+
[z]	Fric	Pre-alv	+	+	[iy]	High	Front	+	-
[sh]	Fric	Post-alv	-	+	[eh]	Mid	Front	-	+
[zh]	Fric	Post-alv	+	+	[ey]	Mid	Front	+	-
[hh]	Fric	Glottal	-	+	[ae]	Low	Front	+	+
[m]	Nasal	Bilabial	+	+	[ay]	Low	Front	+	-
[n]	Nasal	Alveolar	+	+	[aw]*	Low	Central	+	-
[ng]	Nasal	Velar	+	+	[aa]	Low	Central	+	+
[em]	Nasal	Bilabial	+	-	[ao]	Low	Back	+	+
[en]	Nasal	Alveolar	+	-	[oy]	Mid	Back	+	-
[eng]	Nasal	Velar	+	-	[ow]*	Mid	Back	+	-
[nx]	Flap	Alveolar	+	+	[uh]	High	Back	-	+
[dx]	Flap	Alveolar	+	-	[uw]*	High	Back	+	-

Table 3.4: Articulatory-acoustic feature specification of phonetic segments developed for the American English (N)TIMIT corpus. An asterisk (*) indicates that a segment is lip-rounded. The consonantal segments are marked as "nil" for the feature tense. "*Voi*" is the abbreviation for "voicing," "*Sta*" for "Static," "*Ten*" for "Tense," "CON" for "consonant," "APPR" for "approximant" and "VOW" for "vowel." The phonetic orthography is a variant of ARPABET.

Lip-rounding	Manner	Place	Static	Voc-height	Voc-tense	Voicing
82.9	85.0	71.2	76.6	80.9	81.9	88.9

Table 3.5: Overall frame-level AF classification accuracy (percent correct) on the NTIMIT corpus.

	Consonantal Segments					Vocalic Segments				N-S
Reference	Lab	Alv	Vel	Den	Glo	Rho	Frt	Cen	Bk	Sil
Labial	*60*	24	03	01	01	01	02	02	01	05
Alveolar	06	*79*	05	00	00	00	03	02	00	05
Velar	08	23	*58*	00	00	00	04	01	01	05
Dental	29	40	01	*11*	01	01	05	03	01	08
Glottal	11	20	05	01	*26*	02	15	10	03	07
Rhotic	02	02	01	00	00	*69*	10	09	06	01
Front	01	04	01	00	00	02	*82*	07	02	01
Central	02	03	01	00	01	02	12	*69*	10	00
Back	03	02	01	00	00	04	17	24	*48*	01
Silence	03	06	01	00	00	00	00	00	00	*90*

Table 3.6: A confusion matrix illustrating classification performance for place-of-articulation features from manner-independent training. The data are partitioned into consonantal and vocalic classes. Silence is classified as non-speech (N-S). All numbers are percent of total frames of the reference features.

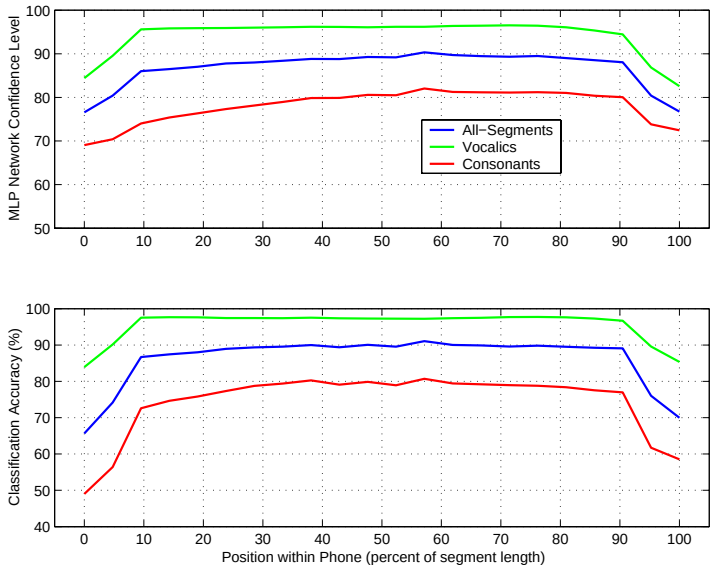

Figure 3.6: The relation between frame classification accuracy for manner of articulation on the NTIMIT corpus (bottom panel) and the MLP output confidence level (i.e., maximum MLP output magnitude) as a function of frame position within a phonetic segment (normalized to the duration of each segment). Frames closest to the segmental boundaries are classified with the least accuracy; this performance decrement is reflected in a concomitant decrease in the MLP confidence magnitude.

	Vocalic		Nasal		Stop		Fricative		Flap		Silence	
Ref	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best
Vocalic	*96*	*98*	02	01	01	01	01	00	00	00	00	00
Nasal	14	10	*73*	*85*	04	02	04	01	01	00	04	02
Stop	09	08	04	02	*66*	*77*	15	09	00	00	06	04
Fric	06	03	02	01	07	03	*79*	*89*	00	00	06	04
Flap	29	30	12	11	08	04	06	02	*45*	*53*	00	00
Silence	01	01	02	00	03	01	05	02	00	00	*89*	*96*

Table 3.7: The effect of the "elitist" approach for selecting frames with a high confidence of manner classification. All numbers are in terms of percent of total frames of the reference features. "All" refers to the manner-independent system using all frames of the signal, while "Best" refers to the frames exceeding a 70% threshold. The confusion matrix illustrates the pattern of classification errors.

ignation and we call such a method of delineating the relative importance of frames the "elitist" approach. By establishing a network-output threshold of 70% (relative to the maximum) for frame selection, the selected frames (with maximum MLP output greater than the threshold) yield manner-of-articulation classification performance between 2% and 14% (absolute) greater than that applied to all frames, as illustrated in Table 3.7 and Figure 3.7. Most of the frames discarded are located in the interstitial region at the boundary of adjacent segments. The overall accuracy of manner classification for the selected frames is 93% (compared to 85% for all frames). A potential utility of this approach is to provide a quantitative basis for differential treatment of the various regions of speech signal; the more confidently classified regions are likely to provide more useful information for recognition.

3.3.3 Manner-Specific Training

In the experimental results presented in Table 3.6 for manner-independent classification place-of-articulation information was correctly classified in 71% of the frames; the accuracy for individual place features ranged between 11% and 82%. There are ten distinct places of articulation across the manner classes (plus silence), making it difficult to effectively train networks expert in the classification of each place feature. There are other problems as well. For example, the loci of maximum articulatory constriction for stops differ from those associated with fricatives. And the articulatory constriction has a different manifestation for consonants and vowels. The number of distinct places of articulation for any given manner class is usually just three or four. Thus, if it were possible to identify manner features with a high degree of assurance it should be possible, in principle, to train an articulatory-place classification system in a manner-specific manner that could potentially enhance place-feature extraction performance.

Figure 3.8 illustrates a manner-specific training scheme for place-of-articulation classification. Separate MLPs are trained to classify place-of-articulation features for each of the five manner classes – stops, nasals, fricatives, flaps and vowels (the latter includes the

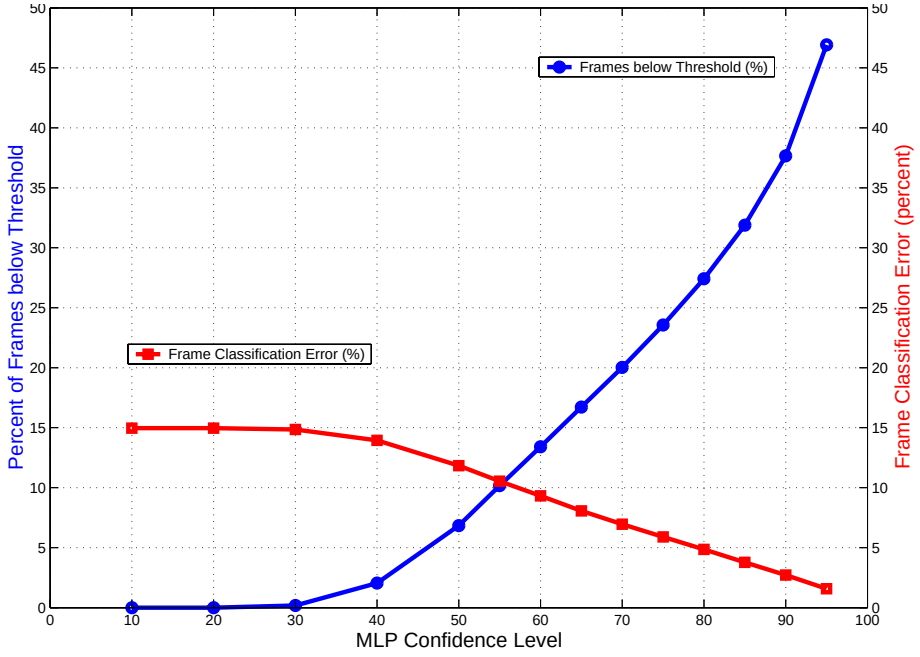

Figure 3.7: Trade-off between the proportion of frames falling below threshold and frame-error rate for the remaining frames for different threshold values (MLP confidence level – the maximum MLP output value at each frame) for manner classification on the NTIMIT corpus.

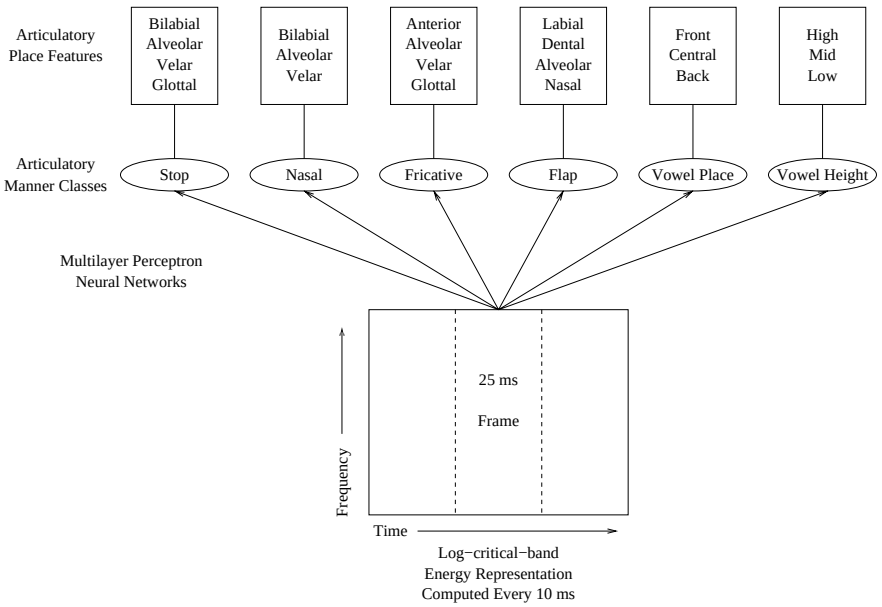

Figure 3.8: The manner-dependent, place-of-articulation classification system for the NTIMIT corpus. Each manner class contains between three and four place-of-articulation features. Separate MLP classifiers were trained for each manner class.

approximants), and each MLP is trained only on frames associated with the corresponding manner class. The place dimension for each manner class is partitioned into three basic features. For consonantal segments the partitioning corresponds to the relative location of maximal constriction – anterior, central and posterior (as well as the glottal feature for stops and fricatives). For example, "bilabial" is the most anterior feature for stops, while the "labio-dental" and "dental" loci correspond to the anterior feature for fricatives. In this fashion it is possible to construct a relational place-of-articulation pattern customized to each consonantal manner class. For vocalic segments front vowels were classified as anterior, and back vowels as posterior. The liquids (i.e., [l] and [r]) were assigned a "central" place given the contextual nature of their articulatory configuration. Other non-place AF dimensions, such as vocalic height (cf. Figure 3.8), can also be classified in a manner-specific fashion.

Table 3.8 illustrates the performance of such manner-specific, place classification. In order to characterize the potential efficacy of the method, manner information for the test materials was derived from the reference labels for each segment rather than from automatic classification of manner of articulation. Under this evaluation condition, manner-specific, place classification performance for most of the manner classes are significantly better than that of manner-independent classification. This gain in classification performance is most likely derived from two specific factors – (1) a more homogeneous set of training material for manner-specific, place classification and (2) a smaller number of place-feature targets for each manner class. It should be noted that the aim of manner-specific training is to enhance the classification performance of features that exhibit manner-dependency, rather

than to provide a prescription for phonetic-segment classification (as was the goal in other research, such as the hierarchical connectionist acoustic modeling of Fritsch [39]).

3.4 Cross-linguistic Transfer of AFs

As described earlier in this chapter, a potential advantage of modeling speech with AFs is cross-linguistic transferability. Because of the universal set of articulators (vocal folds, tongue, lips, etc.) and auditory apparatus (ear, auditory pathway, etc.) shared by speakers of all languages of the world, many of the same articulatory features are present in different languages. For example, virtually all languages make the distinction between vocalic and consonantal sounds, as well as between voiced and unvoiced sounds. Such common acoustic properties, as expressed in different languages, are likely to share similar acoustic correlates in the speech signal and can be exploited to build cross-linguistic acoustic models more efficiently than a detailed phonetic or phonemic approach, particularly helpful for languages with little corpus material. Of course languages are different and there are sounds present in one language but not another, even at the broad articulatory feature level. For example, the "trill" (usually an [r]), which is very common in some Indo-European languages, is usually not found in American English. It is therefore of interest to ascertain to what extent such cross-linguistic transfer of AF classification succeeds (or fails).

As a preliminary means of developing AFs for cross-linguistic training in ASR, we have applied the AF-classification system originally designed for American English to spontaneous Dutch material – the VIOS corpus [127][142][17]. Dutch and English are historically related languages (both belong to the West Germanic language family) with approximately 1500 years of time depth separating the two [7]. This close relationship between the two languages makes it particularly suitable for the initial testing of cross-linguistic transferability.

VIOS is a Dutch corpus composed of human-machine dialogues within the context of railroad timetable queries conducted over the telephone [127]. A subset of this corpus (3000 utterances, comprising ca. 60 minutes of material) was used to train an array of networks of multi-layer perceptrons (MLPs), with an additional 6 minutes of data used for cross-validation purposes. Labeling and segmentation at the phonetic-segment level was performed using a special form of automatic alignment that explicitly models pronunciation variation derived from a set of phonological rules [72]. An eighteen-minute component of VIOS, previously hand-labeled at the phonetic-segment level by students in the department of Language and Speech Pathology at the University of Nijmegen, was used as a test set in order to ascertain the accuracy of AF-classification performance. This test material was segmented at the phonetic-segment level using an automatic-alignment procedure that is part of the Phicos recognition system [124] trained on a subset of the VIOS corpus. The phonetic inventory of the VIOS corpus is listed in Table 3.9, along with the AF equivalents for each segment.

AFs for Dutch were systematically derived from phonetic-segment labels using the mapping pattern illustrated in Table 3.9 for the VIOS corpus. The feature dimensions, Front-Back and Rounding applied solely to vocalic segments. The rhoticized segments,

	Anterior		Central		Posterior		Glottal	
Reference	M-I	M-S	M-I	M-S	M-I	M-S	M-I	M-S
Stop								
Anterior	*66*	*80*	17	13	04	06	01	02
Central	07	13	*76*	*77*	06	09	01	02
Posterior	11	12	19	14	*61*	*74*	01	01
Glottal	09	12	16	13	04	07	*29*	*68*
Fricative								
Anterior	*46*	*44*	40	55	01	00	01	00
Central	04	02	*85*	*96*	00	01	03	00
Posterior	01	01	31	43	*62*	*57*	00	00
Glottal	16	15	30	49	06	02	*19*	*34*
Nasal								
Anterior	*64*	*65*	20	31	02	04	-	-
Central	12	09	*69*	*86*	03	05	-	-
Posterior	10	05	32	39	*28*	*56*	-	-
Vowel								
Anterior	*82*	*83*	07	14	02	03	-	-
Central	12	11	*69*	*80*	10	09	-	-
Posterior	17	16	24	35	*48*	*50*	-	-
Vowel Height								
Low	*77*	*83*	13	16	01	01	-	-
Mid	15	18	*58*	*73*	12	09	-	-
High	02	5	11	22	*73*	*73*	-	-

Table 3.8: Confusion matrix associated with the manner-specific (M-S) classification for place-of-articulation feature extraction for each of the four major manner classes, plus the non-place AF dimension "vowel height." Place classification performance for the manner-independent (M-I) system is shown for comparison. All numbers are percent of total frames of the reference features.

CONS	*Manner*	*Place*	*Voice*	*VOW*	*F/B*	*Place*	*Round*
[p]	Stop	Bilabial	-	[i]	Front	High	-
[b]	Stop	Bilabial	+	[u]	Back	High	+
[t]	Stop	Alveolar	-	[y]	Front	High	+
[d]	Stop	Alveolar	+	[I]	Front	High	-
[k]	Stop	Velar	-	[e:]	Front	High	-
[f]	Fricative	Lab-dent	-	[2:]	Front	Mid	+
[v]	Fricative	Lab-dent	+	[o:]	Back	Mid	+
[s]	Fricative	Alveolar	-	[E]	Front	Mid	-
[z]	Fricative	Alveolar	+	[O]	Back	Mid	+
[S]	Fricative	Velar	-	[Y]	Back	Mid	-
[x]	Fricative	Velar	+	[@]	Back	Mid	-
[m]	Nasal	Bilabial	+	[Ei]	Front	Mid	-
[n]	Nasal	Alveolar	+	[a:]	Front	Low	-
[N]	Nasal	Velar	+	[A]	Back	Low	-
				[Au]	Back	Low	+
				[9y]	Front	Low	+
APPR	*Manner*	*Place*	*Voice*	*APPR*	*F/B*	*Place*	*Voice*
[w]	Vocalic	Labial	+	[w]	Back	High	+
[j]	Vocalic	High	+	[j]	Front	High	+
[l]	Vocalic	Alveolar	+	[l]	Central	Mid	+
[L]	Vocalic	Alveolar	+	[L]	Central	Mid	+
[r]	Vocalic	Rhotic	+	[r]	Central	Mid	+
[R]	Vocalic	Rhotic	+	[R]	Central	Mid	+
[h]	Vocalic	Glottal	+	[h]	Central	Mid	+

Table 3.9: Articulatory-acoustic feature specification of phonetic segments developed for the Dutch VIOS corpus. The approximants (*APPR*) are listed twice, on the left for the manner-independent features, and on the right for manner-specific place features. "*F/B*" refers to "Front-back." The phonetic orthography is derived from SAMPA.

	VIOS–VIOS		NTIMIT–VIOS	
FEATURE	+ Silence	- Silence	+ Silence	- Silence
Voicing	89	85	79	86
Manner	85	81	73	74
Place	76	65	52	39
Front-Back	83	78	69	67
Rounding	83	78	70	69

Table 3.10: Comparison of AF-classification performance (percent correct at the frame level) for two different systems – one trained and tested on Dutch (VIOS–VIOS), the other trained on English and tested on Dutch (NTIMIT–VIOS). Two different conditions are shown – classification with silent intervals included (+Silence) and excluded (-Silence) in the test material.

[r] and [R], were assigned a place feature (+rhotic) unique unto themselves in order to accommodate their articulatory variability [85][133]. Each articulatory feature dimension also contained a class for silence. In the manner-specific classification the approximants (i.e., glides, liquids and [h]) were classified as vocalic with respect to articulatory manner rather than as a separate consonantal class.

Classification experiments were performed on the VIOS test material using MLPs trained on the VIOS and NTIMIT corpora, respectively (cf. Table 3.10). Because ca. 40% of the test material was composed of silence, classification results are partitioned into two separate conditions, one in which silence was included in the evaluation of frame accuracy (+silence), the other in which it was excluded (-silence).

Classification performance of articulatory-acoustic features trained and tested on VIOS is more than 80% correct for all dimensions except place of articulation. Performance is lower for all feature dimensions when silence is excluded. Overall, this performance is comparable to that associated with other American English [14] and German [76] material.

Classification performance for the system trained on NTIMIT and tested on VIOS is lower than the system both trained and tested on VIOS (Table 3.10). The decline in performance is generally ca. 8-15% for all feature dimensions, except for place, for which there is a somewhat larger decrement (26%) in classification accuracy. Voicing is the one dimension in which classification is nearly as good for a system trained on English as it is for a system trained on Dutch (particularly when silence is neglected). The manner dimension also transfers reasonably well from training on NTIMIT to VIOS. However, the place-of-articulation dimension does not transfer particularly well between the two languages.

One reason for the poor transfer of place-of-articulation feature classification for a system trained on NTIMIT and tested on VIOS pertains to the amount of material on which to train. Features which transfer best from English to Dutch are those trained on the greatest amount of data in English. This observation suggests that one potentially effective means of improving performance on systems trained and tested on discordant corpora would be to evenly distribute the training materials over the feature classes and

dimensions classified.

3.5 Robustness of AFs

Acoustic interference poses a significant challenge to current-generation ASR systems. ASR systems that work well under pristine acoustic conditions generally perform much more poorly at low signal-to-noise ratios (SNRs). In contrast, human listeners typically experience little (if any) degradation of intelligibility under comparable circumstances, except for SNRs of less than 0 dB [86]. The robust nature of human speech decoding may reflect the brain's application of multiple processing strategies, spanning a broad range of time constants and structural units, providing complementary perspectives on the signal's phonetic and lexical representation [50][51]. Previous research has demonstrated that ASR systems incorporating broad articulatory features are more robust to acoustic interference than systems using only phonetic segments [76].

This section describes experiments designed to reinforce this notion of robustness of AFs by testing the AF classification system on speech in a wide variety of noise backgrounds. We compare the phonetic-segment classification performance of a system with an intermediate stage of AF classification and a system without the AF-classification stage. We also compare two training methods: (1) training only on "clean" speech (i.e., speech that has been recorded under pristine, high-SNR conditions), (2) training on speech embedded in a variety of noise backgrounds over a wide dynamic range of SNRs. The mixed-training scheme has been shown to perform well for both matched (included in the training set) and novel noise conditions (not included in the training set) [15], and is similar to the multi-condition training for noise-robustness adopted by the Aurora Evaluation of Distributed Speech Recognition Systems [4]. In the following chapter we will extend this study to include supra-segmental (syllable-level) information into the classification system.

3.5.1 Corpus Material with Noise

The experiments described in this section were performed on the Numbers95 [12] corpus, the same set of materials as described in Section 3.2.2. Various forms of acoustic interference, derived from the NOISEX corpus [131], were mixed, in additive fashion, with the Numbers95 speech material. The NOISEX material was originally recorded with 16-bit resolution at 19.98 kHz but was down-sampled to 8 kHz for the current study. A subset of the noise backgrounds was mixed with the speech material over a range of SNRs (as indicated in Table 3.11). The signal-to-noise ratio was calculated from the normalized power (computed over the entire length of the utterance) for both the speech signal and the noise background using a procedure described in [74].

3.5.2 Experimental Results

The experiment setup is the same as the MLP-based AF-classification (and subsequent phonetic-segment-classification) system described in Section 3.2. The classifica-

tions are based on entirely automatically derived data, and the "elitist" approach and the manner-specific training are not applied here. Frame-level phonetic-segment classification performance (percent accuracy) are compared across four systems:

- direct phonetic classification (without AF-classification) trained on "clean" speech only (PhnClean);
- phonetic classification via an intermediate stage of AF classification trained on "clean" speech only (AFClean);
- direct phonetic classification trained on speech material embedded in both white and pink noise over a 30-dB range of SNRs, as well as on "clean" speech (PhnMix);
- phonetic classification via an intermediate stage of AF classification trained on speech material embedded in both white and pink noise over a 30-dB range of SNRs, as well as on "clean" speech (AFMix).

The results are summarized in Table 3.11. The mixed-training scheme is very effective against acoustic interference. For both direct-phone-based and AF-based phonetic-segment classifications the mixed-training system dramatically improved classification accuracy not only for noise conditions included in the mixed-training set but also for novel noise conditions absent from the mixed-training set. And in no condition does the mixed-training system perform worse than the corresponding "clean"-trained system. The conditions where the mixed-training system fails to significantly outperform the "clean"-training system are those in which the latter is already performing close to the optimum performance associated with the classification framework used.

AFs exhibit more robustness than phonetic segments (cf. Table 3.11). For the "clean"-only training condition incorporating an intermediate AF-classification stage reduces the phone-classification error by an average of 9.6% (relative, and an average of 4.5% absolute) relative to the direct-phone-based system. This significant error reduction is maintained for the mixed-training condition where the average error reduction of the AF-based system (compared to the direct-phone-based system) is 15.3% (relative, and an average of 5.6% absolute), and this pattern of performance improvement is observed in both seen and unseen noise backgrounds.

Test			Clean Training		Mixed Training	
Noise/Condition		SNR	PhnClean	AFClean	PhnMix	AFMix
1	Clean	-	73.40	78.12	74.85	79.24
2	Pink	0	17.64	21.90	57.69	64.14
3		10	47.06	55.21	69.16	74.59
4		20	68.15	72.92	74.39	78.95
5		30	72.67	77.32	75.32	79.67
6	White	0	15.04	18.94	55.50	61.22
7		10	34.86	45.15	66.94	72.36
8		20	60.86	67.54	73.05	77.81
9		30	71.15	76.05	75.05	79.49
10	Mixture of White	0	16.62	20.35	56.98	63.22
11	and Pink Noise*	10	43.01	52.00	68.45	73.95
12		20	66.57	71.73	74.16	78.71
13		30	72.42	77.05	75.33	79.70
14	Speech Babble*	0	27.48	28.36	39.87	45.21
15		10	55.05	57.78	62.66	68.32
16	Buccaneer (190 knots)*	0	15.32	19.68	52.70	59.24
17	Jet Cockpit (450 knots)*	0	17.16	20.60	51.57	58.48
18	F-16 Jet Cockpit*	0	17.62	23.52	52.63	58.81
19	Destroyer Eng Room*	0	17.42	20.83	46.46	51.09
20	Destroyer Op. Room*	0	29.92	34.28	51.11	58.02
21	Leopard 2 Mil. Vehicle*	0	54.23	56.52	55.33	62.69
22	M109 Tank*	0	41.31	43.01	59.73	66.13
23	Machine Gun*	0	55.24	59.25	57.14	63.44
24		10	62.62	67.90	64.14	70.62
25	Car Factory (Floor)*	0	21.86	25.05	47.28	52.47
26		10	50.17	55.36	65.23	70.58
27	(Production Hall)*	0	35.03	37.64	59.57	65.73
28	Volvo (Interior)*	0	67.19	70.39	69.75	74.83
29		10	70.63	74.62	71.71	76.85
30	Hi-Freq Radio Channel*	0	13.95	17.94	52.59	57.99

Table 3.11: Phonetic-segment classification performance (percent frame accuracy) compared across four systems. Conditions (10-30) marked with an asterisk (*) are those that the mixed-training system has not been trained on. "PhnClean" and "PhnMix" are results of direct phone classification (without intermediate AF-classification); "AFClean" and "AFMix" are results of phone classification via an intermediate stage of AF classification. "Clean Training" refers to training on clean data only; "Mixed Training" refers to training on both clean data and speech embeded in white and pink noises over a 30-dB range (conditions 1-9).

3.6 Summary

This chapter has focused on automatic extraction of articulatory-acoustic features from speech input. Motivated by the AF-related analysis results from the linguistic dissection in the previous chapter, analyses and experiments described in this chapter provided evidence to support incorporating AFs in models of speech processing.

- An TFM/MLP neural-network-based AF-classification system was described in detail with experimental evaluation on the Numbers95 corpus; the AF-classification system was also extended to perform automatic labeling of phonetic segments. Good performance was obtained on AF classification, as well as on phonetic labeling and segmentation.
- AF-classification experiments on the more comprehensive and phonetically balanced NTIMIT corpus were described, and in particular, an "elitist" approach was described to delineate regions of speech with high confidence in AF classification. Moreover, a manner-specific training scheme for enhancing the place-of-articulation classification was also described.
- The cross-linguistic transferability of AF training was assessed quantitatively by testing (American English) NTIMIT-corpus-trained AF-classification networks on a Dutch corpus (VIOS). Experiment results showed that certain AF dimensions (e.g. voicing and manner of articulation) transfer better than others (e.g. the place of articulation).
- Further evidence supporting the use of AFs was provided by the robustness of the AFs as demonstrated in experiments involving speech in noisy background, particularly when the AF-classification system was trained on speech embedded in a variety of noise backgrounds over a wide dynamic range of SNRs.

In the following two chapters further analysis of the AF-deviation patterns from canonical forms will examine the close relationship among AFs, syllable structure and stress accent, which can be exploited to capture the complex phenomenon of pronunciation variation in spontaneous speech in a parsimonious fashion.

Chapter 4

Speech Processing at the Syllable Level

From linguistic dissection of Switchboard-corpus LVCSR systems (cf. Chapter 2), we have observed that syllable-level information plays an important role in word recognition of spontaneous American English discourse. Syllable structure was found to be an important factor in determining word errors, especially word deletions. Tolerance of articulatory-feature errors differs depending on the segments' position within the syllable; and furthermore, much prosodic information that is important for word recognition, such as stress-accent level and speaking rate, is directly tied to the syllabic representation of speech. This suggests that syllable-level information may have a significant impact on speech recognition performance and it may be beneficial to model such syllable-related factors explicitly in ASR systems.

In fact, there has been an increasing interest in the notion that the syllable may be the binding unit of speech around which information at various linguistic tiers is organized [30][51], in contrast to the traditional phonetic-segment perspective of spoken language. This chapter argues for a syllable-centric view of speech perception from three different aspects – (1) the stability of the syllable in the speech signal and the importance of syllable-level information in speech perception, (2) the efficiency of modeling pronunciation variation in a syllable-based representation and the close link between many important kinds of prosodic information (such as stress-accent) and the syllable, as well as (3) the systematic variation of the articulatory-acoustic features (AFs) as a function of syllable position and syllable structure. First, a brief introduction of the syllable is provided.

4.1 What is a Syllable?

In his introductory phonetics textbook [78], Ladefoged concedes that there is no agreed phonetic definition of a syllable. He notes that although nearly everybody can identify syllables, it is difficult to define a syllable with precision. There have been various attempts to define the syllable either in terms of properties of sounds, such as sonority or

prominence, or from the perspective of the speaker, principally the notion that a syllable is a unit of organization for the sounds in an utterance. However, none of these attempts has yielded a completely satisfactory definition (ibid). This thesis does not attempt to provide a solution to this intriguing linguistic problem, but for the convenience of discussion, we simply assume that a syllable is the smallest articulatorily coherent span of speech in the sense that every speech utterance must contain at least one syllable.

Structurally, a syllable potentially consists of three parts[1] – an onset, a nucleus and a coda, where both the onset and coda elements are optional (e.g. the syllable in the word "I" contains only a nucleus). The nucleus is almost always vocalic[2]. The onset in English, as well as the coda, when present, can consist of one or more consonants (a consonant cluster). For example, the word "six" (in its canonical pronunciation) has a single-consonant onset ([s]), a vocalic nucleus ([ih]) and a consonant-cluster coda consisting of two consonants ([k] and [s]). In this case, we may refer to this structure as a CVCC syllable. The phenomenon of ambisyllabicity slightly complicates the picture, with certain segments acting as both the coda of the preceding syllable and the onset of the following syllable, such as the [r] in the word "zero" in some pronunciations. Although we discuss syllable structures by their vocalic and consonantal constituents, this is not an excuse to consider a syllable simply as a sequence of phones, as has been warned by Greenberg with an analogy that the syllable can be likened to a linguistic "wolf" in phonetic clothing [51]. As he remarks, "what distinguishes the syllable from this phonetic exterior is its structural integrity, grounded in both the production and perception of speech and wedded to the higher tiers of linguistic organization."

Because of the possibility of having consonant clusters, syllable structure in English can be very complex. For example, the monosyllabic word "strengths," in its canonical pronunciation, may be of the form CCCVCCC. However, such a complex syllable structure is relatively rare in natural speech. It was found that the "simple" (without consonant cluster) syllable structures – CV, CVC, VC and V – together account for over 75% of the lexicon and over 83% of the syllable tokens in the Switchboard corpus [49][51]. A similar statistic can be observed from the Switchboard material used in the phonetic evaluation (cf. Chapter 2), as shown in Figure 2.5[3]. Interestingly, this preference for simple syllabic structures put spontaneous spoken English on par with languages that are traditionally known to have more homogeneous syllable structures, such as Japanese [1] and Mandarin Chinese. For example, standard Mandarin Chinese does not allow any consonant in the coda position except nasals, and the majority of Mandarin Chinese syllables contain no more than one consonant in the onset position [81].

[1] In tonal languages such as Chinese, however, the tone is also an important part of a syllable.

[2] When a nucleus is not vocalic, it is often a syllabic consonant such as in the second syllable of the word "button" in certain pronunciations (e.g. [b ah q *en*] as often heard in New York City dialect).

[3] Note that many polysyllabic words, such as those of the form CVCVC, also contain syllables with no consonant cluster.

4.2 The Stability and Importance of the Syllable in Speech Perception

Syllable-based modeling of speech has been widely used in ASR systems for languages that are considered more explicitly syllabic (e.g Mandarin Chinese [81] and Japanese [97]). In recent years, there have also been several studies of incorporating syllable-level information in English-language ASR systems, which to date, are still dominated by phone-based (including context-dependent phones such as tri- and quinta-phone) approaches [107]. For example, Wu describes building ASR systems incorporating information from syllable-length time scales in her dissertation [144]; Ganapathiraju et al. have built a syllable-based and syllable/phone hybrid ASR system [40] for the Switchboard corpus [42]; and Jones et al. report a syllable-based ASR system for a British English corpus [67].

4.2.1 Stability of Syllables in Speech Corpora

A commonly cited reason for adopting a syllable-based approach is the greater stability of syllables relative to phones [51]. A study by Ganapathiraju et al. [40] showed that the syllable deletion rate (compared to the canonical pronunciation) on the manually transcribed portion of the Switchboard corpus [42][49] was below 1%[4], while the phone-deletion rate for the same material was ca. 12%. It should be noted that the transcribers were carefully trained not to insert into transcriptions anything that was not truly present in the speech signal [49]. On the Numbers95 corpus [12], which presumably is more canonically pronounced than the Switchboard material, on average, due to the nature of the material, the syllable-deletion rate was below 0.6% and the phone-deletion rate was ca. 4% (computed on approximately 2.5 hours material). Both these statistics suggest that there is approximately an order of magnitude difference between syllable- and phone-deletion rates. This stability of the syllable is also manifested in automatic systems for AF extraction (cf. Chapter 3), particularly for manner-of-articulation classification, where vocalic segments (usually taking the position of the non-optional syllable nuclei) enjoy the lowest error rate among various phonetic segments.

4.2.2 Acoustic-based Syllable Detection and Segmentation

The relative stability of the syllable is also supported by evidence that it is possible to achieve reasonable performance of syllable detection and segmentation from the acoustic signal alone. For example, Shire [118] developed a perceptually motivated method of estimating syllable onsets and achieved a 94% onset detection rate (with a five-frame tolerance) with a false positive rate of 15% (i.e., onset detected where there was none). The onset detection method was adopted by Wu in her work on integrating syllabic onsets into ASR system and achieved a ca. 10% (relative) word-error reduction on the Numbers95 corpus[145][144].

[4]The syllable-deletion rate for the subset of the Switchboard corpus that was used in the Year-2000 phonetic evaluation (cf. Chapter 2) was 1.05%.

In our previous work [117], Shastri, Greenberg and I have developed a temporal-flow-model (TFM) [137] (also cf. Section 3.2.1) neural-network-based syllable detection and segmentation system using the perceptually inspired Modulation-filtered Spectrogram (ModSpec) pre-processing [60][75][74] (cf. Section 4.2.3 for a discussion of the relationship between modulation spectrum and syllable duration, which is a significant aspect that the ModSpec is modeling). Two types of TFM networks were used: the global TFM network was similar to that described in Section 3.2.1 for AF classification; the tonotopically organized TFM network provided a spectrally differentiated receptive field for different hidden nodes. The training targets for the networks were Gaussian functions approximating the syllable-peak trajectory over time. The trained networks were able to produce outputs incorporating syllabic information. Syllable-onset detection (on the Numbers95 corpus) derived from network outputs using a two-level thresholding algorithm achieved an 88.5% detection rate (with a five-frame tolerance), with only 4.4% false positive rate [5].

In his dissertation work on segmentation and recognition of continuous speech, Prasad [106] developed a syllable-segmentation algorithm based on a minimum-phase, group-delay function of short-term energy. The algorithm was demonstrated to perform well on corpora of both American English and on a number of Indian languages (Tamil and Telugu). The evidence above suggests that there are significant acoustic correlates of the syllable in the speech signal, which is the likely reason that acoustic-based measures of speaking rate (e.g., MRATE) [91][92] exhibit some correlation with the linguistic measure (syllable per second).

4.2.3 Significance of Syllable Duration

Research focusing on acoustic correlates of speech intelligibility has found that significant attenuation of the key components of the modulation spectrum results in serious degradation of speech intelligibility. Drullman et al. have found by temporally filtering the spectral envelope of speech [34][33]. that intelligibility of spoken Dutch material does not require modulation energy above 16 Hz. In their experiments using Japanese syllables with filtered time trajectories of the spectral envelope, Arai et al. showed that the modulation frequency region between 1 and 24 Hz is most important for speech intelligibility [3]. In other studies, Arai and Greenberg have found that speech intelligibility (of American English TIMIT sentences) is correlated with the magnitude of the low-frequency (3-6 Hz) modulation spectrum from psychoacoustic experiments using a technique of cross-channel spectral de-synchronization [2][54].

This converging evidence raises interesting questions as to whether there is any linguistic significance to the particular distribution of modulation spectral magnitude that is important for speech intelligibility. A potentially revealing comparison between the modulation spectrum and the frequency histogram of syllabic durations for spontaneous English discourse was presented in an analysis by Greenberg et al.[49], on manually transcribed Switchboard corpus material [42][49], and is reprinted in Figure 4.1. The modes of both the modulation frequency and (the reciprocal of) the syllable duration are at ca. 4-5 Hz and both have significant magnitude between 2 and 8 Hz. Except for the longer tail exhibited by

[5] The system was optimized for minimizing the combined false positive and false negative rate.

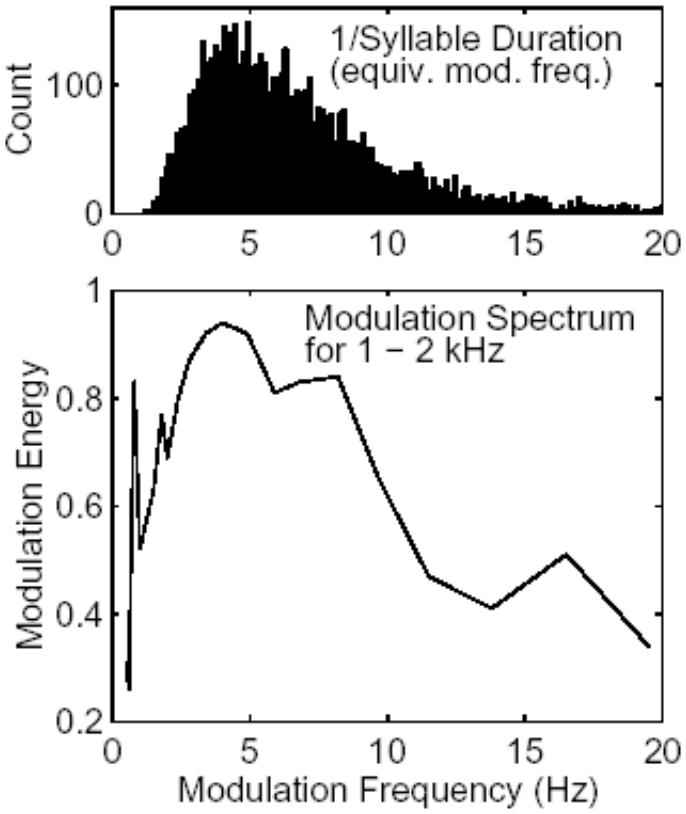

Figure 4.1: Modulation spectrum and frequency histogram of syllabic durations for spontaneous English discourse (adapted from [49]). Top panel: histogram pertaining to 2925 syllabic segments from the Switchboard corpus. Bottom panel: modulation spectrum for two minutes of connected, spoken discourse from a single speaker.

the modulation spectrum, the two distributions are remarkably similar, suggesting that it may be encoding information at the syllable level, and thus also suggesting the importance of syllable-length units in speech intelligibility. A similar comparison was performed on spontaneous Japanese material and the conclusion is essentially the same as for the English material [1].

4.2.4 Syllables and Words

An English word can potentially contain many syllables, partially because of the various derivational and inflectional affixes that can be attached (for example, consider the word *uninformatively*) In the Switchboard lexicon only ca. 22% of the lexicon is monosyllabic, almost 40% of the lexicon contains two syllables and ca. 24% contains three syllables [49][51]. An implication of this potential complexity of the English lexicon is a

seemingly non-transparent relationship between syllables and words. However, an analysis of corpus tokens reveals a rather different trend; about 80% of the word tokens in the Switchboard corpus are monosyllabic in form, while most of the remaining word tokens have but two syllables [51]. The distribution of syllable structure in the phonetic evaluation material described in Chapter 2 shows similar patterns (cf. Figure 2.5). This preference of brevity in syllable structure suggests that syllable and word have a closer relationship in spontaneous spoken English than may at first seem and that the correct identification of syllables would take a system a long way toward good performance in word recognition.

4.3 Pronunciation Variation, Prosody and the Syllable

Through linguistic dissection of the Switchboard LVCSR systems, it was shown that current generation ASR systems generally do not adequately model pronunciation variation found in the spontaneous speech corpus (cf. Section 2.2.6). The number of pronunciation variants per word in the Switchboard corpus according to the manual transcription [49] can often be an order of magnitude greater than the number of different pronunciations contained in the lexical models of the ASR systems. As we have shown in Figure 2.12, systems incorporating more sophisticated pronunciation models tend to perform better with respect to word recognition. One implication of this finding is to increase the number of pronunciation variants per word in an ASR system's lexicon in order to improve the word recognition performance. However, it has been noted that simply adding more variants to the lexical models could increase confusibility among words and therefore result in degraded recognition performance [114][123]. The source of this problem, as concluded by McAllaster et al. using fabricated data from the Switchboard corpus, is likely due to the mismatch between the pronunciation models and data encountered in the recognition task [87]. Therefore, to achieve a substantial reduction in word error, the acoustic models must be very accurate and well-tuned to the representations of the recognition lexicon. To accomplish this, it is essential to characterize pronunciation variation accurately and efficiently.

However, directly examining pronunciation variation at the phonetic-segment level often leaves a rather arbitrary and complex picture. Consider the example illustrated in Table 4.1 with various phonetic realizations of the word "that" from the Year-2001 Switchboard material (cf. Section 2.1.1). Among the 226 instances of this word there are a total of 63 different pronunciations[6]. Taking [dh ae t] as the canonical pronunciation, there are a total of 176 phonetic-segment substitutions (26% of the 678 canonical phonetic segments), 101 deletions (15%) and 3 insertions (1%). These statistics provide little insight in and of themselves into the pattern of pronunciation variations for the word "that" other than the overall proportion of substitutions, deletions and insertions of phonetic segments.

Now, let us partition the data according the position of each segment within the syllable (cf. Table 4.1 and Table 4.2). Some potentially interesting patterns of pronunciation variation emerge:

[6] According to [49], there are a total of 117 different pronunciations of the word "that" among 328 instances in the original STP material (ca. four hours). Most of the pronunciations are transcribed accurately although a very small portion of the pronunciation variations may be due to temporal misalignment between word and phone segments.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
dh ae t	53	-	-	-	th aw t	1	S	S	-
dh ae	31	-	-	D	s ax t	1	S	S	-
dh ae dx	13	-	-	S	n eh	1	S	S	D
dh eh t	10	-	S	-	n ax	1	S	S	D
dh ax	9	-	S	D	n aw	1	S	S	D
dh aw	9	-	S	D	n ae dx	1	S	-	S
n ae t	8	S	-	-	n aa	1	S	S	D
n ae	7	S	-	D	l ih	1	S	S	D
dh ax t	7	-	S	-	l ae dx	1	S	-	S
dh eh	5	-	S	D	k ih dh	1	S	S	S
dh ah t	5	-	S	-	iy	1	D	S	D
dh ih t	4	-	S	-	hh ih t	1	S	S	-
th ae t	3	S	-	-	eh t	1	D	S	-
d ae t	3	S	-	-	eh dx	1	D	S	S
dh ax dx	3	-	S	S	d ax p	1	S	S	S
ae	3	D	-	D	dh iy	1	-	S	D
t aw	2	S	S	D	dh ih d	1	-	S	S
n ah t	2	S	S	-	dh eh dx	1	-	S	S
d ae	2	S	-	D	dh ah d	1	-	S	S
dh ih	2	-	S	D	dh ae d	1	-	-	S
dh ah dx	2	-	S	S	dh ae ch	1	-	-	S
ah dx	2	D	S	S	dh ae b	1	-	-	S
z d ae	1	S,I	-	D	dh aa t	1	-	S	-
z ah p	1	S	S	S	ax dx	1	D	S	S
t dh ae	1	I,-	-	D	ax dh	1	D	S	S
t b ae t	1	S,I	-	-	ax	1	D	S	D
t ax	1	S	S	D	aw	1	D	S	D
t ae	1	S	-	D	ae w	1	D	-	S
th eh t	1	S	S	-	ae t	1	D	-	-
th eh	1	S	S	D	nx ax	1	S	S	D
th ax t	1	S	S	-	nx ae	1	S	-	D
th ax	1	S	S	D					

Table 4.1: Pronunciation variants of the word "that" found in the Switchboard corpus material used in the Year-2001 diagnostic phonetic evaluation (cf. Chapter 2). For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([dh ae t]) with respect to syllable onset, nucleus and coda. "S" is "substitution," "D" is "deletion," "I" is "insertion" and "-" is "no deviation."

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	70.7	21.8	6.1	1.3	229
Nucleus	59.7	40.3	0	0	226
Coda	46.0	15.5	38.5	0	226
Overall	58.9	25.8	14.8	0.4	681

Table 4.2: Summary of the phonetic deviation (from canonical), in percentage of total segments (last column) at each syllable position (and overall), for the word "that" (cf. Table 4.1) from the Year-2001 diagnostic phonetic evaluation material.

- the onsets exhibit the smallest deviation from the canonical, with far more substitutions than deletions and with only three instances of insertions;
- the nuclei have many more deviations from canonical than the onsets, and all of them are substitutions;
- the codas exhibit the largest number of deviations from the canonical; and there are far more deletions than substitutions.

Although computed for a single monosyllabic word, these patterns roughly agree with the statistics computed on all syllables from the Switchboard corpus [51], where 84.7% onsets, 65.3% nuclei and 63.4% codas are canonically pronounced, and are also in agreement with the principles of pronunciation variation for spontaneous spoken English described by Greenberg (*ibid*, section 7).

However, of course, the patterns summarized above are not sufficient to explain the complex phenomenon of pronunciation variation in natural speech. As enumerated by Wester [141], the sources of pronunciation variation can be attributed to either inter-speaker or intra-speaker variability. The inter-speaker variability may be due to factors such as vocal tract differences, age, gender, dialect, etc.; the intra-speaker variability may depend on speaking style and speaking rate, stress accent and intonation patterns, emotional state of the speaker, environmental conditions, idiolectal variation, etc. It is observed that many of the factors affecting pronunciation come from linguistic levels higher than the phonetic segment. This suggests that it would be very helpful to model pronunciation variation at several different linguistic levels, especially the suprasegmental tiers. In the following chapter, it will be shown that incorporating stress-accent patterns yield further insight into the phenomenon of pronunciation variation in spontaneous speech.

4.4 Articulatory-acoustic Features and the Syllable

As described in the previous chapter, articulatory-acoustic features (AFs) offer a number of potential advantages for models of spontaneous speech. One of the most important is the systematic relationship between AFs and syllable position. This echoes

what was observed from the linguistic dissection of the Switchboard LVCSR systems (cf. Chapter 2), which showed that ASR systems' tolerance of AF errors varies as a function of syllable position, syllable structure and the particular AF dimension of interest.

Different AFs are realized differently across the syllable. Manner of articulation features are mostly coterminous with the traditional phonetic segment, such that it is very rare to have two consecutive segments associated with the same manner class. For example, on the Switchboard corpus, at least 93% of the segments have a manner of articulation different from that of the previous segment when both intra- and inter-syllabic segmental boundaries are considered; the rate of manner-of-articulation change is 99.6% when only intra-syllabic segmental boundaries are considered. This suggests that segmentation of the acoustic signal based solely on manner-of-articulation would be very close to phonetic segmentation, particularly within the syllable. In contrast, other AF dimensions, such as voicing and place-of-articulation, change much more slowly, evolving at a rate comparable to that of the syllable. Voicing changes ca. 31% of time across segmental boundaries when both intra- and inter-syllabic segmental boundaries are considered; the rate of voicing change is ca. 30% when only intra-syllabic segmental boundaries are considered. Unlike manner-of-articulation, there is relatively little difference between the rates of voicing change across intra-syllabic and inter-syllabic segmental boundaries. When partitioned into coarse anterior, central and posterior places, place-of-articulation evolves slightly faster than voicing. The rate of place-of-articulation change is ca. 55% for both intra- and inter-syllabic segmental boundaries, still much slower than manner-of-articulation changes.

As described previously, syllable nuclei are almost always associated with a vocalic manner of articulation, while onsets and codas are generally consonantal. Place of articulation also exhibits different distributional patterns across the syllable. For example, Figure 4.2 illustrates the distribution of place features (partitioned into anterior, central and posterior, plus the "place chameleons" such as [l] and [r] that adapt their place according to the vowels in context) with respect to position within the syllable for both canonical and realized (transcribed) segments, from the Switchboard corpus [42][49][7]. In both the canonical and transcribed segments, the syllable onset segments have a relatively even place distribution with a slight preference of anterior place over central and posterior places. However, the coda segments behave very differently, with a decided preference of central place over anterior and posterior places. The general preference of central place (coronal) has been noted previously by several researchers for many different languages [103][70][69]. Interestingly, as shown in Figure 4.2, a significant portion of the central coda segments, (as well as the place chameleons) are deleted in the transcribed data relative to the canonical, while the numbers of anterior and posterior segments are relatively stable. A hypothesis for explaining the tendency of central-place coda deletion was offered by Greenberg et al. [56] linking the identity of the preceding vocalic nucleus and the coda consonant by a possible sharing of acoustic cues in the mid-frequency (ca. 1500-2200 Hz) region of the spectrum.

The intimate relationship between AFs and syllable structure is further evidenced by a large gain in AF classification performance when syllable position information was incorporated, as described in the following experiments. Recall the experiments performed on the Number95 corpus [12] to evaluate the robustness of AF classification for speech in

[7] The data were originally computed by Hannah Carvey at ICSI.

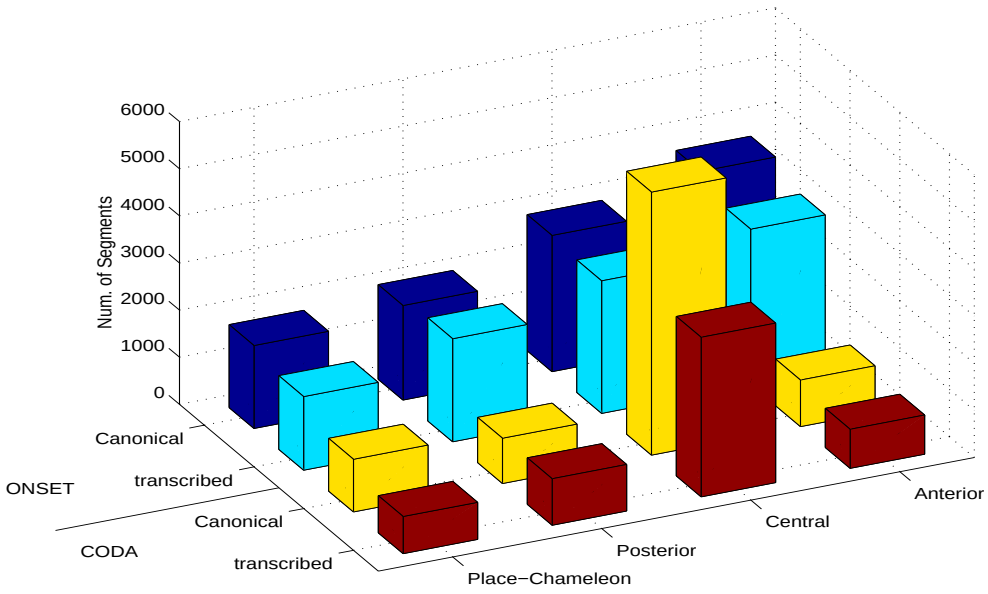

Figure 4.2: Distribution of place features (partitioned into anterior, central and posterior, plus the "place chameleons" such as [l] and [r] that adapt their place according to the vowels in context) as a function of the position within the syllable for both the canonical and realized (transcribed) segments, from the Switchboard corpus [42][49]. Note the large proportion of deleted central coda segments.

noisy background [15] in the previous chapter (cf. Section 3.5). From those experiments (cf. Table 3.11), it was found that AFs are relatively more robust with respect to additive noise compared to phonetic segment and that the mixed-training regime under a variety of noise backgrounds and over a large dynamic range of SNRs significantly improved AF and phone classification performance on both "clean" and noisy speech (in both unseen and previously seen noises). In the current experiments, another input feature pertaining to the position of a frame within the syllable was also fed to the neural-network-based AF classifiers, along with the log-compressed critical-band energy front-end features [16].

This new feature of syllable position can be derived from two separate sources. In the first experiment it was derived from the manual phonetic transcripts that are automatically syllabified using Dan Ellis' (then at ICSI) adaptation of Bill Fisher's (NIST) syllabification program TSYLB2 [31]. For each frame a number linearly spaced between 0 and 1 indicates relative position of the current frame within the syllable, with 0 being the initial frame and $1 - 1/N$ being the final frame where N is the total number of frames within the syllable. These fabricated data establish an upper-bound on the accuracy of the syllable position estimate and can be used to infer the potential contribution of syllable position in AF classification. In the second experiment the syllable position feature was estimated by an MLP neural network from the acoustic signal with Modulation-filtered Spectrogram [74] pre-processing (also cf. Section 4.2.2). The training targets for the MLPs were the transcript-derived syllable position features from the first experiment. In both experiments, the MLP networks for AF classifications were trained and tested with the new syllable position feature included in the input.

The results of the two experiments are shown in Tables 4.3-4.5 for each of the five AF dimensions, as well as for phonetic-segment classification using the results of the AF classification. Both experiments used the mixed-training scheme, same as that described in Section 3.5. "NoSyl" refers to the baseline condition where no syllable position feature is used (i.e. same as the AFMix condition in Section 3.5); "HandSyl" refers to the results of the first experiment using transcript-derived syllable position feature; and "SylMSG" refers to the results of the second experiment using the MLP/ModSpec estimate of syllable position features. The test condition numbers refer to the noise conditions listed in Table 3.11.

For the first experiment ("HandSyl" in Tables 4.3-4.5) a large improvement in classification performance from the baseline condition ("NoSyl") is evident for every noise conditions (either seen or unseen during training) across the various AF dimensions, as well as for the phonetic-segment (via AF classification). The error rate reduction is often between 20 and 30% over the baseline, which as described in the previous chapter (cf. Section 3.5), already enjoys a substantial improvement from the "clean"-training scheme. For the second experiment ("SylMSG" in Tables 4.3-4.5) where the syllable position feature is automatically computed from the acoustic signal, there are still significant gains in performance for most of the noise conditions across the various AF dimensions. However, the phonetic-segment classification (using automatic AF classification results) yields mixed results; "SylMSG" performance is better than that of the "NoSyl" for some noise conditions but not for others. It is not entirely clear what the cause of this discrepancy is. A possible explanation is that the evaluation of classification accuracy at the AF level considers only the feature with the maximum output as the winner along each AF dimension; however, the inputs to the

Test Noise	Manner			Place		
	NoSyl	SylMSG	HandSyl	NoSyl	SylMSG	HandSyl
1	81.83	82.44	87.95	76.42	76.50	82.92
2	66.21	68.07	74.47	61.89	62.36	70.05
3	76.28	77.60	83.38	71.70	72.00	78.85
4	81.16	81.75	87.64	76.27	75.93	82.73
5	82.26	82.72	88.40	76.95	76.72	83.43
6	65.24	68.38	73.59	60.31	61.61	68.36
7	74.02	76.29	81.61	69.81	70.79	77.20
8	79.91	80.72	86.36	75.02	75.09	81.84
9	81.99	82.33	88.19	76.75	76.58	83.27
mean (1-9)	*76.54*	*77.81*	*83.51*	*71.68*	*71.95*	*78.74*
10*	65.62	67.78	73.86	61.15	61.72	69.38
11*	75.54	77.22	82.76	71.16	71.66	78.41
12*	80.95	81.56	87.31	76.04	75.73	82.57
13*	82.26	82.68	88.42	76.96	76.73	83.46
14*	51.65	56.93	61.56	46.64	52.29	57.20
15*	70.37	73.37	78.76	64.89	68.47	74.03
16*	63.30	66.46	72.02	57.63	60.11	67.01
17*	61.15	64.61	70.61	57.03	58.93	66.55
18*	62.67	65.34	71.51	57.22	59.24	66.34
19*	59.14	64.42	68.52	52.80	57.78	62.37
20*	59.06	64.87	70.03	53.87	60.30	65.98
21*	64.28	65.06	76.24	58.58	63.24	71.58
22*	67.55	69.11	77.16	62.58	65.19	72.83
23*	66.91	68.74	74.15	61.69	64.04	69.18
24*	72.60	74.20	80.33	66.68	68.81	74.58
25*	57.91	62.21	67.62	54.02	57.27	63.32
26*	72.99	75.82	81.21	68.20	70.62	76.74
27*	68.27	69.16	76.84	63.21	64.75	72.31
28*	78.36	78.69	85.04	71.26	72.93	79.34
29*	79.73	81.18	86.52	73.21	74.69	80.87
30*	64.00	66.69	72.74	58.05	59.52	66.58
mean (10-30)	*67.82*	*70.29*	*76.34*	*62.52*	*64.95*	*71.46*

Table 4.3: Comparison of frame-level accuracy (percent) of manner and place classification for mixed-training system without syllable position (NoSyl), with ModSpec-based automatic syllable position estimates (SylMSG), and with transcription-derived syllable position feature (HandSyl). The test noise index refers to the conditions listed in Table 3.11. Test noise conditions 1-9 are included in the training data and conditions 10-30 are included (marked with an asterisk).

Test Noise	Front-back			Rounding		
	NoSyl	SylMSG	HandSyl	NoSyl	SylMSG	HandSyl
1	82.81	82.87	88.37	83.31	83.80	89.35
2	68.61	72.61	78.57	69.38	73.53	79.87
3	78.38	79.87	85.51	79.12	80.89	86.61
4	82.35	82.73	88.43	83.21	83.72	89.58
5	83.29	83.30	88.83	83.92	84.27	89.92
6	66.58	70.59	76.49	67.48	71.77	77.83
7	76.16	78.59	84.17	76.88	79.55	85.18
8	81.24	82.16	87.66	82.08	83.01	88.81
9	82.98	83.35	88.79	83.89	84.13	89.87
mean (1-9)	*78.04*	*79.56*	*85.20*	*78.81*	*80.52*	*86.34*
10*	67.80	71.72	77.95	68.71	72.69	79.26
11*	77.76	79.62	85.24	78.44	80.57	86.15
12*	82.16	82.64	88.31	82.97	83.56	89.36
13*	83.24	83.32	88.87	84.01	84.29	89.91
14*	57.33	64.08	66.68	58.40	64.72	68.51
15*	71.85	76.13	80.21	72.95	76.54	82.42
16*	64.62	69.75	75.48	65.98	70.90	76.99
17*	65.00	69.69	75.76	65.61	70.48	76.87
18*	65.02	69.00	75.17	66.28	70.03	76.87
19*	59.62	66.34	70.25	61.99	67.53	72.00
20*	63.95	71.11	75.07	64.86	71.38	76.96
21*	66.34	73.29	77.41	68.25	72.61	80.00
22*	69.39	74.39	79.98	71.04	74.29	81.91
23*	70.71	73.53	77.42	71.69	74.53	78.05
24*	75.03	76.92	81.53	75.51	77.96	81.99
25*	62.84	67.45	72.64	63.64	68.36	73.86
26*	75.23	78.12	83.46	75.97	79.03	84.88
27*	69.92	73.64	79.80	71.79	74.17	81.49
28*	78.14	80.49	85.01	78.93	81.17	86.28
29*	80.48	81.63	86.89	81.29	82.67	87.59
30*	64.01	67.41	74.49	65.44	68.90	76.30
mean (10-30)	*70.02*	*73.82*	*78.93*	*71.13*	*74.59*	*80.36*

Table 4.4: Comparison of frame-level accuracy (percent) of front-back and lip-rounding classification for mixed-training system without syllable position (NoSyl), with ModSpec-based automatic syllable position estimates (SylMSG), and with transcription-derived syllable position feature (HandSyl). The test noise index refers to the conditions listed in Table 3.11. Test noise conditions 1-9 are included in the training data and conditions 10-30 are included (marked with an asterisk).

Test Noise	Voicing			Phone (via AF)		
	NoSyl	SylMSG	HandSyl	NoSyl	SylMSG	HandSyl
1	89.48	89.89	93.98	79.24	78.51	85.60
2	77.07	80.15	84.74	64.14	62.08	73.47
3	85.03	86.77	90.81	74.59	73.64	82.66
4	88.79	89.53	93.67	78.95	77.76	85.71
5	89.72	90.11	94.24	79.67	78.69	86.09
6	77.47	81.92	85.26	61.22	60.32	70.51
7	84.19	86.66	90.30	72.36	72.19	80.68
8	88.14	89.05	93.12	77.81	77.11	84.82
9	89.47	89.87	93.98	79.49	78.65	85.97
mean (1-9)	*85.48*	*87.11*	*91.12*	*74.16*	*73.22*	*81.72*
10*	76.74	80.05	84.59	63.22	61.18	72.49
11*	84.67	86.67	90.56	73.95	73.22	82.17
12*	88.64	89.41	93.55	78.71	77.62	85.55
13*	89.71	90.05	94.21	79.70	78.72	86.12
14*	63.47	69.71	72.49	45.21	48.53	55.69
15*	77.45	82.70	86.27	68.32	69.31	76.69
16*	74.74	79.22	83.37	59.24	58.91	68.80
17*	72.14	76.17	81.44	58.48	56.71	68.45
18*	74.38	78.36	82.87	58.81	58.48	67.81
19*	72.88	78.77	80.95	51.09	54.21	61.27
20*	69.43	75.82	79.49	58.02	59.27	67.85
21*	69.68	76.25	81.66	62.69	60.51	73.51
22*	76.48	79.58	84.75	66.13	65.19	75.54
23*	75.23	78.88	81.27	63.44	62.88	71.36
24*	79.99	82.32	86.21	70.62	69.22	78.03
25*	69.40	74.64	78.24	52.47	54.43	63.46
26*	81.49	85.07	88.78	70.58	71.96	79.72
27*	77.49	79.95	85.05	65.73	64.91	74.97
28*	84.26	87.19	91.80	74.83	74.22	82.52
29*	87.70	89.02	93.56	76.85	76.29	84.10
30*	77.12	81.10	85.02	57.99	57.46	68.12
mean (10-30)	*77.29*	*81.00*	*85.05*	*64.58*	*64.44*	*73.53*

Table 4.5: Comparison of frame-level accuracy (percent) of voicing and phonetic-segment (using the results of AF classification) classification for mixed-training system without syllable position (NoSyl), with ModSpec-based automatic syllable position estimates (SylMSG), and with transcription-derived syllable position feature (HandSyl). The test noise index refers to the conditions listed in Table 3.11. Test noise conditions 1-9 are included in the training data and conditions 10-30 are included (marked with an asterisk).

subsequent phone classification networks are in the form of raw MLP outputs from the AF classification networks, resulting in a potential mismatch. Overall, these results are very encouraging and suggest that there is a significant relation between AFs and syllable position which could be exploited to improve the performance of automatic systems for speech processing.

4.5 Summary

This chapter has focused on the central role played by the syllable in spoken language and provided evidence in support of the syllable being the binding unit of speech, around which information at various linguistic tiers is organized.

- The stability and importance of the syllable in speech was emphasized from several perspectives: deletion statistics from manually annotated spontaneous speech corpora; an acoustic-based syllable detection and segmentation experiment using TFM neural networks and Modulation-filtered Spectrogram features; the significance of syllable duration in speech perception; the close relationship between words and syllables in spoken English.

- Through a concrete example of word pronunciation instances extracted from spontaneous speech material syllable information was shown to be very helpful in describing the observed pronunciation variation patterns. In particular, it was shown that nuclei and codas of syllables are more likely to deviate from canonical forms than onsets. This example will be described in further detail in the next chapter with respect to stress-accent information.

- The intimate relationship between articulatory-acoustic features and the syllable was reinforced by a significant gain in AF and phonetic classification accuracy when syllable-position information (either derived from manual syllable segmentation or automatically estimated from the acoustic signal) was incorporated for speech in both clean and noisy backgrounds.

In the following chapter it will be shown through both statistics on spontaneous speech and concrete examples extended from the current chapter that stress accent, in conjunction with syllable position, helps characterize pronunciation variation patterns of spontaneous speech. It will further be shown that a parsimonious description of pronunciation variation phenomena may be obtained by considering the realization of articulatory-acoustic features, within the context of syllable position and stress accent.

Chapter 5

Stress Accent in Spontaneous American English

Imagine that you hear a piece of speech uttered according to the canonical dictionary pronunciation for each phoneme associated with a word but without any apparent prosodic prominence. Chances are that you would have difficulty following what is said even though it would have been perfectly intelligible if it were spoken naturally. In fact, many early speech synthesis systems without prosodic modeling could produce such speech; and a few people with a rare skill of uttering such speech could have been employed by advertising agencies to provide the "tiny prints" following a TV or radio commercial that is not intended to be heard distinctively. The problem is not that you have trouble making out the sound of each phoneme but rather it is the lack of informational cues to parse the utterance into a succession of shorter phrases and to specify the linguistic relationships among them.

A major prosodic cue of spoken English is stress accent. The next section defines what is meant by stress accent and provides a brief survey of previous work regarding the acoustic correlates of stress-accent patterns, as well as our interpretation of the factors affecting stress accent based on spontaneous speech material. The following section discusses the pattern of interactions among stress accent, syllable position and articulatory-acoustic features (AFs) in relation to pronunciation variation. The last section describes the development of an automatic stress-accent labeling system for spontaneous spoken English and experimental evidence to support the perceptual basis of stress accent proposed in the first section.

5.1 Stress Accent in Spontaneous American English

English is a stress-accent language [5] in that it uses many different acoustic parameters to engender prosodic prominence essential for lexical, syntactic and semantic disambiguation [83]. This is in contrast to non-stress-accent (e.g. pitch-accent) languages, such as Japanese, that rely far more on pitch-related cue for providing accentual information [5].

It is important to note that the term "stress accent" is different from lexical stress (or word stress) found in the pronunciation component of a dictionary, but rather reflects the perceived prominence (stress-accent level) of the phonetic realization of a syllable in a speech utterance. Although sometimes the perceived stress accent may coincide with lexical stress, there are many other potential factors affecting perceived stress accent, such as emphasis and turn-holding (in conversation), etc.

5.1.1 The Perceptual Basis of Stress Accent

The perceived level of stress accent is a subjective matter in that it is quite possible that different listeners would give different stress-accent judgments for the same phonetic realization of syllables in an utterance. However, this is not to say that there is no consistent set of acoustic cues associated with stress-accent level. In fact, listeners agree most of the time on stress-accent markings for spontaneous (conversational) speech [121][57][63]. Traditionally, it was generally accepted that variation in fundamental frequency (f_0), which closely relates to pitch change[1], is the primary acoustic cue for spoken English [20][41]. However, studies based on experimental data have called the traditional view into question. For example, Beckman showed through psychoacoustic study of *laboratory* speech that duration and amplitude play a more important role in the perception of stress accent than previously thought [5]. More recent studies using statistical methods have shown that the acoustic basis of stress accent in spontaneous American English (using OGI Stories corpus [11]) is largely derived from amplitude and duration (as well as their product), with f_0 variation playing a largely subsidiary role [121][122], and a similar pattern has also been found in spontaneous spoken Dutch discourse [130]. Such findings are also supported by the experimental results of automatic stress-accent labeling on the Switchboard corpus, as will be described in the final section of this chapter. But a more interesting finding coming out of the statistical analysis of a subset of the Switchboard corpus with manual stress-accent labels, as well as from the automatic stress-accent labeling experiments, is an intimate relationship between vocalic identity and stress-accent level, at least in spontaneous American English dialogs [63][58].

5.1.2 Vocalic Identity and Stress Accent

Traditional linguistic theories hold that stress accent is a linguistic parameter functionally orthogonal to the phonetic tier [20], in that the realization of phonetic constituents is largely independent of the stress-accent level associated with the syllable. Thus, a non-orthogonal relationship between vocalic identity and stress-accent level not only challenges this traditional belief but also opens the door to more sophisticated interactions between stress accent and phonetic realization, as will be described in the next section.

Statistical analysis of the relationship between vocalic identity and stress accent [63][58] was performed on a 45-minute subset of the Switchboard corpus [42][49] consisting of 9,992 words, 13,446 syllables and 33,370 phonetic segments, comprising 674 utter-

[1]It is important not to confuse the concepts between the physical measure of f_0 and the subjective sensation of pitch [78].

ances spoken by 581 different speakers. This material was manually labeled by linguistically trained individuals at the word-, syllable- and phonetic-segment levels. Two individuals (distinct from those involved with the phonetic labeling) marked the material with respect to stress accent [63][58]. Three levels of stress accent were distinguished - (1) fully accented [level 1], (2) completely unaccented [level 0] and (3) an intermediate level [0.5] of accent. The transcribers were instructed to label each syllabic nucleus on the basis of its perceptually based stress accent rather than using knowledge of a word's canonical stress-accent pattern derived from a dictionary. The transcribers met on a regular basis with the project supervisor to insure that the appropriate criteria were used for labeling.

All of the material was labeled by both transcribers and the stress-accent markings averaged. In the vast majority of instances the transcribers agreed precisely as to the stress-accent level associated with each nucleus – inter-labeler agreement was 85% for unaccented nuclei, 78% for fully accented nuclei (and 95% for any level of accent, where both transcribers ascribed some measure of stress accent to the nucleus) [63][58]. In those instances where the transcribers were not in complete accord, the difference in their labeling was usually a half- (rather than a whole-) level step of accent. Moreover, the disagreement was typically associated with circumstances where there was some genuine ambiguity in stress-accent level (as ascertained by an independent, third observer). In other words, it was rare for the transcribers to disagree as to the presence (or absence) of accent.

The systematic relationship between stress accent and vocalic identity is most evident in the distinct vowel spaces associated with the vocalic nuclei of fully accented and unaccented syllables [55][56]. A vowel space is a convenient conceptualization of the distribution of vowel qualities in a two-dimensional articulatory feature space [78]. The first dimension relates to the front-backness of tongue position and is closely associated with the difference between the second and first formant frequencies ($f_2 - f_1$); the second dimension relates to the height of the tongue body and is more closely correlated with the first formant frequency (f_1). A typical "vowel triangle" for American English is shown in Figure 5.1.

Figure 5.2 compares the vowel spaces associated with vocalic nuclei of fully accented syllables (upper panel) and that of completely unaccented syllables (lower panel). The data pertain to realized (transcribed) vocalic segments and a similar comparison for canonical vocalic segments (realized as such) can be found in [55][56]. For fully accented syllables (Figure 5.2, upper panel), the distribution of vocalic segments is relatively even across the articulatory space, except for very small numbers of [ix], [ax] and [ux] that are usually manifest in unaccented syllables, as well as [oy] which rarely occurs in the corpus. A dramatically different distribution of vowels is observed in the unaccented syllables (Figure 5.2, lower panel) where an overwhelmingly large number of vowels occur in the high-front ([ih],[iy],[ix]) or high-central ([ax]) portion of the vowel space, with most of the remaining segments positioned in the mid-front ([eh]) and mid-central ([ah]) locations. The number of diphthongs ([ay],[ey],[aw],[oy],[ow],[uw]) is significantly smaller in the unaccented syllables than in fully accented syllables (except for [iy]).

The analysis above suggests that the relationship between stress accent and vocalic identity (particularly vowel height) is clearly non-arbitrary. In fact, some of the previously discovered perceptual basis of stress accent is found to have a systematic relationship to vocalic identity. For example, as shown in Figure 5.3, the duration of vocalic nuclei are

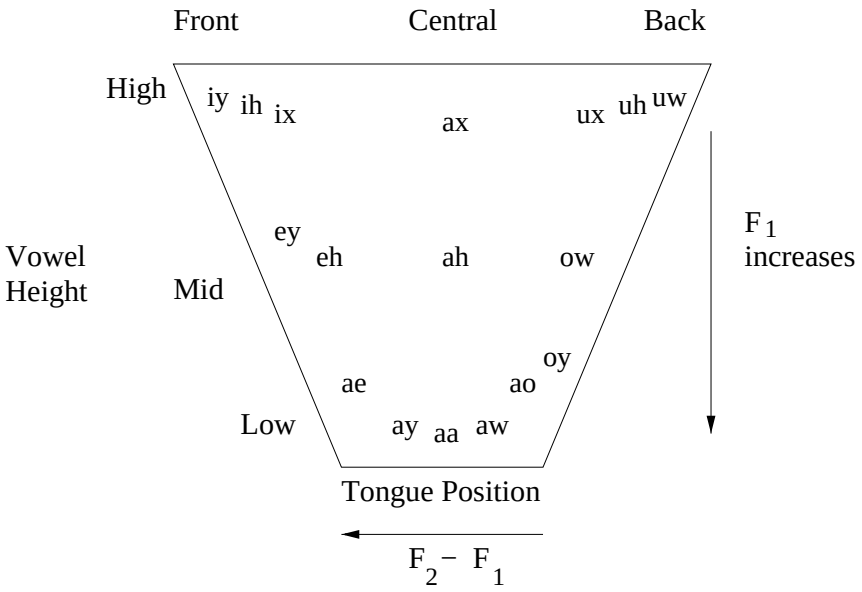

Figure 5.1: A typical "vowel triangle" for American English. The dynamic trajectories of the diphthongs ([iy], [uw], [ey], [ow], [oy], [ay], [ow]) are not shown for illustrative clarity.

consistently longer in fully accented syllables than in unaccented ones, and the differences are especially large for diphthongs and tense monophthongs ([ae],[aa],[ao]) [58]. As will be described in the final section of this chapter vocalic nucleus duration (as well as its ratio to syllable duration), the normalized vocalic nucleus energy and vocalic identity (either in terms of the transcribed label or spectro-temporal features derived from the acoustic signal) are found to be the most informative cues in an automatic stress-accent labeling system.

5.2 Stress Accent and Pronunciation Variation

The previous section presented a systematic relationship between vocalic identity and stress-accent level, suggesting that stress accent is far from an independent parameter of the speech utterance, orthogonal to the phonetic composition of the syllable. In this section the link between stress accent and the phonetic realization of speech is further investigated. We will describe how pronunciation in spontaneous speech is affected by stress-accent level (and vice versa) for different components of the syllable, as well as how pronunciation variation patterns are manifested across various articulatory-acoustic feature (AF) dimensions. The impact of stress accent on ASR performance has been described in Section 2.2.4 of the linguistic dissection of LVCSR systems, where a large variation in word-error rate (particularly the word deletion rate) was observed across stress-accent level. The interaction patterns of stress accent, syllable position and AFs are likely to help in developing more parsimonious models of pronunciation variation and thus improve ASR performance if incorporated appropriately.

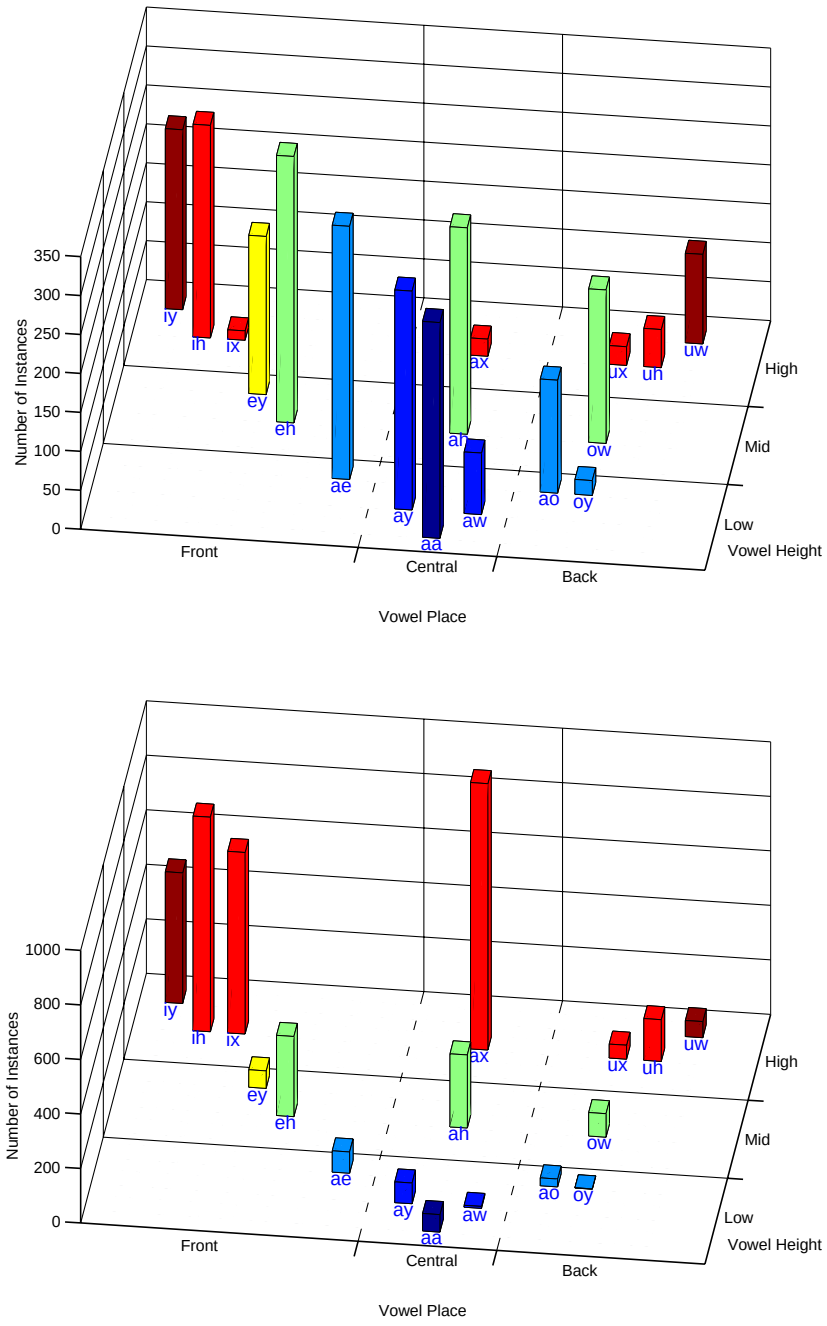

Figure 5.2: The distribution of realized (transcribed) vocalic segments associated with fully accented (upper panel) and unaccented (lower panel) syllables, from a 45-minute subset of the Switchboard corpus with manual stress-accent labels. Vowels in fully accented syllables have a relatively even distribution (with a slight preference to the front and central regions); vowels in unaccented syllables are highly concentrated in the high-front and high-central regions.

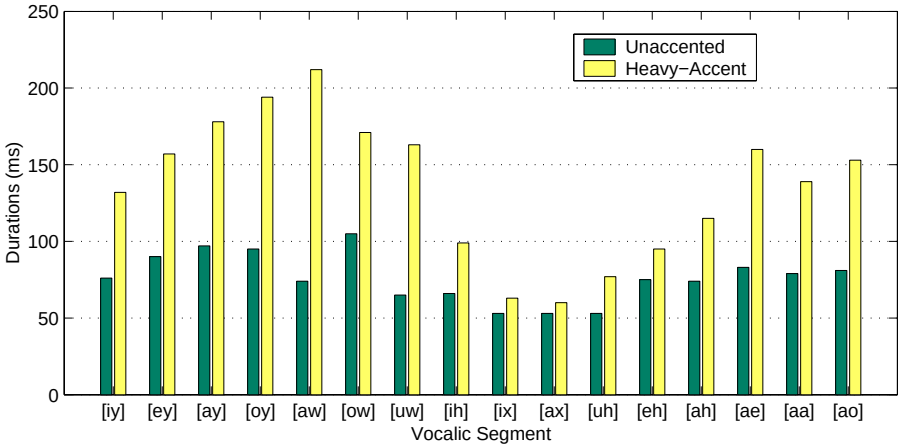

Figure 5.3: The relationship between segment duration and vocalic identity, partitioned into heavily accented and unaccented syllables (from [58]). The duration of vocalic nuclei are consistently longer in fully accented syllables than in unaccented ones, and the differences are especially large for diphthongs and tense monophthongs ([ae],[aa],[ao]).

5.2.1 Pronunciations of "That" – Revisited

Recall the example of the 63 different pronunciations of the word "that" presented in the previous chapter (cf. Section 4.3, Table 4.1) from the Year-2001 phonetic evaluation data. Partitioning the pronunciation deviations by position within the syllable (cf. Table 4.2) yields further insight into the pronunciation variation patterns than afforded by the overall deviation statistics. In this section, let us further partition the data by the stress-accent level associated with each instance of the word "that." Note that this exercise would be much less meaningful if we use lexical stress rather than perceived stress accent, as there is simply no differential markings of lexical stress for the different instances of the word "that" in a dictionary. Tables 5.1, 5.3 and 5.5 show the phonetic realizations (transcribed) associated with the unaccented, lightly accented and fully accented instances of "that," respectively, as well as the type of deviation partitioned according to position within the syllable. The corresponding summary information on pronunciation deviation is shown in Tables 5.2, 5.4 and 5.6.

Different patterns of deviations are observed across the stress-accent levels for each of the syllable positions:

- in onset position, the deviation pattern is relatively stable across stress-accent levels but with more deletions for unaccented syllables than instances with some degree of stress accent;
- in nucleus position, the rate of deviation (all substitutions) is very high (66.3%) for unaccented syllables, which is more than double the deviation rate of the lightly accented ones (30.9%), and the fully accented instances exhibit virtually no deviations

from canonical pronunciation;

- in coda position, unaccented syllables exhibit slightly more deviations (both substitution and deletion) than lightly accented syllables, while the deviation rate of the fully accented syllables is much smaller than that of the less accented ones.

These patterns roughly agree with those that observed on the subset of the Switchboard corpus with manual stress-accent labels (as described in Section 5.1.2) except that the magnitude of the differences as a function of stress-accent level is greater in the latter (cf. Figure 5.4, data from [55][56]). A likely reason for the less striking differences of deviation patterns in the instances of "that" is the disproportionally larger numbers of unaccented and lightly accented instances than that of the fully accented ones due to the special syntactic and semantic role often assumed by this word. It should also be noted that the word "that" is unusual with respect to the onset segment (i.e. [dh] in the canonical pronunciation), which tends to have a greater number of deviations from the canonical pronunciation than other onset segments. To get a perspective using a different word, refer to Appendix A for the sample pronunciations of the word "but" extracted from the same set of material. Finally, the Year-2001 phonetic evaluation data, from which the instances of the word "that" were extracted, differ in certain respect from the subset of the Switchboard data that were collected earlier.

From Figure 5.4 the deviation patterns of the unaccented syllables are significantly different from that of the fully accented ones, while the lightly accented syllables usually assume a pattern intermediate between the accent poles but their realization is often closer to the fully accented syllables [55]. Such deviation patterns suggest that it would be beneficial to explicitly model pronunciation variation with respect to stress accent.

5.2.2 Impact of Stress Accent by Syllable Position

It is evident from the previous discussion that the patterns of pronunciation variation are differentially realized as a function of the position of the segment within the syllable and stress-accent level. In many cases the phonetic realization deviates from the canonical pronunciation concurrently across several articulatory dimensions (e.g. a vocalic nucleus may change both its height and horizontal position) and the resulting pattern of variation, if considered at the phonetic-segment level, may appear complex and difficult to model. The statistics described in this section may help in the development of more parsimonious models of pronunciation variations by considering the variation patterns along several AF dimensions with respect to position within the syllable and stress-accent level. In the following figures (5.5-5.19), the AF-realization of each segment (in terms of the proportion of AF labels associated with the canonical pronunciation) is displayed (as well as segment deletions) for onset, nucleus and coda positions separately for each of several AF dimensions. In each condition only the statistics associated with the fully accented and unaccented syllables are displayed for illustrative clarity.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
dh ae t	11	-	-	-	n eh	1	S	S	D
dh ax	9	-	S	D	n ax	1	S	S	D
dh ax t	7	-	S	-	n ae t	1	S	-	-
dh ae	7	-	-	D	n aa	1	S	S	D
dh eh t	4	-	S	-	nx ax	1	S	S	D
dh eh	4	-	S	D	l ih	1	S	S	D
dh ih t	3	-	S	-	k ih dh	1	S	S	S
dh ax dx	3	-	S	S	iy	1	D	S	D
dh ae dx	3	-	-	S	dh iy	1	-	S	D
n ah t	2	S	S	-	dh ih	1	-	S	D
n ae	2	S	-	D	dh eh dx	1	-	S	S
dh ah t	2	-	S	-	dh ah dx	1	-	S	S
ah dx	2	D	S	S	dh ae ch	1	-	-	S
ae	2	D	-	D	ax dx	1	D	S	S
z ah p	1	S	S	S	ax dh	1	D	S	S
th ax t	1	S	S	-	ax	1	D	S	D
s ax t	1	S	S	-					

Table 5.1: Unaccented instances of the word "that" from the Year-2001 phonetic evaluation material (cf. Table 4.1). For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([dh ae t]) with respect to syllable onset, nucleus and coda. "S" is "substitution," "D" is "deletion," "I" is "insertion," and "-" is "no deviation."

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	72.5	17.5	10.0	0	80
Nucleus	33.8	66.3	0	0	80
Coda	40.0	18.8	41.3	0	80
Total	48.8	34.2	17.1	0	240

Table 5.2: Summary of phonetic deviations (from canonical) in terms of percentage of total segments (last column) in each syllable position (and overall), for the unaccented instances of the word "that" (cf. Table 5.1) from the Year-2001 diagnostic phonetic evaluation material.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
dh ae t	30	-	-	-	l ae dx	1	S	-	S
dh ae	20	-	-	D	eh t	1	D	S	-
dh aw	9	-	S	D	eh dx	1	D	S	S
dh ae dx	9	-	-	S	d ae t	1	S	-	-
n ae t	6	S	-	-	d ax p	1	S	S	S
dh eh t	6	-	S	-	dh ih t	1	-	S	-
n ae	4	S	-	D	dh ih d	1	-	S	S
dh ah t	3	-	S	-	dh ih	1	-	S	D
th ae t	2	S	-	-	dh eh	1	-	S	D
d ae	2	S	-	D	dh ah dx	1	-	S	S
t aw	2	S	S	D	dh ah d	1	-	S	S
t dh ae	1	I,-	-	D	dh ae d	1	-	-	S
t ax	1	S	S	D	dh ae b	1	-	-	S
t ae	1	S	-	D	dh aa t	1	-	S	-
th eh t	1	S	S	-	ae w	1	D	-	D
th eh	1	S	S	D	ae	1	D	-	D
th ax	1	S	S	D	aw	1	D	S	D
th aw t	1	S	S	-	ae t	1	D	-	-
n aw	1	S	S	D	hh ih t	1	S	S	-
n ae dx	1	S	-	S	z d ae	1	S,I	-	D
nx ae	1	S	-	D					

Table 5.3: Lightly accented instances of the word "that" from the Year-2001 phonetic evaluation material (cf. Table 4.1). For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([dh ae t]) with respect to syllable onset, nucleus and coda. "S" is "substitution," "D" is "deletion," "I" is "insertion," and "-" is "no deviation."

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	69.6	24.0	4.8	1.6	125
Nucleus	69.1	30.9	0	0	123
Coda	44.4	14.5	40.3	0	124
Total	61.2	23.2	15.1	0.5	371

Table 5.4: Summary of phonetic deviations (from canonical), in terms of percentage of total segments (last column) in each syllable position (and overall), for the lightly accented instances of the word "that" (cf. Table 5.3) from the Year-2001 diagnostic phonetic evaluation material.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
dh ae t	12	-	-	-	th ae t	1	S	-	-
dh ae	4	-	-	D	n ae t	1	S	-	-
d ae t	2	S	-	-	n ae	1	S	-	D
t b ae t	1	S,I	-	-	dh ae dx	1	-	-	S

Table 5.5: Fully accented instances of the word "that" from the Year-2001 phonetic evaluation material (cf. Table 4.1). For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([dh ae t]) with respect to syllable onset, nucleus and coda. "S" is "substitution," "D" is "deletion," "I" is "insertion," and "-" is "no deviation."

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	70.8	25.0	0	4.2	24
Nucleus	100.0	0	0	0	23
Coda	73.9	4.3	21.7	0	23
Total	81.4	10.0	7.14	1.4	70

Table 5.6: Summary of phonetic deviations (from canonical), in terms of percentage of total segments (last column) in each syllable position (and overall), for the fully accented instances of the word "that" (cf. Table 5.5) from the Year-2001 diagnostic phonetic evaluation material.

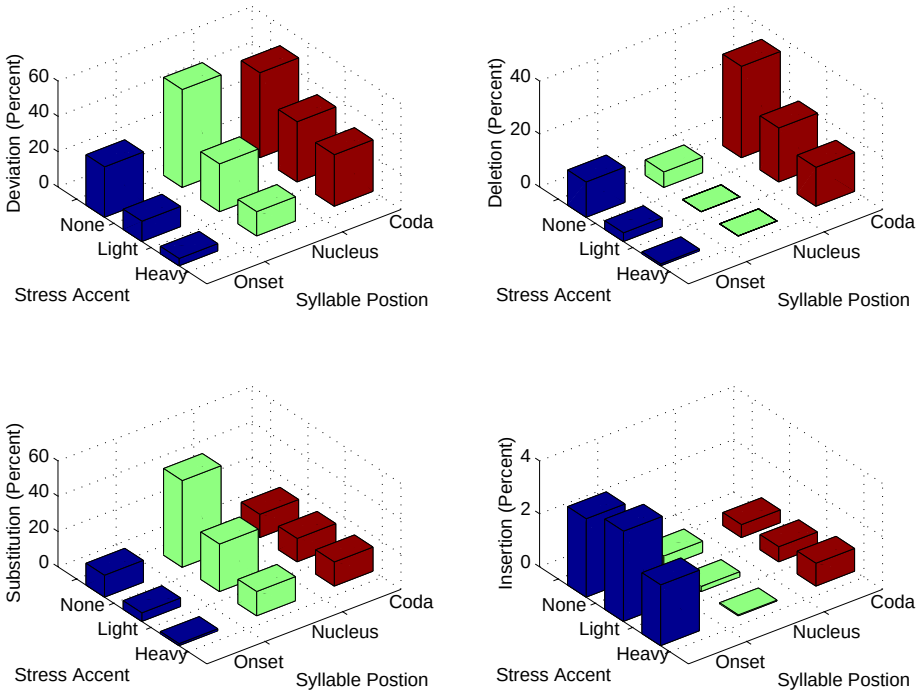

Figure 5.4: The impact of stress accent on pronunciation variation in the Switchboard corpus, partitioned by syllable position and the type of pronunciation deviation from the canonical form. The height of the bars indicates the percent of segments associated with onset, nucleus and coda components that deviate from the canonical phonetic realization. Note that the magnitude scale differs for each panel. The sum of the "Deletions", (upper right panel) "Substitutions" (lower left) and "Insertions" (lower right) equals the total "Deviation from Canonical" shown in the upper left panel. (From [55][56].)

While manner of articulation at onset is mostly canonical in fully accented syllables (Figure 5.5, left panel), a significant proportion of nasals (23%) and stops (15%) are realized as flaps in unaccented syllables (Figure 5.5, right panel). The only substitutions of manner at coda are also from either nasals or stops (into flaps). Unlike onsets, there is little difference between fully accented (Figure 5.6, left panel) and unaccented (right panel) syllables with respect to substitutions of manner at coda. The proportion of manner substitution for unaccented syllables is smaller at coda than at onset. Codas, however, possess a significantly greater proportion of deletions than onsets, particularly in unaccented syllables. For example, 63% coda approximants and 55% coda stops are deleted in unaccented syllables, while 20% onset approximants and 11% onset stops are deleted in unaccented syllables. Among various manner features at coda, approximants have the largest difference in deletion rates between fully accented (63% deleted) and unaccented (11% deleted) syllables, suggesting the preservation of coda approximants may be a potential cue for stress accent.

Voicing feature is relatively stable at onset (cf. Figure 5.7), particularly of fully accented syllables. There is a greater proportion of unvoiced segments realized as voiced (at onset of unaccented syllables and at coda of all syllables) than vice versa (cf. Figure 5.7 and 5.8). There is a large (and roughly equal) proportion of segment deletion for both voiced and unvoiced codas of unaccented syllables. The relatively low proportion of voicing substitution suggests that voicing feature does not play a significant role in pronunciation variation of spontaneous speech. Majority of voicing deviations (83% onset and 62% coda voicing deviations from canonical) are accompanied by a concomitant deviation in manner of articulation, also suggesting a subordinate role of voicing feature to manner of articulation [53].

Because of the differential manifestation of the place constriction for different manner segments, the statistics for place of articulation (for onsets and codas) are partitioned into different manner classes. Figures 5.9 and 5.10 show the place deviation patterns for fricatives at onset and coda, respectively. Aside from deletions, most fricative place features are stable, except for a small number of interchanges (in both directions) between alveolars (e.g. [s],[z]) and palatals (e.g. [sh],[zh]) at onset. There are significant proportions of deletions at onset of unaccented syllables: 45% glottal fricatives (e.g. [hh]) and 20% dental fricatives (e.g. [th],[dh]). Both labial (e.g. [f],[v]) and dental fricatives have over 30% deletions at coda of unaccented syllables.

The non-deleted stops at both onset and coda (Figures 5.11 and 5.12) are relatively canonical with respect to place; the only deviations observed are some alveolars (and labials) realized as glottal stops (i.e. [q]) at coda. Although the canonical forms of stops are relatively evenly distributed across different places at syllable onsets, the vast majority (ca. 75%) of coda stops are alveolar [56]. Interestingly, in unaccented syllables, a large fraction (ca. 60%) of alveolar stops are deleted with respect to canonical, leading to a high likelihood of overall stop-segment deletion in unaccented codas. A likely reason for the tendency of alveolar coda deletion is the sharing of acoustic cues in the mid-frequency (ca. 1500-2200 Hz) region of the spectrum with the preceding vocalic nucleus [56]. This tendency also appears in nasal codas (see below).

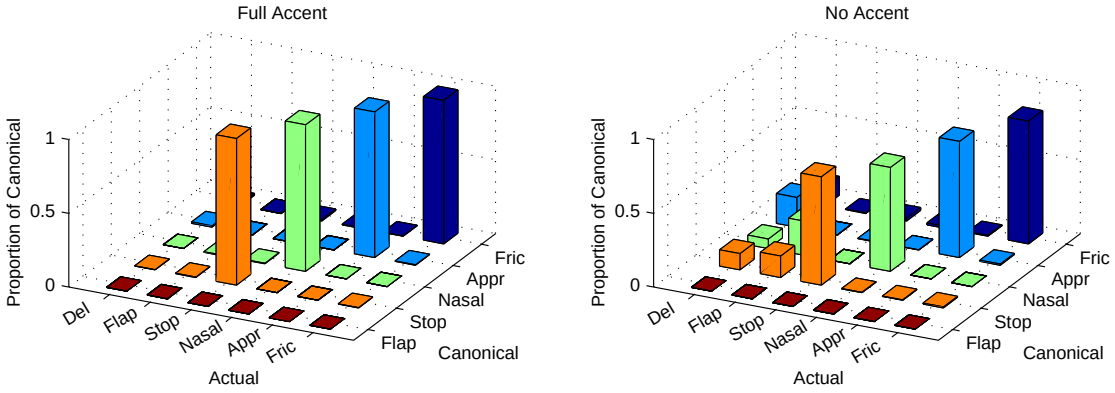

Figure 5.5: The realization of manner of articulation in onset position (proportion of manner labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Appr" is approximants, and "Fric" is fricatives. Note that there are no flap segments in canonical pronunciations.

The nasals at onset (as well as at coda of fully accented syllables) are mostly canonically realized. (Figures 5.13 and 5.14). However, at coda of unaccented syllables a high proportion (27%) of velar nasals (e.g. [ng]) are realized as alveolars (e.g. [n]), but not vice versa. Alveolar and labial nasals at coda exhibit a significant proportion of deletions, particularly of unaccented syllables.

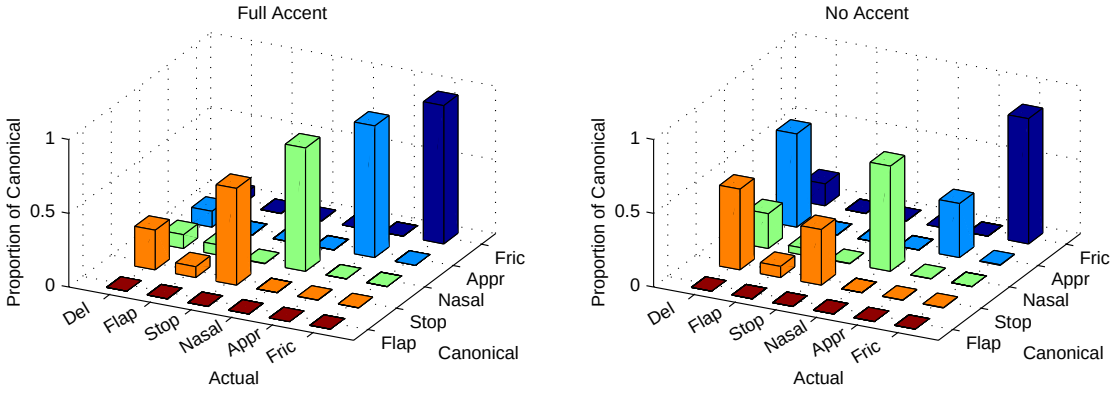

Figure 5.6: The realization of manner of articulation in coda position (proportion of manner labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Appr" is approximants, and "Fric" is fricatives. Note that there are no flap segments in canonical pronunciations.

Syllable Nucleus

As discussed previously (cf. Figure 5.4), syllable nuclei are far more likely to exhibit segmental substitutions (from canonical) than onsets and codas. However, segmental deletions are rare for nuclei even in unaccented syllables. Figure 5.15 compares the patterns of vocalic height deviation from canonical in fully accented (left panel) and unaccented syllables (right panel). While vowel height is largely canonical in fully accented syllables (except for some "mid" to "low" movement), the unaccented syllables exhibit a large number of height movements, particularly from "low" to "mid" (33%) and "high" (34%), and from "mid" to "high" (24%). This height-raising pattern largely agrees with the tendency of having high and mid-vowels in unaccented syllables discussed in the previous section (cf. Section 5.1.2). The only height movement in the reverse direction is a small proportion of high vowels that are realized as mid-vowels (but not as low vowels).

The vocalic segments in unaccented syllables also exhibit an anterior (front) preference of horizontal place. As shown in Figure 5.16 (left panel), a significant proportion of back vowels are realized as either central (32%) or front (27%). There are also many more central vowels realized as front (19%) than realized as back (4%). This pattern also agrees with the tendency of having front and central vowels in unaccented syllables discussed in the previous section (cf. Section 5.1.2).

Stress accent has a large impact on the realization of lip-rounding features. As shown in Figure 5.17, while the unrounded vowels mostly remain unrounded, the proportion of rounded vowels becoming unrounded is much greater in unaccented syllables (73%) than in fully accented syllables (22%). Since the total number of unrounded vowels (if pronounced canonically) in the Switchboard corpus is much larger (by at least five times) than the number of rounded vowels, this implies that very few vocalic segments are realized

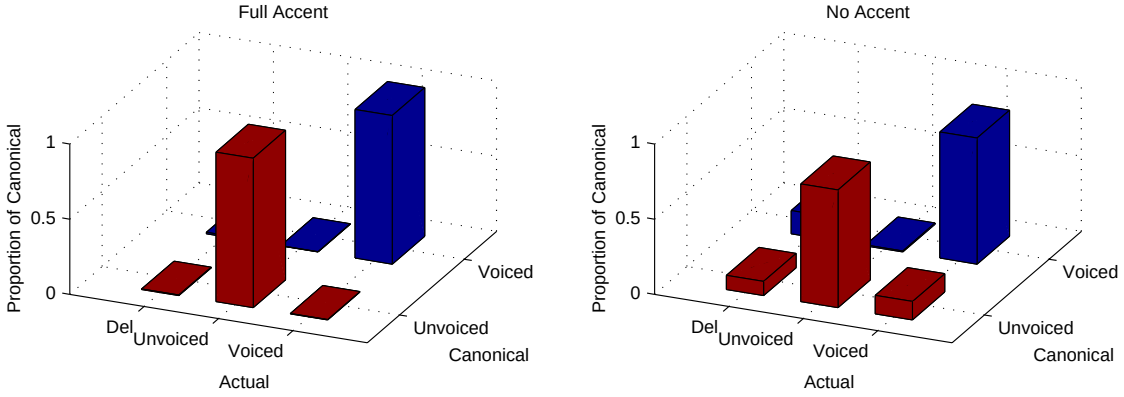

Figure 5.7: The realization of voicing in onset position (proportion of voicing labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion.

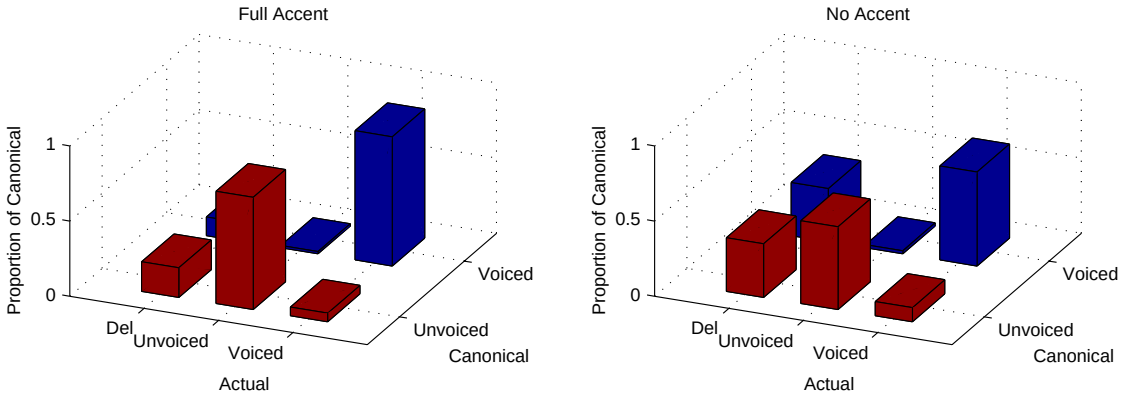

Figure 5.8: The realization of voicing in coda position (proportion of voicing labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion.

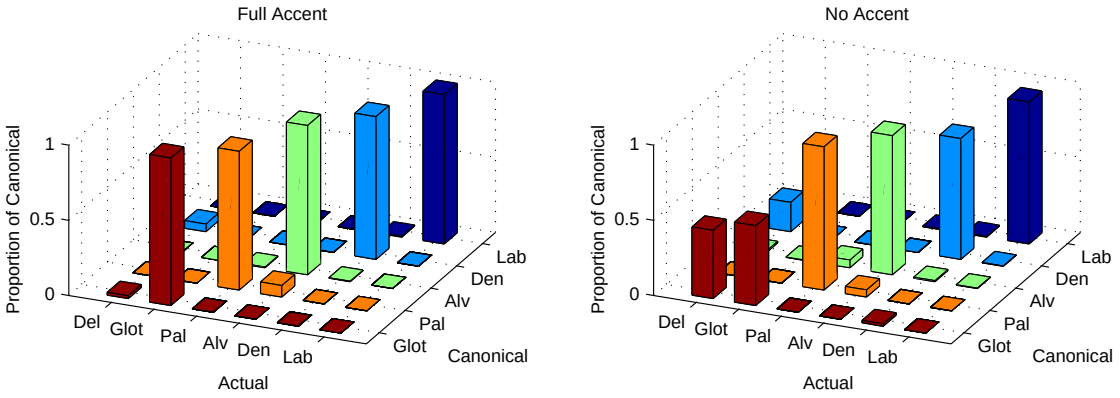

Figure 5.9: The realization of place of articulation for all fricatives in onset position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Glot" is glottal, "Pal" is palatal, "Alv" is alveolar, "Den" is dental and "Lab" is labial.

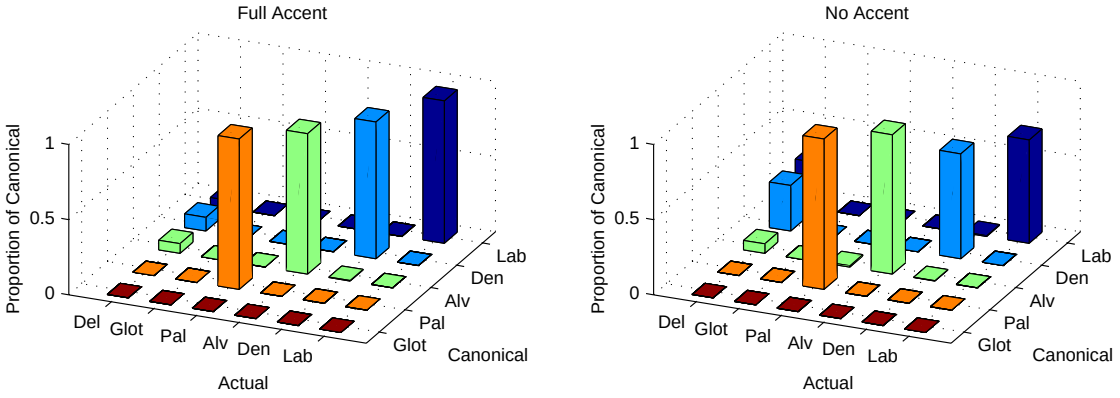

Figure 5.10: The realization of place of articulation for all fricatives in coda position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Glot" is glottal, "Pal" is palatal, "Alv" is alveolar, "Den" is dental and "Lab" is labial. Note that there are no glottalic segments in canonical pronunciations of coda fricatives.

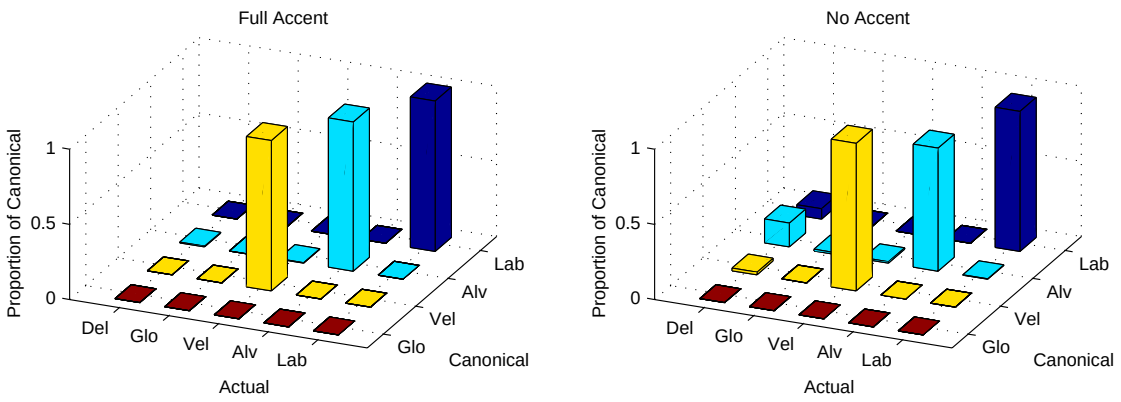

Figure 5.11: The realization of place of articulation for all stops in onset position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Glo" is glottal, "Vel" is velar, "Alv" is alveolar and "Lab" is labial. Note that there are no glottalic segments in canonical pronunciations of onset stops.

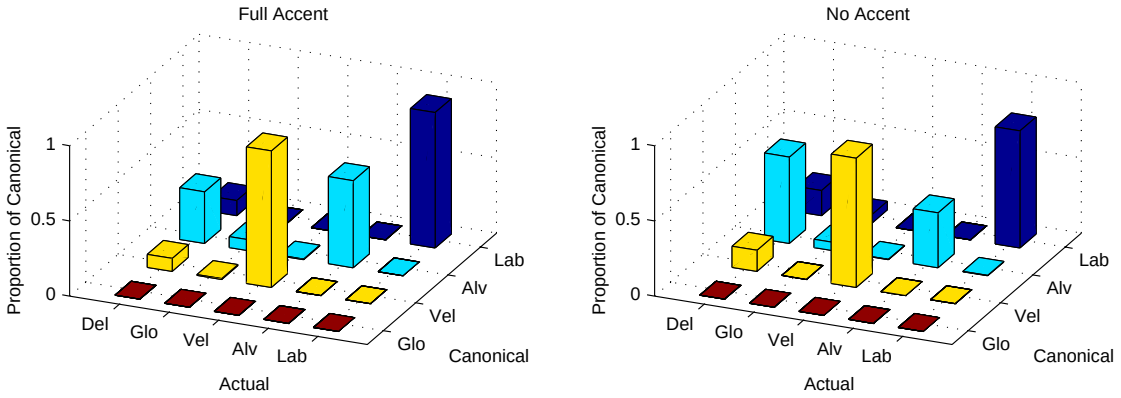

Figure 5.12: The realization of place of articulation for all stops in coda position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Glo" is glottal, "Vel" is velar, "Alv" is alveolar and "Lab" is labial. Note that there are no glottalic segments in canonical pronunciations of coda stops.

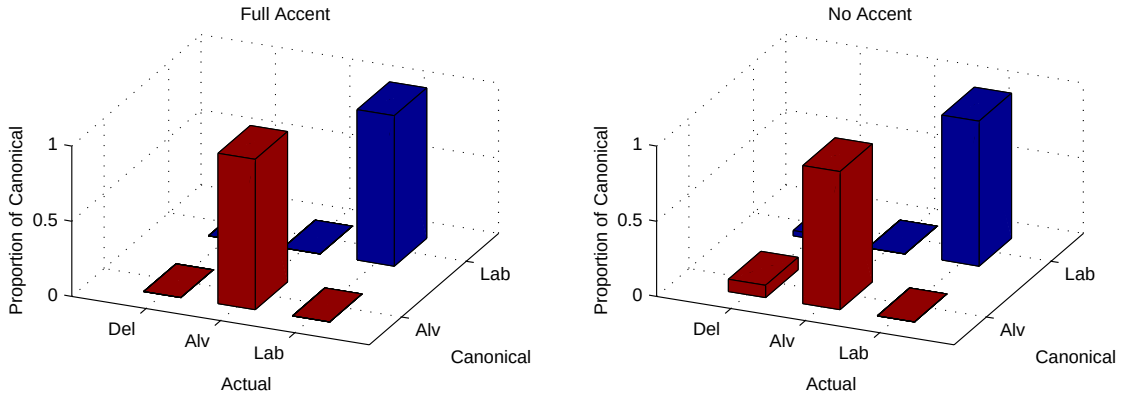

Figure 5.13: The realization of place of articulation for all nasals in onset position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Vel" is velar, "Alv" is alveolar and "Lab" is labial.

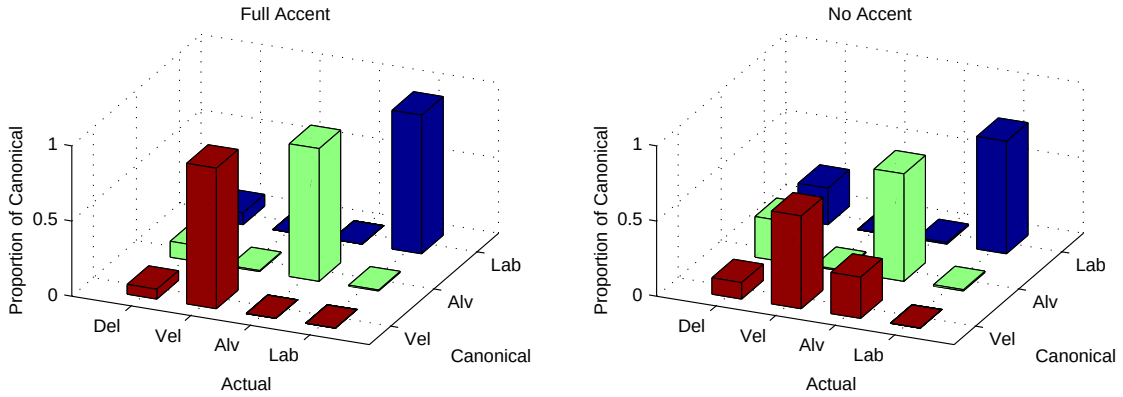

Figure 5.14: The realization of place of articulation for all nasals in coda position (proportion of place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Vel" is velar, "Alv" is alveolar and "Lab" is labial.

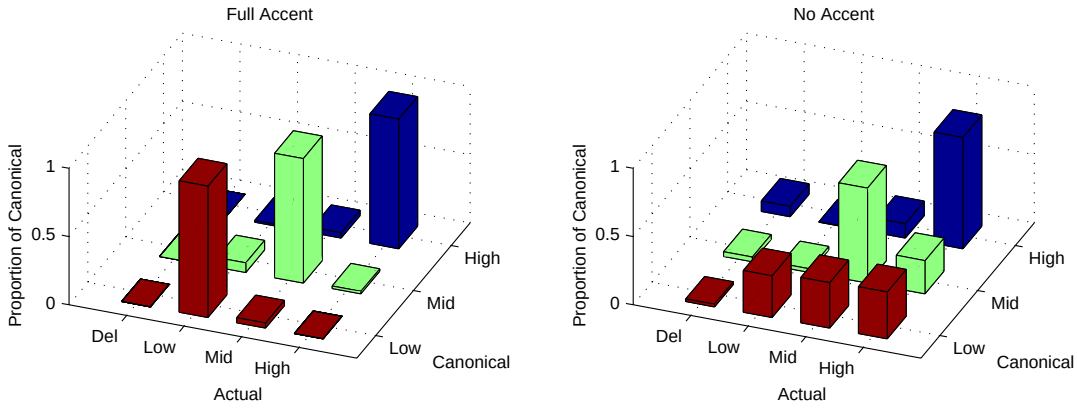

Figure 5.15: The realization of vocalic height in nucleus position (proportion of vocalic height labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion.

as lip-rounded (particularly in unaccented syllables). There is also an interesting correlation between lip-rounding deviation and horizontal-vowel-place deviation from canonical for unaccented syllables. While only 15% lip-rounding deviations are accompanied by a horizontal-vowel-place deviation for fully accented syllables, 82% lip-rounding deviations are accompanied by a horizontal-vowel-place deviation for unaccented syllables. The situation is just opposite for between lip-rounding and vocalic height. While 68% lip-rounding deviations are accompanied by a vocalic-height deviation for fully accented syllables, only 15% lip-rounding deviations are accompanied by a vocalic-height deviation for unaccented syllables.

The last two comparisons pertain to the tenseness of a vowel (cf. Figure 5.18) and whether the spectrum is relatively static (monophthongs) or dynamic (diphthongs) (cf. Figure 5.19). In the first instance, the vowels in fully accented syllables are relatively stable with respect to tenseness, while for unaccented syllables a large proportion (51%) of tense vowels[2] are realized as lax vowels. In the second instance, a large proportion (35%) of diphthongs are realized as monophthongs in unaccented syllables and there are far more diphthongs becoming monophthongs than vice versa in either fully accented or unaccented syllables. The overwhelming tendency for unaccented syllables to contain lax vowels and monophthongs is consistent with the vocalic duration distribution discussed in the previous section (cf. Figure 5.3) since both of these two AF dimensions are closely related to the duration of vocalic segments.

[2]All diphthongs are grouped with the tense vowels.

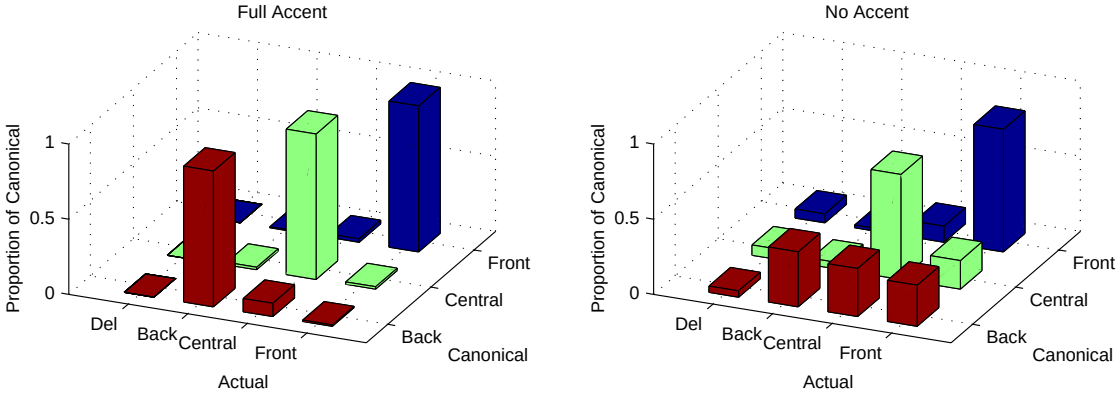

Figure 5.16: The realization of horizontal vocalic place (front-central-back) in nucleus position (proportion of vocalic place labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion.

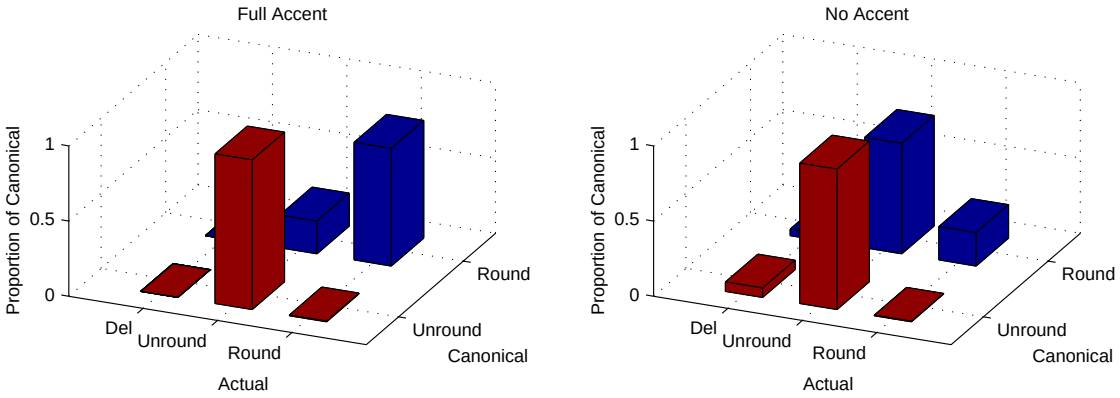

Figure 5.17: The realization of lip-rounding in nucleus position (proportion of rounding labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion.

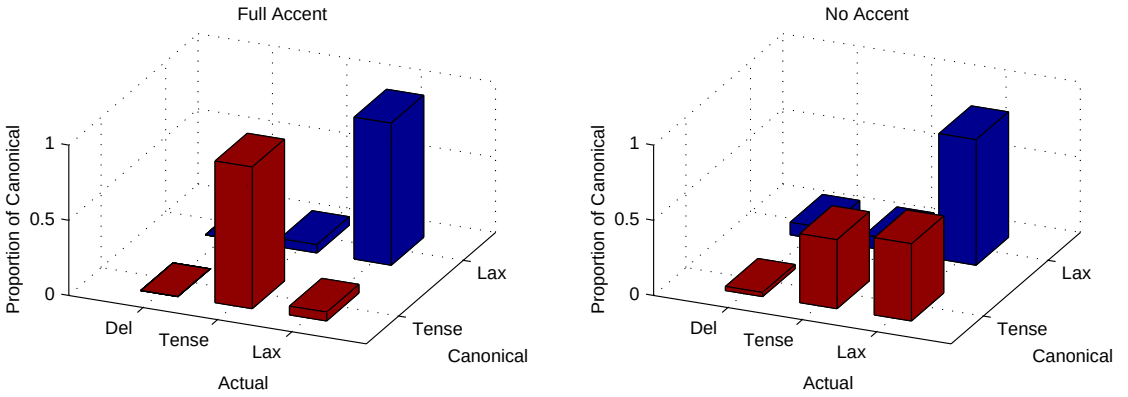

Figure 5.18: The realization of tense/lax features in nucleus position (proportion of tense/lax labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. All diphthongs are considered to be tense vowels. "Del" is deletion.

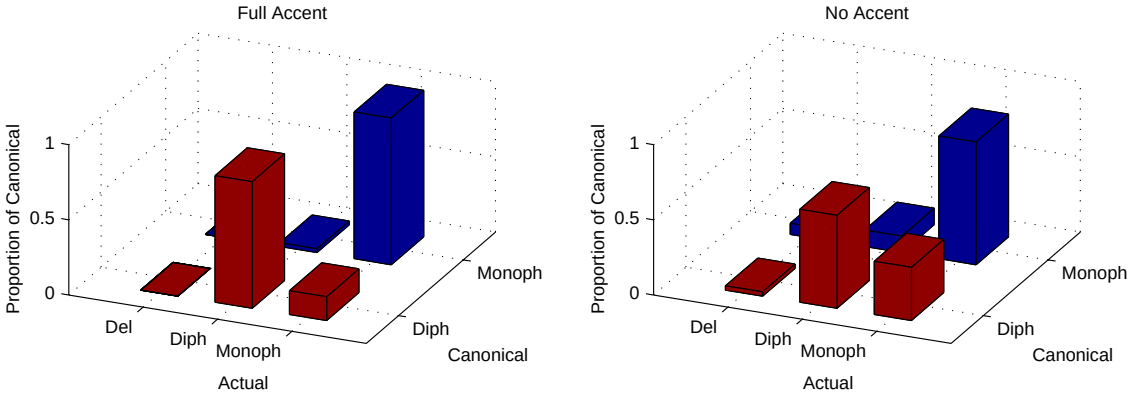

Figure 5.19: The realization of static/dynamic features (monophthong vs. diphthong) in nucleus position (proportion of static/dynamic labels as the canonical pronunciation), for fully accented (left panel) and unaccented (right panel) syllables. "Del" is deletion, "Diph" is diphthong and "Monoph" is monophthong.

5.3 Automatic Stress-Accent Labeling of Spontaneous Speech

The pronunciation variation patterns presented in the previous section suggest a potentially significant advantage of incorporating stress-accent information in ASR systems. This would require an automatic means of estimating stress-accent levels of each syllable (and of course, methods of automatic detection of the syllables as well). In [121][122], Silipo and Greenberg describe experiments of automatically detecting prosodic stress accent of syllables on the OGI Stories [11] corpus using features pertaining to syllable-nucleus duration, energy and f_0-related information. This section describes the development of a neural-network-based automatic stress-accent labeling system (AutoSAL) for spontaneous American English [58] using the Switchboard corpus [42][49].

Besides being used by an ASR system, the development of the AutoSAL system has additional advantages. Just like manually transcribing speech phonetically, manual labeling of stress accent is a very tedious and time-consuming task. In order to label accurately and consistently, the transcribers are required to have special linguistic training and substantial practice. The prohibitively high cost makes it unfeasible to totally rely on human transcribers to label stress accent for large quantities of speech material in novel corpora and diverse languages, which would facilitate useful statistical analysis of stress accent such as that performed in this thesis project, as well as form the basis of training material for ASR systems incorporating stress-accent information. Thus, an accurate means of automatic labeling of stress accent would be very helpful.

Another motivation of developing the AutoSAL system is to assess the utility of various features in the estimation of stress-accent levels. As discussed earlier in this chapter, the perceptual basis of stress accent is not entirely agreed upon by linguists. Further experiments of automatic stress-accent labeling would certainly help advance our understanding of the stress-accent phenomenon.

5.3.1 System Description

The AutoSAL system uses multi-layer-perceptron (MLP) neural networks to estimate the stress-accent level of each syllable based on a number of features derived from the vocalic nucleus segment and the surrounding consonantal contexts. In the experiments described in this section, the input features to the MLP network include some or all of the following features:

- Duration of the vocalic nucleus (in 10-ms-frame units)
- The integrated energy of the vocalic nucleus represented as a Z-score (i.e., in terms of standard-deviation units above or below the mean) normalized over a three-second interval of speech (or less for utterances shorter than this limit) centered on the mid-point of the nucleus
- The average of the critical-band, log-energy (cf. Section 3.2.1), as well as the corresponding delta and double-delta features pertaining to the interval of the vocalic

nucleus

- Vocalic identity – this feature has 25 possible outputs, each corresponding to a specific vocalic-segment label
- Vocalic height (0 for low, 1 for mid and 2 for high)
- Vocalic place (0 for front, 1 for central and 2 for back)
- The ratio of the vocalic-nucleus duration relative to the duration of the entire syllable
- Gender of the speaker (male or female)
- Minimum-maximum (dynamic range) of vocalic f_0[3]
- Mean vocalic f_0
- Static/Dynamic Property of Nucleus (Diphthong/Monophthong)

Each of the input features can be derived either automatically from acoustics or from existing transcripts at phonetic-segment or syllable levels depending on availability. The MLP network contains a single hidden layer of 40 units and is trained with a standard online back-propagation algorithm [112] adapted to speech processing [90]. Each input feature is normalized to have zero mean and unit variance before feeding into the MLP network.

The training data were derived from the 45-minute Switchboard material with manual stress-accent labels (cf. Section 5.1.2). Each vocalic nucleus has an associated five-level stress-accent value based on the average accent levels from the two transcribers. Computation of the human/machine concordance is based on a 5-tier system of stress accent (accent levels of 0, 0.25, 0.5, 0.75 and 1). In the manually transcribed material 39.9% of the syllables are labeled as being entirely unaccented (Level-0 accent), and 23.7% of the syllables labeled as fully accented (Level-1 accent). The remaining nuclei are relatively equally distributed across accent levels (0.25: 12.7%; 0.5: 13%; 0.75: 10.7%).

5.3.2 Experiments on the Switchboard Corpus

Several dozen different combinations of input features (as described above) were tested for their utility in estimating stress accent. Table 5.7 shows the different feature combinations that have been used. Due to the limited amount of material with manually derived stress-accent labels, the testing was performed using a four-fold, jack-knifing procedure: (1) the data were partitioned into four groups; (2) three groups were used for training and the remaining group for testing; (3) the groups were rotated and the procedure was repeated; (4) the average of test performances on the four groups was taken.

The performance associated with various input feature combinations (cf. Table 5.7) are presented in Figure 5.20 in terms of a normalized measure of concordance with the manual labels. This normalized measure is obtained by linearly scaling the stress-accent label classification accuracy associated with using the various feature combinations into a

[3]The f_0 is calculated using a gender-dependent ensemble autocorrelation method [121].

No.	Feature Set Specification		
1	Vocalic place (front-central-back)[Voc-Place]		
2	Nucleus/syllable duration ratio [N_S-Dur-Ratio]		
3	Speaker gender [Gender]		
4	Minimum-maximum (dynamic range) of vocalic f_0 [f_0-Range]		
5	Mean vocalic f_0 [f_0-Mean]		
6	Static/dynamic property of nucleus (Diphthong/Monophthong)[Voc-Dyn]		
7	Vocalic height (high-mid-low) [Voc-Height]		
8	Average vocalic-segment spectrum [Voc-Spec]		
9	Vocalic identity [Voc-ID]		
10	Vocalic-segment duration [Voc-Dur]		
11	Voc-Spec+delta features [Voc-Spec_D]		
12	Normalized energy (of the nucleus relative to the utterance) [Z-Energy]		
13	Voc-Spec+delta and double-delta features [Voc-Spec_D_DD]		
14	(4)+(5)	30	(1)+(7)+(12)
15	(1)+(7)	31	(1)+(7)+(13)
16	(4)+(9)	32	(2)+(4)+(10)
17	(4)+(10)	33	(4)+(10)+(12)
18	(4)+(12)	34	(9)+(10)+(12)
19	(9)+(10)	35	(10)+(12)+(13)
20	(2)+(10)	36	(1)+(7)+(10)+(12)
21	(4)+(13)	37	(4)+(10)+(12)+(13)
22	(9)+(12)	38	(3)+(10)+(12)+(13)
23	(9)+(13)	39	(9)+(10)+(12)+(13)
24	(12)+(13)	40	(2)+(10)+(12)+(13)
25	(10)+(12)	41	(3)+(9)+(10)+(12)+(13)
26	(10)+(13)	42	(2)+(4)+(9)+(10)+(12)
27	(1)+(6)+(7)	43	(2)+(3)+(9)+(10)+(12)
28	(1)+(7)+(9)	44	(2)+(3)+(4)+(5)+(9)+(10)+(12)+(13)
29	(1)+(7)+(10)	45	(2)+(3)+(9)+(10)+(12)+(13)

Table 5.7: Various input features (and feature combinations) used in developing the automatic stress-accent labeling (AutoSAL) system. The specifications of feature sets 14-45 refer to combinations of singleton feature sets 1-13, e.g. set #14 is [f_0-Range] + [f_0-Mean]. Features listed pertain to those shown in Figure 5.20.

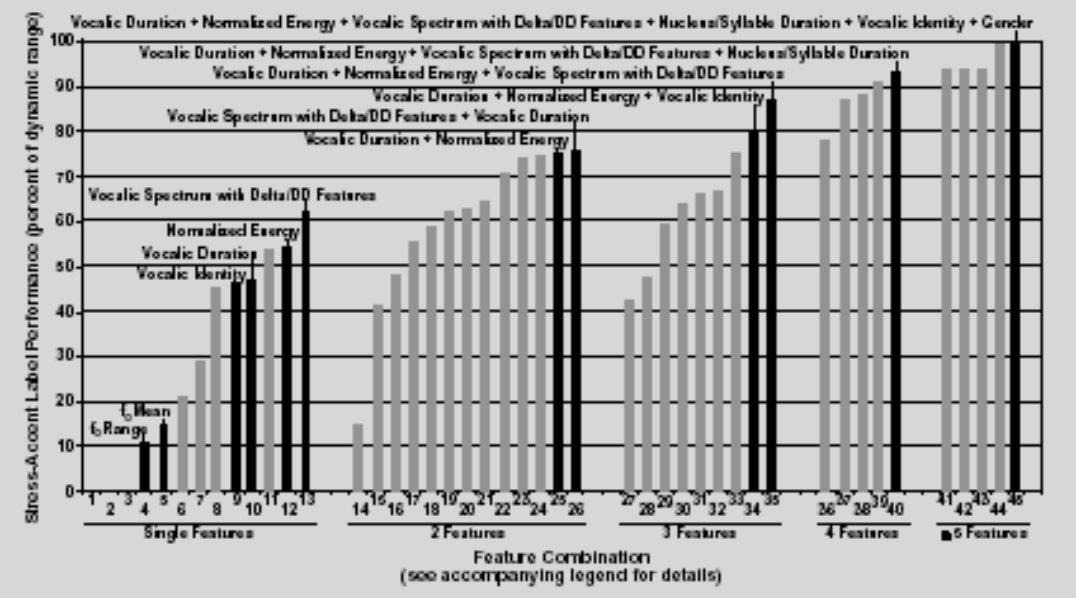

Figure 5.20: Normalized concordance (linearly scaled between 0 and 100) between manually transcribed stress-accent labels and the AutoSAL-generated labels using different combinations of input features listed in Table 5.7. A concordance of 0 is roughly equivalent to chance performance (ca. 40% concordance, using only the prior distribution of stress-accent levels) and 100 is comparable to the concordance between two human transcribers (ca. 67.5%). These results are based on an analysis using a tolerance step of 0 (i.e., an exact match between human and machine accent labels was required for a hit to be scored) and a three-accent-level system (where 0.25 and 0.75 accent outputs were rounded to 0.5). (Figure from [52].)

dynamic range between 0 and 100, where 0 pertains to the concordance of the most poorly performing feature set (#1, roughly equivalent to the chance performance by using only the prior distribution of different stress-accent levels) and 100 pertains to the concordance of the best performing feature set (# 45), comparable to the overall concordance between two human transcribers. Figure 5.21 shows the concordance of the AutoSAL output (using input feature combination #45) with the (average) manually labeled stress-accent material in terms of two levels of tolerance – a quarter and a half step. A syllable is scored as correctly labeled if the AutoSAL system output is within the designated tolerance limit. Such a metric is required in order to compensate for the inherent "fuzziness" of stress accent in spontaneous material, particularly for syllables with some degree of accent. The average concordance is 77.9% with a quarter-step tolerance and 97.5% with a half-step tolerance.

Features pertaining to vocalic identity (either the vocalic identity derived from segment labels or the vocalic spectrum with delta features), vocalic-segment duration, and normalized energy are most closely associated with stress accent. The contribution of f_0-based features is relatively small, especially when the three most effective features delineated above are already present. These observations provide further evidence concerning the

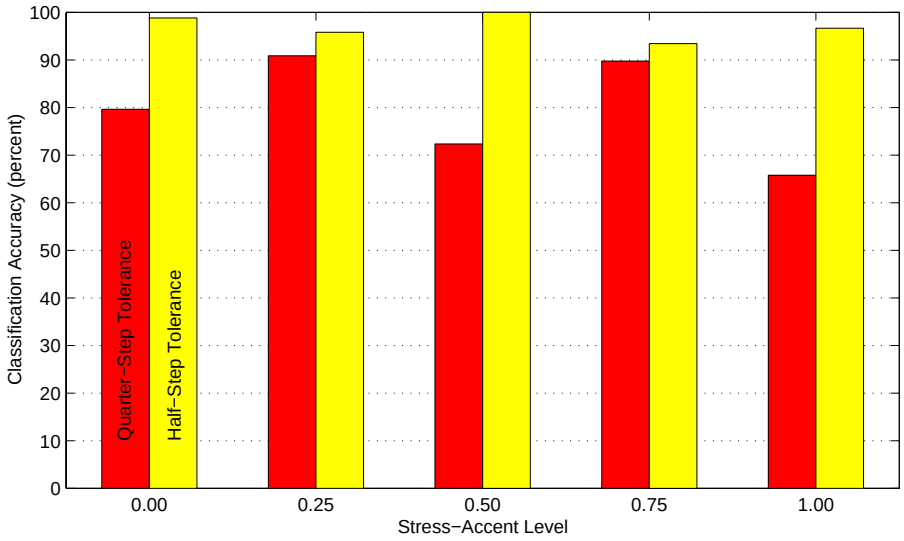

Figure 5.21: Classification accuracy of the automatic (MLP-based) stress-accent labeling (AutoSAL) system for the Switchboard corpus using two degrees of accent-level tolerance – quarter-step and half-step, on a five-level stress-accent scale. The results were obtained using the best performing feature set (#45 in Table 5.7). (From [58].)

perceptual basis of stress accent discussed in the first section of this chapter. Both the ratio of vocalic duration to syllable duration and speaker gender are also contributing features, with information complementary to the three most effective features.

5.4 Summary

Stress accent is an important component of spoken English and has great impact on the pronunciation variation of spontaneous speech. This chapter first introduced the background of the stress accent with its perceptual basis in spoken language, and in particular, the relationship between vocalic identity and stress accent in spontaneous speech was emphasized.

The link between stress accent and the phonetic realization of spontaneous speech was investigated in detail, first by revisiting the word pronunciation variation example (introduced in the previous chapter) to include the effect of stress-accent levels, and then by describing the general articulatory-acoustic feature deviation patterns (from the canonical forms) as a function of both syllable position and stress accent. Some gross AF realization patterns are:

- Without stress accent, vocalic nuclei often move "up" (in height) and "forward" (in horizontal vowel place) and tend to have reduced duration, less lip-rounding and relatively stable spectra.

- Onsets are the most canonical and the only deviations from canonical forms are usually deletions and flapping.
- Codas often have a large number of deletions compared to the canonical forms, especially for coda segments of central places (e.g. [t], [d], [n]).

It was shown that pronunciation variation of spontaneous speech can be largely captured by the systematic deviation patterns of AFs from the canonical forms within the context of particular syllable position and stress-accent levels.

To take advantage of stress-accent information in automatic speech recognition, a neural-network-based system was developed to automatically label stress accent for spontaneous speech. The Switchboard-corpus-trained system was able to perform at a level comparable to human transcribers. A large number of feature combinations were assessed for their contribution to stress-accent labeling and the most salient features included the duration, energy and vocalic identity (in terms of vocalic labels or spectral features), and the pitch-related features that we chose were found to play only a minor role.

The findings from this chapter, especially the systematic patterns of AF deviation from the canonical forms within the context of syllable position and stress accent, in conjunction with the discussion of AF and syllable processing in the previous two chapters, provide a basis for a multi-tier model of speech recognition to be introduced in the next chapter.

Chapter 6

A Multi-tier Model of Speech Recognition

The linguistic dissection of LVCSR systems described in Chapter 2 identified a number of acoustic and linguistic factors that affect word-recognition performance. In subsequent chapters, a detailed study of articulatory-acoustic features (AFs), syllables and stress accent was described based on empirical data from several American English corpora. In particular, the studies showed that the complex phenomenon of pronunciation variation in spontaneous speech may be characterized succinctly using information pertaining to AFs within the context of syllable position and stress accent. Motivated by such results, a multi-tier model of speech recognition is described in this chapter for spontaneous speech. The model adopts a syllable-centric organization for speech and uses features pertaining to a number of AF dimensions to describe the detailed phonetic realizations organized with respect to syllable position and stress-accent level. Although the model has obvious limitations, it should be viewed as a first step toward bridging the gap between automatic recognition and the reality of spontaneous speech using insights gained through statistical analysis of spontaneous spoken material.

The following section provides a general description of the multi-tier model, independent of any particular corpus, followed by a discussion of its feasibility and functionality. In the sections to follow, an implementation of a testbed system is described based on the model proposed. Rather than trying to develop a scalable, powerful performance system, the implementation is designed to provide preliminary answers to questions posed about the multi-tier model through controlled experiments within a highly constrained task domain. Experimental results and analysis based on the implementation will be described in the following chapter.

6.1 Model Description

In Chapter 4 evidence was presented in support of the syllable as the fundamental binding unit of speech, around which information from other linguistic tiers is organized.

This position is integrated within the multi-tier model, which considers speech to be organized as a sequence of syllables (in contrast to the conventional phonetic-segment-based organization assumed by most ASR systems).

In the multi-tier model each syllable contains a nucleus element, which is almost always a vocalic segment. In addition, the nucleus may be preceded by an optional onset element and followed by an optional coda element; both the onset and coda elements may contain one or more consonantal segments (i.e. a consonantal cluster).

Each syllable carries a certain level of accent (or lack of accent), which could be stress- or nonstress-based (e.g. pitch accent) depending on the nature of the specific language. The accent level can be coarse but should at least distinguish among completely unaccented syllables, fully accented syllables, and possibly syllables of intermediate accent. The onset, nucleus and coda of each syllable are described by features along several quasi-orthogonal AF dimensions, such as manner of articulation, place of articulation (manner-specific or manner-independent), voicing, vocalic height, tenseness (lax vs. tense) and spectral dynamics (monophthong vs. diphthong), lip-rounding, duration, etc.

Instead of storing every possible pronunciation of each lexical entry in the vocabulary, the model lexicon consists of simple representations of words (or short phrases) with only a single canonical baseform (or in cases of distant or non-systematic variations in pronunciation, a small number of such forms). Each baseform contains one or more syllables, described by the canonical AF specification and accent levels. However, the small number of canonical baseforms does not necessarily restrict the multi-tier model from representing a large number of pronunciation variants for each lexical entry. During recognition, pronunciation transformation rules (e.g. in terms of statistical characterization of AF transformation from canonical to realized forms) take many potential systematic variations of the baseforms into account, effectively expanding the pronunciation coverage. These transformation rules may depend on a number of contextual cues, such as syllable position and accent levels. In order to explain the capability of human listeners to understand novel realizations of familiar lexical entries, such a succinct representation of the lexicon with a dynamically expanding pronunciation may be more intuitively appealing than explicitly storing all possible pronunciation variants. It is very likely that human listeners require only a few sample pronunciations and variation patterns generalized from previous experience. In addition, such a representation incorporating explicit pronunciation variation patterns may reduce the problem of lexical-phonetic mismatch.

The recognition process starts by classifying features along each of the AF dimensions. Not all AF dimensions are equally important for particular lexical classification and not all AFs can be determined with equally high confidence. This is especially true for speech in adverse acoustic conditions or spoken in an unusual way. It is thus important to treat features differentially according to their potential contribution to recognition and the level of classification accuracy that can be obtained.

Among various AF dimensions, the manner of articulation, especially the vocalic class, is particularly important and plays a key role in approximate syllable detection and segmentation. This initial syllable processing also provides for accent-level estimation and serves as the basis for subsequent syllable-position and accent-based pronunciation variation

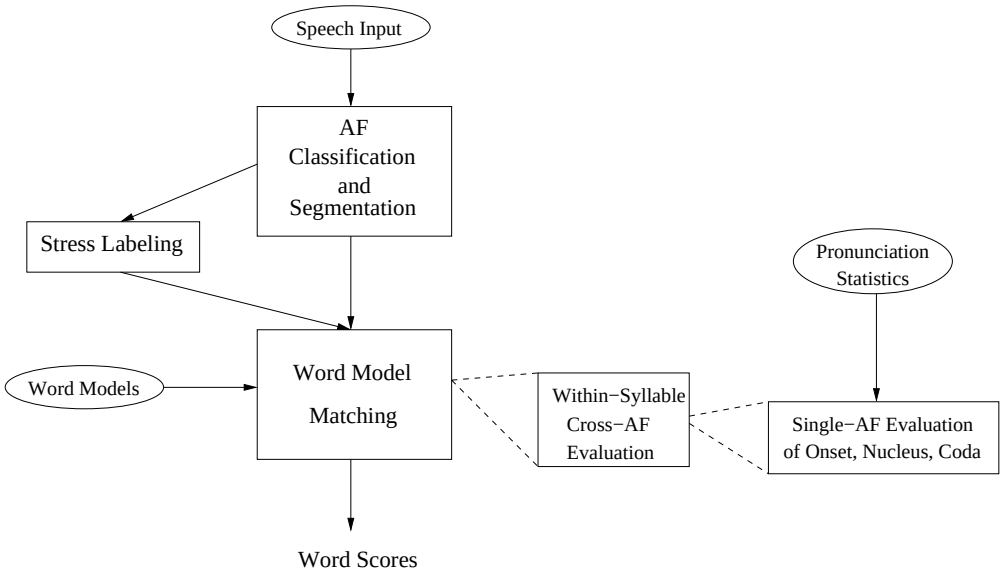

Figure 6.1: A very high-level overview of the recognition model. See text for detail.

modeling.

The syllables extracted are evaluated relative to syllable sequences of potentially matching lexical hypotheses. For each detected and hypothesized syllable pair, the evaluation is first performed along each AF dimension and the scores are then combined across the dimensions appropriately, taking into account the differential contributions of various AFs. For each AF dimension, the evaluation within a syllable is further decomposed into onset, nucleus and coda matching where the pronunciation variation patterns are interpreted according to the contextual information (summarized in syllable position and accent levels). A score is derived for each lexical hypothesis from the matching scores of its syllable constituents, and competing hypotheses are compared using these individual scores. This process is summarized in schematic form in Figure 6.1.

Because of the limited scope of this thesis, higher-level knowledge pertaining to the language model, and semantic and pragmatic processing is not explicitly considered in the multi-tier model, although they are extremely important in the recognition (and understanding) of speech. The multi-tier model is unlikely to capture all the variability of spontaneous speech; moreover, high-level processing is crucial for providing additional evidence to reduce confusibility and to facilitate efficient search among alternative hypotheses. Such high-level information may in fact significantly modify how the lower-level processing is performed.

The overall recognition process can be viewed as a fusion of heterogeneous information sources, distributed across time and space. The utility for recognition of different information sources is sufficiently variable both temporally and spatially, and hence, not all information sources should be given equal weight at each time frame. Differential treatment of the information sources relative to their contributions to the recognition task is likely to

be the most efficient and effective method for dealing with such variable information. Moreover, the appropriate information-fusion scheme should be adaptable to acoustic conditions, speaking style and other factors influencing the interpretation of the speech signal.

6.2 Questions Regarding the Multi-tier Model

The high-level description of the multi-tier model raises many questions pertaining to both its feasibility and functionality relative to conventional phonetic-segment-based models, and suggests directions for improvement. The discussion in previous chapters provided some motivation for the multi-tier model from different perspectives, such as articulatory-acoustic features, syllables and stress accent, with a special focus on capturing pronunciation variation phenomena in spontaneous speech. It is important to understand how the different components of the multi-tier model interact to produce recognition results and what level of accuracy is required at each level to achieve optimal performance.

The syllable is adopted as the binding unit across linguistic levels in the multi-tier model because of its stability and its systematic relationship to both the lower-tier units such as AFs and the supra-segmental features pertaining to stress accent. It is, therefore, of interest to ascertain how inaccuracies in syllable detection and segmentation affect recognition performance, how information at different syllable positions contributes to recognition performance, and how syllable-level information interacts with stress accent to affect the interpretation of AF classification.

As described in the previous chapter, stress accent plays a significant role in AF distribution and pronunciation variation of spontaneous speech. It is of interest to ascertain the efficacy of stress-accent modeling in recognition, especially its utility in pronunciation modeling.

Articulatory-acoustic features are the basic building block of the multi-tier model; their manifestation depends on the specific configuration of syllable and stress accent. In order to optimally use the information at hand, it is very useful to know the relative contributions of various AF dimensions to reocognition performance, both individually and in combination with other features. It is also of interest to know whether parsimonious pronunciation-variation modeling using AFs is indeed effective for recognition of spontaneous speech.

To answer such questions, experiments with real data are required. In the following section a simple test-bed system is described based on the multi-tier model for performing preliminary experiments on a highly constrained task. The results of the experiments are analyzed in the following chapter and initial answers to the questions posed above are discussed based on this analysis.

6.3 Test-bed System Implementation

This section describes the development of a simple test-bed system for evaluating the multi-tier model. The purpose of the implementation is to demonstrate the functionality

of the multi-tier model, as well as provide some preliminary answers to the questions posed in the previous section. The test-bed implementation is not meant to be a scalable, powerful performance system capable of fully automatic recognition on large, unconstrained tasks. Rather, the implementation is intended to be simple and transparent, easy to control and manipulate with both real and fabricated data in a highly constrained task domain. The system closely follows the multi-tier model, but is necessarily simplified in those instances where implementation is difficult with currently available resources, and in some cases in order to maintain a transparent system structure. Efforts have been made wherever possible to make the system's components modular and intuitive in order to permit a systematic dissection of functionality and performance. Limited learning capability is built-in wherever convenient to enable data-driven analysis, but not to an extent that would obscure the transparency and interpretation of the system structure and functionality.

The experiments were performed on the OGI Numbers95 corpus with word segment boundaries [12][1] (also cf. 3.2.2). This corpus contains the numerical portion (mostly street addresses, phone numbers and zip codes) of thousands of spontaneous telephone conversations (cf. Table 7.2 for a list of vocabulary), partitioned into different utterances of between one and ten words with an average of 3.9 words per utterance. An example of a typical utterance in this corpus is "*nine hundred forty six.*" The speakers contained in the corpus are of both genders and represent a wide range of dialect regions and age groups. The training data set contains ca. 2.5 hours of material with a separate 15-minute cross-validation set. The test data set contains ca. 1 hour of material.

The isolated nature and the small vocabulary of the recognition material make it relatively easy to implement and perform controlled experiments, thereby eliminating the necessity (and the effect) of language modeling and elaborate search algorithms (as well as associated pruning techniques). Performance is likely to be sufficiently good so as to accommodate meaningful empirical analysis. However, the task is not a trivial one; the corpus preserves many characteristics observed in large-vocabulary spontaneous speech corpora, particularly much of the pronunciation variation patterns described in the previous chapter. In the past several dissertation projects (e.g. [74][144][88]) at ICSI have successfully used the Numbers95 corpus for developing novel algorithms for various components of ASR systems.

This section first provides a high-level overview of the test-bed system, with reference to the description of the multi-tier model in the previous section. Various components of the system are then described in detail, with particular reference to the Numbers95 corpus.

6.3.1 Overview

The overall system implementation largely follows the model description in the previous sections. A flow-diagram of the test-bed system is given in Figures 6.2 and 6.3. The inputs to the system include the raw speech pressure waveform, (word-boundary infor-

[1]The original OGI Numbers95 corpus distribution does not include word boundary information. The current word boundaries are derived from the manual phonetic transcripts using Viterbi-alignment with a multiple-pronunciation lexicon developed by Dan Gildea at ICSI; the results were verified manually.

mation is only applied at a subsequent stage of hypothesis evaluation), a lexicon of syllable-AF-stress-accent-based canonical word models, and a set of transformation statistics from canonical to realized forms for various AFs derived from training data. Additionally, word-boundary information is made available as input to the hypothesis evaluation stage.

The front-end feature processing computes the log-compressed critical-band energy features as described in Section 3.2.1, their deltas and their double-deltas (here deltas and double-deltas refer to the first and second temporal derivatives, respectively). AF classification is based on the MLP-based procedure described in Chapter 3.2.1 for a number of AF dimensions (manner-specific or manner-independent). The classification results of the manner-of-articulation dimension are used to perform a manner-based segmentation with a Viterbi-like forward dynamic programming procedure. This procedure partitions the speech utterance into coherent regions of manner features, which to a great extent are co-terminous with the phonetic segments, as described in Section 4.4. The identified vocalic segments are classified as the nuclei of syllables. For each syllable detected, the manner-partitioned segments (especially the vowels) are used for automatic estimation of stress-accent levels using the Switchboard-corpus-trained AutoSAL system as described in Chapter 5. For the sake of simplicity, other AF dimensions are synchronized to manner segmentation, and for every segment a classification score is summarized for each feature along each AF dimension.

Word-boundary information is applied to delimit the AF information within each word segment to compare against the canonical word models in the lexicon. For each word hypothesis, a variety of syllable-sequence alignments with the inputs are considered and a matching score is obtained based on the matching results at the syllable level (including the possibility of syllable insertion and deletion). To evaluate each pair of reference and hypothesis syllables, the matching scores across various AF dimensions are combined using a fuzzy-measure and fuzzy-integral-based technique, described in detail in Section 6.3.5. This technique was chosen (1) for its ability to model the importance of and the interactions among the AF dimension scores, (2) for its interpretability and (3) for the lack of simplifying assumptions. Within each syllable, the computation of the score for each AF dimension is further decomposed into matchings at onset, nucleus and coda positions within the syllable. At this stage of processing, pronunciation variation modeling of the AFs with respect to syllable position and stress accent are considered through the use of statistics collected from training data, in the form of a feature-transformation matrix that transforms the canonical form to the realized (transcribed) form.

Once all hypotheses are evaluated, the results are returned and the best matching hypothesis (or hypotheses) identified. In the following sections each system component is described in detail, and training methods for certain components are also presented.

6.3.2 AF Classification and Segmentation

The front-end feature processing and AF classification adopts the procedures described in Chapter 3. The input speech signal is represented as a sequence of log-compressed critical-band energy features every 10 ms (a frame) over a 25-ms window, along with their first and second time derivatives (deltas and double-deltas). For each of the AF dimensions considered, an MLP is trained to perform feature classification at each frame such that the

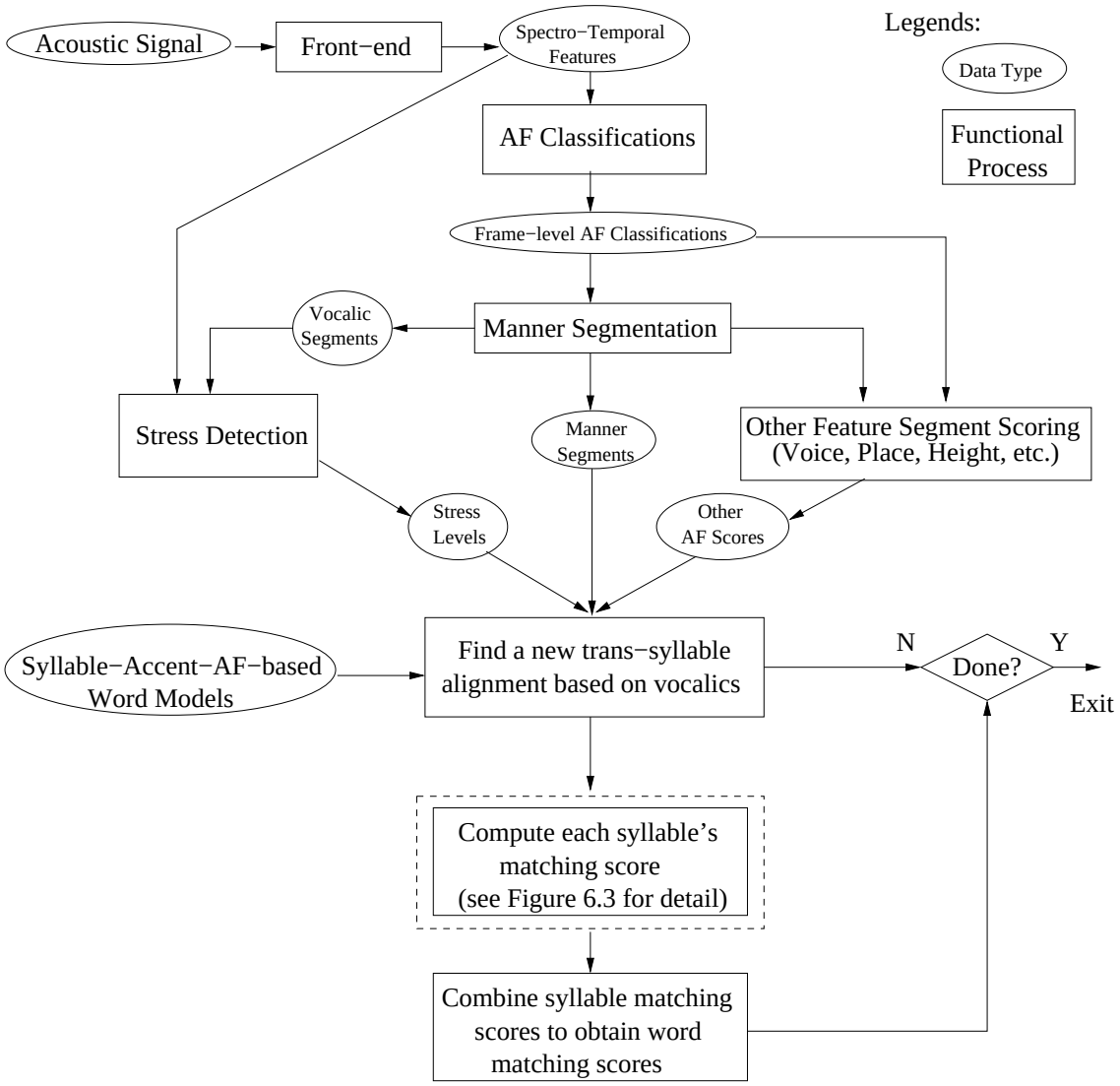

Figure 6.2: A flow diagram of the test-bed implementation of the proposed multi-tier model. Details of the process within the dashed box are in Figure 6.3.

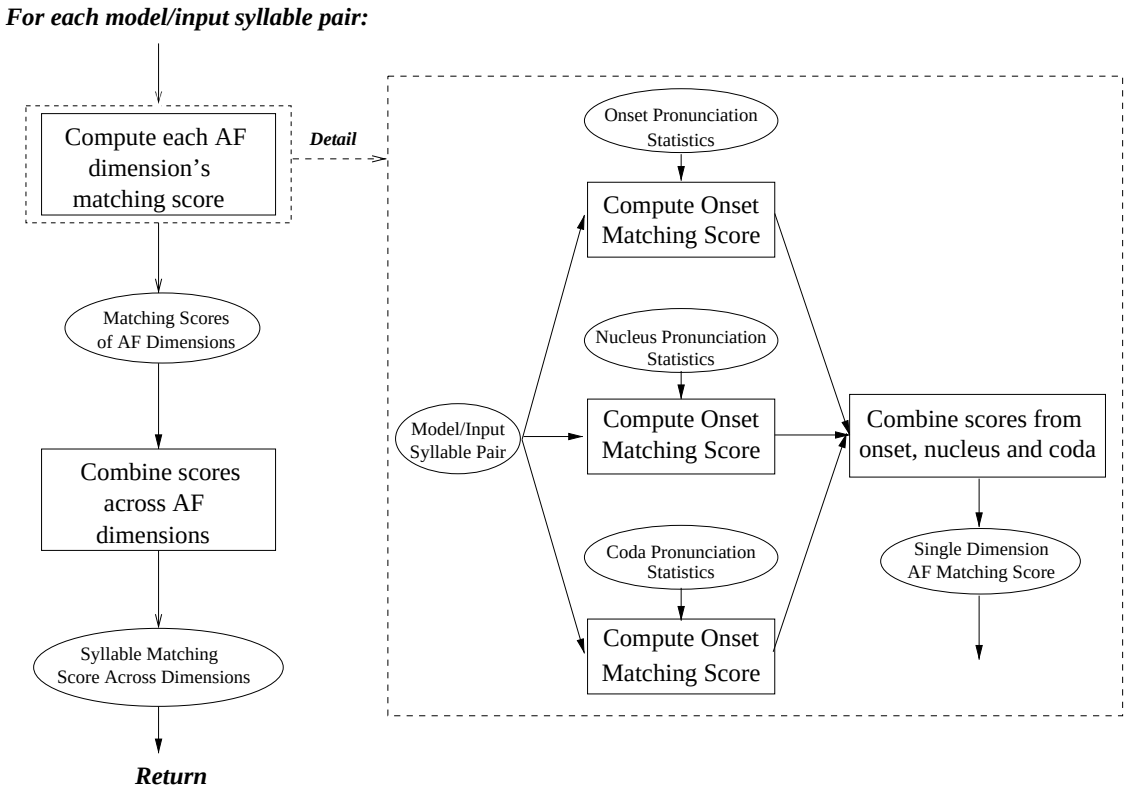

Figure 6.3: Details of computing the syllable matching score (as the process within the dashed box in Figure 6.2 – a flow diagram of the test-bed implementation of the multi-tier model). Legends are the same as that in Figure 6.2.

outputs represent the posterior probability estimates of each feature along the AF dimension. The AF dimensions include manner of articulation, place of articulation and voicing for manner-independent classifications. Manner-specific classifications are performed for place of articulation (partitioned into stops, fricatives, approximants, etc.) and vocalic-only feature dimensions, such as lip-rounding, vowel height, spectral dynamics (monophthong vs. diphthong) and tenseness (tense vs. lax). Thus, the AF classification output at each time frame is a set of vectors, each for a particular AF dimension consisting of posterior-probability estimates of constituent features.

Manner of articulation is a key dimension, whose classification results are segmented into roughly homogeneous manner segments using a Viterbi-like decoder, where the "lexicon" consists of manner features (e.g. vocalic, stop, fricative, etc.). This amounts to smoothing and integrating frame-level classification outputs in adjacent frames to identify segments of coherent manner features. Although it is certainly possible to produce different segmentations, for simplicity the current implementation only considers a single manner segmentation that tends to maintain a relative balance between the numbers of deleted and inserted segments. This simplification is controlled in the experiments to be described by comparing recognition results obtained from the same initial manner segmentation (whether it is automatically computed or derived from transcripts in fabricated data conditions).

The identified vocalic segments are interpreted as the vocalic nuclei of syllables; however, the syllable boundaries are not explicitly defined, so the syllable positions of consonantal segments are not specified at this stage. In real speech the AF dimensions are not necessarily synchronized with each other and may overlap. This feature asynchrony has been explored in various models of articulatory-feature-based speech processing (e.g. [129][10]). However, for simplicity's sake, the current implementation assumes that segmentation in various AF dimensions are synchronized to that in the manner dimension. This simplification is not without merit. As discussed in Section 4.4, manner-of-articulation segments are largely co-terminous with traditionally defined phonetic segments, so that synchronization to manner segmentation means that other AF dimensions are coordinated with phonetic-segment intervals. It should be noted that this is not to equate the implementation with the phonetic-segment-based systems, since each AF dimension in the model has the flexibility of independently evolving and the segments are explicitly tied to a syllable representation.

With this form of segmentation a summarized score is derived for each feature along each AF dimension for every segment[2]. Thus, each segment contains a vector of feature "confidence scores" along each AF dimension (cf. Table 6.1 for an example). Compared to the frame-level MLP outputs the total amount of information retained at this stage is greatly reduced but what remains should be the essential information germane to word recognition.

6.3.3 Stress-accent Estimation

Identification of the vocalic nuclei provides a means of estimating stress-accent level using the AutoSAL procedure described in the previous chapter. In the current im-

[2]In the current implementation this is obtained by simply averaging the posterior probabilities of the frames within the segment.

plementation, MLP networks trained on the 45-minute subset of the Switchboard corpus with manual stress-accent labels (cf. Section 5.3) were used as there is no manually labeled stress-accent material for the Numbers95 corpus. The set of features used on the Numbers95 corpus include the duration and normalized energy of the vocalic nucleus (i.e. features 10 and 12 in Table 5.7), the average of the vocalic segment spectra and the associated deltas and double-deltas (feature 13), as well as the ratio of the vocalic nucleus and syllable duration (feature 2). This four-feature set (equivalent to set 40 in Table 5.7) is able to achieve a near-optimal stress-accent labeling performance on the Switchboard corpus (cf. Figure 5.20).

To adopt the Switchboard-trained networks to the Numbers95 corpus, a few modifications in the feature-generating procedure were required. The time-constant for energy normalization was reduced from three seconds (Switchboard) to 600 milliseconds (Numbers95) in order to reduce the effect of pauses and silence period added to the beginning and ending of utterances. Since the exact syllable segmentation is not determined at this stage, the ratio between the vocalic segment duration and the syllable duration must use an estimated syllable duration. For this purpose, the sum of the durations of the vocalic nucleus, the preceding consonantal segment and the following consonantal segment is taken to approximate the syllable duration. Interestingly, an experiment on the Switchboard corpus shows that this approximation performs almost as well as using the manual-transcription-derived syllable duration.

With the above modifications, all the features used in AutoSAL can be automatically derived for the Numbers95 corpus. Manual inspection of selected utterances indicates that AutoSAL performs reasonably well on the Numbers95 corpus despite having been trained on a different corpus.

6.3.4 Word Hypothesis Evaluation

Once AFs and stress-accent material is processed the system is ready to evaluate specific word hypotheses. Table 6.1 shows a typical example of the information contained in the processed input data associated with a preliminary word segmentation. The data are listed in the order of the manner segments where the positions within the utterance ("start" and "end") are indicated in (10-ms) frames. Each segment contains scores associated with different features along each of several AF dimensions. In particular, there is a manner-specific place (MS-place) dimension determined by the initial manner segment label, as well as a manner-independent place (MI-place) dimension that corresponds to the overall manner-independent place classification results. Some AF dimensions, such as "height," are only included for the relevant vocalic segments. The stress-accent label associated with a syllable is attached to the vocalic nucleus for convenience of display although it is in fact a property of the entire syllable.

As described in the multi-tier model description, the word models contain only canonical baseform pronunciations for each word in the lexicon. In the Numbers95 corpus experiments, only a single model is included for each word, corresponding to the most popular pronunciation as extracted from the training set. A typical example of a word model (for the word "six") is shown in Table 6.2. Each word model is described by its

ID=4728zi		start=116		end=158		segments=4
1	start=116		end=129		manner=fricative	
	manner	voc=.016	nas=.048	stp=.003	fri=.901	apr=.000
	voice	voi=.278	unv=.722			
	MS-place	lab=.074	den=.004	alv=.921	glo=.000	
	MI-place	den=.007	lab=.070	cor=.832	ret=.000	vel=.000
		glo=.000	frt=.006	cen=.003	bak=.007	sil=.075
2	start=130		end=135		manner=vocalic	
	manner	voc=.963	nas=.000	stp=.003	fri=.032	apr=.001
	voice	voi=.929	unv=.071			
	MS-place	frt=.997	cen=.001	bak=.003		
	MI-place	den=.000	lab=.000	cor=.017	ret=.000	vel=.001
		glo=.000	frt=.979	cen=.001	bak=.002	sil=.000
	height	low=.001	mid=.008	hi=.991		
	round	rnd=.026	unr=.974			
	static	sta=.976	dyn=.024			
	tense	ten=.026	lax=.974			
	stress	str=.650				
3	start=136		end=143		manner=stop	
	manner	voc=.005	nas=.000	stp=.991	fri=.002	apr=.000
	voice	voi=.111	unv=.889			
	MS-place	alv=.000	vel=1.00			
	MI-place	den=.000	lab=.000	cor=.003	ret=.000	vel=.993
		glo=.000	frt=.003	cen=.000	bak=.000	sil=.000
4	start=144		end=157		manner=fricative	
	manner	voc=.001	nas=.000	stp=.133	fri=.809	apr=.016
	voice	voi=.018	unv=.982			
	MS-place	lab=.345	den=.005	alv=.649	glo=.000	
	MI-place	den=.005	lab=.193	cor=.585	ret=.001	vel=.131
		glo=.000	frt=.000	cen=.000	bak=.051	sil=.035

Table 6.1: A typical example of the information contained in the processed input data associated with a word segment after the initial AF classification, manner-based segmentation and automatic stress-accent labeling. The word in this example is "*six*" ([s ih k s]). "ID" is the Numbers95 utterance ID; "MS-place" is manner-specific place; "MI-place" is manner-independent place. "Start," "end" and "segments" are specified in terms of frames (25-ms window with a 10-ms sliding step).

Word-label: six			Syllables: 1		
Stress: 0.75					
Onset	manner=fri	voice=unv	place=alv		
	duration mean=14		SD=5.0		
Nucleus	manner=voc	voice=voi	place=frt	height=hi	round=unr
	static=sta	tense=lax	duration mean=8		SD=3.7
Coda	manner=stp	voice=unv	place=vel		
	duration mean=9		SD=3.2		
	manner=fri	voice=unv	place=alv		
	duration mean=12		SD=5.9		

Table 6.2: An example of a word model (for the word "six"). The stress-accent level is derived from the mode of stress-accent levels for the instances of "six" found in the training data. The duration mean and standard deviation ("SD") are specified in terms of frames (25-ms window with a 10-ms sliding step).

syllable constituents and associated typical stress-accent level (using the mode of various instances). Each syllable contains AF specifications for each of its onset, nucleus and coda segments, as well as the mean duration and the corresponding standard deviation (in 10-ms frames).

Because the number of syllables detected in the input signal may not precisely match that of the reference word model, an alignment (at the syllable level) between the reference and hypothesis is required in order to evaluate a specific word hypothesis. Each alignment considers the possibility of insertion and deletion of syllables, and penalties are assigned to those occurrences. It turns out that, for experiments on Numbers95 corpus, the overall word recognition rate is relatively insensitive to the magnitude of the penalties assigned to the insertions and deletions over a broad dynamic range. Therefore, in the experiments to be described, the syllable insertion and deletion penalties adopt fixed values determined on a cross-validation data set.

For each syllable alignment the total penalty score from the insertion, deletion and substitution of syllables are computed and the word hypothesis takes the minimum penalty score over all alignments. This alignment requires relatively little computation using a dynamic-programming procedure. The bulk of the computation for evaluating word hypothesis lies in the evaluation of matching syllable pairs as described in the following sections.

6.3.5 Cross-AF-dimension Syllable-score Combination

For each matching pair of syllables a score is computed indicating the confidence of the input syllable matching the reference syllable. To obtain this score, a separate score is first computed for each AF dimension (the details of which are described in the next section), and a multiple-information-aggregation approach is adopted to combine the scores from the various AF dimensions.

As previously discussed, different AF dimensions contribute differently to the recognition task and therefore should be given differential weights in the information combining process. Moreover, the various AF dimensions are not truly orthogonal, and there exists significant coupling among them. Thus, a simple linear combination of the scores cannot truly capture the redundancy and synergy of the information contained in subsets of the AF dimensions; a highly non-linear process is required. Since the relationship among the various AF dimensions can be quite complex, a suitable information aggregation method should be flexible in taking information from heterogeneous sources without pre-specification of their inter-relationship. On the other hand, good interpretability of the combining method is desired, especially for the diagnostic experiments where we would like to ascertain the importance and interactions of various AF dimensions to the combined decision, and this would also help in selecting appropriate features to model and in devising effective adaptation methods. The current implementation adopts a fuzzy-measure/fuzzy-integral-based, multiple-information-aggregation method that possesses a number of properties suitable for this task.

Fuzzy Measures

The concept of fuzzy measure was introduced by Sugeno [128][134] in the early seventies in order to extend the classical (probability) measure by relaxing the additivity property. A formal definition of the fuzzy measure is as follows:

Definition 1 *Fuzzy measure: Let X be a non-empty finite set and Ω a Boolean algebra (i.e. a family of subsets of X closed under union and complementation, including the empty set) defined on X. A fuzzy measure, g, is a set function $g : \Omega \rightarrow [0,1]$ defined on Ω, which satisfies the following properties:*

- *Boundary conditions: $g(\phi) = 0$, $g(X) = 1$.*
- *Monotonicity: If $A \subseteq B$, then $g(A) \leq g(B)$.*
- *Continuity: If $F_n \in \Omega$ for $1 \leq n < \infty$ and the sequence $\{F_n\}$ is monotonic (in the sense of inclusion), then $\lim_{n\rightarrow\infty} g(F_n) = g(\lim_{n\rightarrow\infty} F_n)$.*

(X, Ω, g) is said to be a fuzzy measure space.

This definition of a fuzzy measure differs from that of a probability measure only in terms of the monotonicity property. For example, by additivity, the probability measure of the union of two disjoint subsets of X must be equal to the sum of the probability measures of the two subsets; in contrast, the fuzzy measure of the union can be smaller or larger than the sum of the fuzzy measures of the two subsets as long as the monotonicity property holds – a crucial feature that accommodates synergy and redundancy among information sources. Since additivity is a special case of monotonicity, *the probability measure is, in fact, a special case of a fuzzy measure.* Other special cases of fuzzy measures include the possibility measure [148] and the belief functions of the Dempster-Shafer theory [115], etc.

For the present task, let $X = x_1, \cdots, x_N$ represent the set of N AF dimensions under consideration and the fuzzy measure, g, represents the contribution of each subset of X (i.e. a set of some AF dimensions, including singleton sets) in evaluating the match between a reference syllable and the input. In many situations, it is useful to ascertain the contribution of a particular AF dimension in the entire evaluation process. However, since each AF dimension is involved in many subsets of X, the contribution of the AF dimension cannot be easily read from the fuzzy measures. A concept from cooperative game theory, the Shapley score [116][46], can be applied here to help in the interpretation.

Definition 2 *Shapley score: Let g be a fuzzy measure on X. Shapley score for every $i \in X$ is defined by*

$$v_i \equiv \sum_{K \subset X \setminus \{i\}} \frac{(|X| - |K| - 1)!|K|!}{|X|!}[g(K \cup \{i\}) - g(K)] \tag{6.1}$$

where $|X|$ and $|K|$ are the cardinality of X and K, respectively.

Intuitively, a Shapley score computes the additional value that i brings to various subsets of $X \setminus \{i\}$ (the set X excluding the i and including the empty set), normalized appropriately (cf. [116] for a derivation of Equation 6.1). A Shapley score, v_i, can be interpreted as an average value of the contribution that information source, i *alone*, provides in all different combinations of information sources and it can be verified that Shapley scores sum to $g(X) = 1$. This concept has also been extended to computing the interaction of a pair of information sources [93].

Definition 3 *Two-way interaction index: Let g be a fuzzy measure on X. The two-way interaction index of elements $i, j \in X$ is defined by*

$$I_{ij} \equiv \sum_{K \subset X \setminus \{i,j\}} \frac{(|X| - |K| - 2)!|K|!}{(|X| - 1)!}[g(K \cup \{i,j\}) - g(K \cup \{j\}) - g(K \cup \{i\}) + g(K)] \tag{6.2}$$

where $|X|$ and $|K|$ are the cardinality of X and K, respectively.

The interaction index, I_{ij}, provides an indication of the interaction between the pair of information sources i and j. When $I_{ij} < 0$, there exists a negative interaction (*redundancy*) between information sources, i and j, in that the value of the pair i and j is less than the sum of the values of i alone and j alone when they are included into sets of information sources. On the other hand, if $I_{ij} > 0$, there exists a positive interaction (*synergy*) between i and j in that the value of the pair i and j exceeds the sum of the values of i alone and j alone when included in sets of information sources. In cases where $I_{ij} = 0$, the value gained by a set of information sources from including the pair i and j is just equal to the sum of the gains from i alone and j alone, and thus there is no interaction between the pair. This definition has been further extended to the interaction of any subset of X by Grabisch [45]:

Definition 4 *Interaction index: Let g be a fuzzy measure on X. The interaction index of any subset $A \subset X$ is defined by*

$$I(A) \equiv \sum_{B \subset X \setminus A} \frac{(|X| - |B| - |A|)!|B|!}{(|X| - |A| + 1)!} \sum_{C \subset A} (-1)^{|A \setminus C|} g(C \cup B) \tag{6.3}$$

.

The concept of the Shapley score and the interaction index makes it much easier to interpret the importance and contribution of various information sources and can also be used for the task of feature selection where a fuzzy measure is used. For example, if a certain information source has a small Shapley score and mostly negative interactions with other sources, it may be safely removed from consideration without significant impact on recognition performance.

Fuzzy Integral

To combine scores obtained from various information sources with respect to some fuzzy measure a technique based on the concept of fuzzy integral can be adopted. There is actually more than one kind of fuzzy integral [96]; the one adopted here is the Choquet integral proposed by Murofushi and Sugeno [94][3].

Definition 5 *(Choquet) Fuzzy integral: Let (X, Ω, g) be a fuzzy measure space, with $X = \{x_1, \cdots, x_N\}$. Let $h : X \to [0, 1]$ be a measurable function. Assume without loss of generality that $0 \leq h(x_1) \leq \cdots \leq h(x_N) \leq 1$, and $A_i = \{x_i, x_{i+1}, \cdots, x_N\}$. The Choquet integral of h with respect to the fuzzy measure g is defined by:*

$$\int_C h \circ g = \sum_{i=1}^{N} [h(x_i) - h(x_{i-1})] g(A_i) \tag{6.4}$$

where $h(x_0) = 0$. Or equivalently,

$$\int_C h \circ g = \sum_{i=1}^{N} h(x_i)[g_i^N - g_{i+1}^N] \tag{6.5}$$

where $g_i^j = g(\{x_i, x_{i+1}, \cdots, x_j\}), i \leq j$ and 0 otherwise.

An interesting property of the (Choquet) fuzzy integral is that if g is a probability measure, the fuzzy integral is equivalent to the classical Lebesgue integral [32] and simply computes the expectation of h with respect to g in the usual probability framework. The fuzzy integral is a kind of averaging operator in the sense that the value of a fuzzy integral is between the minimum and maximum values of the h function to be integrated. A number of commonly used aggregation operators are special cases of the fuzzy integral [47][43] – for example, the min and max operators, the weighted sum and the ordered weighted

[3] Other kinds of fuzzy integrals include the Sugeno integral, t-conorm integrals, etc. [96]

average. A distinct advantage of the fuzzy integral as a weighted operator is that, using an appropriate fuzzy measure, the weights represent not only the importance of individual information sources but also the interactions (redundancy and synergy) among any subset of the sources. The fuzzy measures and fuzzy integrals have been applied successfully to a number of multi-criteria decision tasks [43][48][104][95] and multiple information aggregation for classifications [47][46][71][19][105].

Learning Fuzzy Measures

The introduction to fuzzy measures and fuzzy integrals above describes how they are used to combine scores obtained along each AF dimension into an aggregated matching score for a syllable. The intuitive interpretation of a fuzzy measure allows for the specification of fuzzy-measure parameters based on knowledge of the importance and interactions of AF dimensions. However, in practice, it is more useful to be able to learn the fuzzy-measure parameters from data in order to discover the contribution patterns automatically.

During training the system has knowledge of which reference syllable in a pool of canonical syllable forms in the lexicon best matches the input syllable. Such information may be used to set up a supervised, discriminant training scheme for learning the fuzzy measures from data by viewing the evaluation of the fuzzy integrals as a part of a syllable classification task. Different algorithms have been proposed for learning fuzzy measures using linear and quadratic programming approaches [47], heuristic measure updating rules based on class confusions [105], as well as gradient-based methods [44]. The current system adopts a gradient-based framework [44] for its efficiency and scalability. Algorithms based on two different error criteria (minimum-squared-error and minimum-cross-entropy) have been derived for the current task within the gradient-based framework. The algorithm and derivations are described in Appendix C.

6.3.6 Within-syllable Single-AF-dimension Matching

Within each syllable, the matching score for a particular AF dimension is first computed for each of the onset, nucleus and coda positions and then averaged. For the vocalic nucleus, the computation of the matching score is straightforward since for each pair of matched syllables there is a matched pair of vocalic nuclei in the reference and the input syllables. At this stage the evaluation of the nucleus matching score considers the AF deviation patterns from the canonical form in order to accommodate more pronunciation variations than that provided by the canonical base-forms in the lexicon. This process is roughly as follows. Let C denote the model representation (the canonical base-form in the lexicon) of the segment along the AF dimension of interest, T, the actually realized form as would be found in a manual transcript, and X, the acoustics. The AF matching score is thus an evaluation of the model in the lexicon (the canonical base-form) based on the acoustic input (for example, the posterior probability $P(C|X)$). Expanding this term we obtain:

$$P(C|X) = \sum_{T} P(C,T|X) \tag{6.6}$$

	Transcribed		
Canonical	Front	Central	Back
Front	83.6	8.1	0.4
Central	16.2	86.9	0.4
Back	0.2	5.0	99.1

Table 6.3: Example of transformation statistics from canonical to transcribed vocalic place (front, central and back) for nuclei of unaccented syllables, derived from the Numbers95 corpus training set. Note that all numbers are in terms of percentage of the *Transcribed* features as required by the formulation $P(C|T,S)$ (see text for detail).

$$= \sum_T P(C|T,X)P(T|X) \tag{6.7}$$

$$\approx \sum_T P(C|T,S)P(T|X) \tag{6.8}$$

where S represents some summary statistics of X. The approximation in Equation 6.8 makes the assumption that S captures most of the important information contained in X that affects the realization T of the canonical form in the model representation C. In the current implementation, S includes information pertaining to syllable position and stress-accent level of the containing syllable. As discussed in the previous chapters, these two pieces of information are both essential factors to pronunciation variation and would likely capture most of the relevant information. Given transcribed training data, the $P(C|T,S)$ can be estimated by simply counting the statistics of the transformation from canonical to realized forms (cf. Table 6.3 for an example) for different S (i.e.. different syllable position and stress-accent level combinations). The term $P(T|X)$ is estimated from the acoustic modeling (e.g. the AF classification outputs). Thus, an AF matching score for each canonical base-form ($P(C|X)$) is essentially a weighted average of the AF matching scores of transcribed forms ($P(T|X)$), where the associated weights ($P(C|T,S)$ are determined by the context.

It should noted that this process only considers a limited range of pronunciation variations. For example, it does not consider the joint variation patterns of adjacent segments. However, the multi-tier model does not preclude the modeling of more complicated pronunciation variation phenomena. The simplified modeling described above only reflects the scope of the current implementation.

The evaluation of the onset and coda is a little more complicated since there may be insertions and deletions of segments, and the boundary between two adjacent syllables within a polysyllabic word is not precisely known. To address this problem, different alignments of the elements in the reference onset or coda are considered in conjunction with the candidate consonantal segments in the input, and the alignment resulting in the fewest mismatches is chosen. The penalty for mismatches considers deletions and insertions, as well as substitutions. The substitution score is computed in the same fashion as for the nucleus evaluation above and possible AF variations are taken into account with respect to syllable position and stress accent. The deletion penalty is handled similarly by using the

deletion statistics of the canonical AF segment (again, partitioned by syllable position and stress-accent level) computed from the training data. For example, a canonical AF segment with a particular syllable position and stress accent that is more likely to be deleted receives a lower deletion penalty. The insertion penalty depends on two factors heuristically – the duration of the inserted AF segment and its distance from the vocalic nucleus. Both of these factors are modeled as an exponential function in the form $d = 1 - \exp(-w \cdot \mu)$ where d is the insertion penalty and $w \geq 0$ a scaling parameter; μ is either the duration of the inserted segment or the (reciprocal of the) distance between the inserted segment and the vocalic nucleus. Essentially, longer durations of inserted segments and shorter distances from the vocalic nucleus yield larger insertion penalties. The parameters involved in these heuristic penalty functions are tuned to the training data.

6.4 Summary

Evidence presented in previous chapters suggests that there are significant advantages in incorporating articulatory-acoustic-feature and stress-accent information in a syllable-centric representation of speech, particularly for efficient modeling of pronunciation variation phenomena in spontaneous speech. Based on such evidence the current chapter proposed a multi-tier model of speech recognition as an alternative to conventional phone-based models. Some key characteristics of the multi-tier model are:

- Speech is organized as a sequence of syllables, containing vocalic nuclei, and optionally, onset and coda segments.

- Each syllable carries a certain level of accent (or lack of accent), which can be coarse but capable of distinguishing among completely unaccented, fully accented and intermediately accented syllables.

- Onset, nucleus and coda segments of each syllable are described by features along several quasi-orthogonal, articulatory-acoustic feature dimensions, such as manner of articulation, place of articulation, voicing, lip-rounding, etc.

- The lexicon contains only a single (or a small number of) canonical baseform representation of words (or short phrases), each containing one or more syllables, described by the canonical AF specification and accent levels.

- During recognition the coverage of pronunciation variants is expanded by taking into account statistical characterization of AF deviations from the canonical forms conditioned on a number of contextual cues, such as syllable position and stress accent.

- The overall recognition process is viewed as a fusion of heterogeneous information sources, distributed across time and space. The utility to recognition varies across information sources and differential weighting is applied according to their relative contribution and reliability of estimation.

Certain useful information sources, such as higher-level linguistic processing, have not been considered explicitly within the multi-tier model but only for simplicity of initial developments and not as a fundamental limitation.

To demonstrate the feasibility and functionality of the multi-tier model and to understand its limitations and directions for improvements, a test-bed implementation was developed to perform controlled experiments on a limited-vocabulary task. A detailed description was provided for key components of the test-bed implementation:

- An array of MLP neural networks are trained to classify each frame of pre-processed, log-compressed critical-band energy features along a number of AF dimensions.
- A Viterbi-like, dynamic-programming procedure is used to obtain a manner-based segmentation of the speech signal from the frame-level output of manner classification. The detected vocalic segments are interpreted as the vocalic nuclei of syllables. Segmentation of various AF dimensions are synchronized to the manner-based segmentation.
- An automatic stress-accent labeling system, trained on manually annotated stress-accent labels (on a subset of the Switchboard corpus), is used to provide stress-accent label for each syllable based on the manner-based segmentation; all input features are automatically derived from the speech signal.
- Matches between input and word hypotheses from the lexicon are evaluated based on syllable-level matching scores.
- For each (input and hypothesis) syllable pair, a matching score is computed along each AF dimension of interest. AF transformation statistics from the canonical to realized forms are computed from the training data and used to provide pronunciation variation information conditioned on syllable position and stress accent.
- A fuzzy-measure and fuzzy-integral-based multiple information aggregation technique is adopted to combine scores from various AF dimensions to form a single syllable matching score. The fuzzy-based technique captures the relative importance, as well as interaction patterns, among any subset of AF dimensions. A gradient-based, discriminant-training algorithm was developed (cf. Appendix C) to learn the fuzzy measures from the training data.

The following chapter will describe controlled experiments performed on the test-bed implementation as well as detailed analysis of the results.

Chapter 7

Multi-tier Recognition – Experiments and Analysis

The previous chapter described a model of speech recognition based on syllable, articulatory-acoustic-feature and stress-accent information. The model uses the syllable as the binding unit for various linguistic tiers of spoken language. Although it has many limitations, the proposed model potentially provides a viable direction for representing spontaneous speech in a parsimonious fashion. The proposed model raises a number of questions regarding its functionality and effectiveness. To obtain some preliminary answers a test-bed implementation was developed to perform experiments on a limited-vocabulary task.

This chapter describes a number of controlled experiments performed on the test-bed implementation using both real (i.e., automatically derived) and fabricated (derived from a manual transcript) data, and discusses the results addressing the questions posed in the previous chapter. As previously noted the implementation was designed to provide a simple, transparent test-bed for performing controlled experiments and analysis rather than to develop a scalable, high-performance system. Consequently, the processing in some of the system components was simplified. The experimental analysis and conclusions drawn from the results of this analysis must take this limitation into account. For example, no higher-level processing was incorporated, and only limited (but the most relevant) conditioning was considered for pronunciation variation modeling.

7.1 Experimental Conditions

As described in the previous chapter (cf. Section 6.3), the experiments were performed on materials from the OGI Numbers95 corpus [12] consisting of digits and numbers extracted from spontaneous telephone dialogues. In addition, word-boundary information was provided to the recognizer to reduce the complexity involved in hypothesis search and to restrict word-error forms to substitutions (for simplicity of analysis). The training data consisted of 3233 utterances (12510 words) from the training set and 357 utterances (1349

words) from a separate cross-validation set. Testing was performed on 1206 utterances (4669 words) from the development test set.

Because of several simplifying assumptions the system was not expected to achieve optimal performance. To gain insights into the sources of recognition error and to be able to focus on particular aspects of the multi-tier model, the experiments described in this section were performed not only on entirely automatically derived data, but also on fabricated data derived from manual transcription, and on a combination of automatic and fabricated data. Details of these three data conditions are described first and will be continually referred to throughout the subsequent description of the controlled experiments.

The first data condition, which will be referred to as the "baseline" in the remainder of this chapter, used all automatically derived data from the acoustic signal, including the classification and segmentation of various AF dimensions and estimation of stress-accent level. This condition establishes how far the current implementation is from optimum and assesses penalties associated with various simplifying assumptions.

The "fabricated" data condition used AF "classification" scores and segmentation derived from the manual phonetic transcription and a fixed phone-to-AF mapping (cf. Table 3.1). In this case fabricating AF "classification" results amounts to assigning, for each AF dimension, a maximum output for the reference feature of each frame and a minimum output for all other features. This condition should make the system perform nearly optimally, and hence, the system performance in this condition can serve as an upper-bound of the performance of the current implementation.

The third data condition, referred to as the "half-way house," used automatically computed AF-classification results but "fabricated" AF segmentation and the a vocalic/non-vocalic segment labeling, derived from manual transcription. This condition serves to assess the contributions of accurate AF segmentation and detection of vocalic nuclei for syllables.

Note that in all three conditions stress-accent estimation was computed using the AutoSAL system (adapted from the Switchboard corpus to OGI Numbers95 corpus). Unless otherwise noted, the AF dimensions considered in the experiments are (1) manner of articulation, (2) place of articulation (manner-independent and manner-specific), (3) voicing, (4) vocalic height, (5) lip-rounding, (6) spectral dynamics, (7) vowel tenseness (cf. Section 3.1 for a brief review of these feature dimensions).

7.2 Overall System Performance

The overall system performance is evaluated by computing word-error rates associated with each data condition (cf. Table 7.1). To obtain a better assessment of the performance, the error rates are computed with different tolerance (top-N-match for $N = 1, 2, 5$) where "top-1-match" refers to the criterion that the correct word must be the top-matched hypothesis to receive a "correct" score, and "top-N-match" (for $N > 1$) refers to that the correct word must be among the top N hypotheses to receive a "correct" score. For the baseline condition, 35% of the misrecognized words (top-1-match) are actually the second top match and 69% are among the second through fifth top matches. A similar pattern is also observed in the recognition results of the half-way house and fabricated data conditions.

	Word-Error Rate%		
Top-N-match	Baseline	Half-way House	Fabricated
1	5.59	1.97	1.29
2	3.64	1.46	1.05
5	1.76	0.86	0.47

Table 7.1: Overall word-error rates (percentage) on the development test set for the three data conditions. Three different levels of tolerance (the correct word being within top-1-match, top-2-match and top-5-match) are shown.

Under all three conditions, word-error rates of polysyllabic words are generally higher than that of the monosyllabic words, and this gap is especially large for the baseline condition in which syllable detection (the determination of vocalic nuclei) is much less accurate than the other two conditions (cf. Table 7.2 for the baseline accuracy of each word and the number of times it occurs in the test set). This pattern is very different than that observed on the Switchboard-corpus recognition output using conventional ASR systems (cf. Figure 2.5 in Chapter 2), where polysyllabic words usually exhibit better recognition performance than monosyllabic words. This difference reflects the explicit modeling of the syllable in the multi-tier model and the greater word-error rates associated with polysyllabic forms are partially an artifact of the uneven distribution of monosyllabic and polysyllabic words. Since ca. 79% word tokens in the data set are monosyllabic, the training procedure that optimizes for overall word accuracy trades off penalties for syllable insertion and deletion in favor of monosyllabic words (e.g., by assigning relatively lower penalties to syllable insertions relative to syllable deletions).

The overall word-error rate (top-1-match, cf. Table 7.1) of the half-way house condition (1.97%) is much closer to the fabricated data condition (1.29%) than to the baseline (5.59%). This suggests that the additional information incorporated into the half-way house condition is extremely important for recognition in the current implementation. As described in the previous section, this extra information pertains to AF segmentation, as well as to knowledge of whether a segment is vocalic (or not), which facilitates more accurate syllable detection. The low word-error rates for the fabricated and half-way house data conditions suggest that if the initial AF classification and segmentation can be performed accurately, the implementation based on the multi-tier model will be able to provide a reasonable performance on this constrained task. It also suggests that relying on a single initial segmentation may be too restrictive and no hard commitment should be made at the early stages of recognition.

From the trained fuzzy measures, Shapley scores (cf. Section 6.3.5) can be computed to express the contribution of each AF dimension to the overall recognition. Figure 7.1 shows the mean Shapley scores derived from the trained fuzzy measures of the baseline condition (average over 15 random trials, along with the range of ± 1 standard deviation). Recall that each Shapley score, v_i, represents an average value of the contribution that the ith AF dimension alone provides to the combined recognition score. This average value is computed by considering the contribution of this dimension occuring in all

Monosyllabic			Polysyllabic		
Word label	Count	Accuracy%	Word label	Count	Accuracy%
oh	373	97.6	zero	211	84.4
one	597	97.0	seven	349	92.0
two	495	99.0	eleven	15	60.0
three	420	98.1	thirteen	11	54.5
four	358	94.4	fourteen	12	83.3
five	394	98.0	fifteen	19	89.5
six	302	98.3	sixteen	9	88.9
eight	317	96.5	seventeen	9	77.8
nine	390	97.7	eighteen	7	85.7
ten	29	72.4	nineteen	9	88.9
twelve	11	0.0	twenty	82	79.3
			thirty	59	72.9
			forty	39	87.2
			fifty	41	78.0
			sixty	22	90.9
			seventy	23	87.0
			eighty	23	69.6
			ninety	23	82.6
			hundred	20	75.0
Total	3686	97.0	Total	983	84.8

Table 7.2: Overall word accuracy of the baseline data condition for each word with its number of occurrences in the development test set. The words are partitioned into monosyllabic and polysyllabic forms for comparison.

possible combinations of AF dimensions. A large value of Shapley score implies the associated AF dimension provides significant utility for recognition; however, a small value of Shapley score does not necessarily imply that the associated AF dimension is of no utility since it may still exhibit a significant positive interactions with other AF dimensions. The magnitude of the interactions can be ascertained by computing the interaction indices (cf. Section 6.3.5). Another reason for caution in interpreting Shapley scores and interaction indices is that they were derived from fuzzy measures trained on real AF dimension scores that often contained a certain amount of error. Therefore, the derived Shapley scores and interaction indices reflect not only the inherent contribution that individual AF dimensions provide to recognition, but also the level of reliability of the AF-dimension scores. For example, if a particular AF dimension is very important for recognition but its matching score estimate is unreliable, it would be very unlikely to be associated with a large Shapley score and interaction indices. From Figure 7.1 it can be observed that both manner and place of articulation have above-average Shapley scores, while lip-rounding and voicing have far below-average scores. This suggests that both manner and place dimensions provide significant contribution to recognition, while lip-rounding and voicing dimensions may not have as much utility, depending on how they interact with other AF dimensions. This observation intuitively makes sense, as the manner and place dimensions are expected to provide the most information pertaining to lexical access.

To gain further insights into the interaction of various AF dimensions in performing recognition, the interaction indices (cf. Section 6.3.5) can be computed for any subset of the AF dimensions from the trained fuzzy measures. Figure 7.2 shows the two-way interaction indices[1] computed from the trained fuzzy measures of the baseline condition. For each pair of AF dimensions, a square (below the minor diagonal) has the corresponding interaction index, color-coded such that red indicates positive interaction (synergy), blue indicates negative interaction (redundancy), and white indicates no interaction. For example, the positive interaction between voicing and manner (+0.096) suggests these two dimensions, when considered together, contribute more to recognition than when they are considered separately. Together with Shapley scores, the interaction indices can be used to infer the utility associated with each AF dimension. For example, the "lip-rounding" dimension has a relatively small Shapley score and mostly small or negative interactions with other AF dimensions, and thus may be removed from consideration without having a great impact on the system performance. On the other hand, although the voicing dimension has a small Shapley score, it may still have significant utility for recognition because of the large positive interaction between voicing and manner (as well as place).

7.3 Testing the Contribution of Stress Accent

In order to ascertain the contribution of stress-accent information, particularly its utility in capturing the pattern of AF deviations from canonical, several experiments were performed with and without conditioning the AF statistics on stress-accent level. For each of the three data conditions, Table 7.3 shows word-error rates associated with each of

[1] Higher-order interactions are also computed but they are more difficult to display and are omitted here.

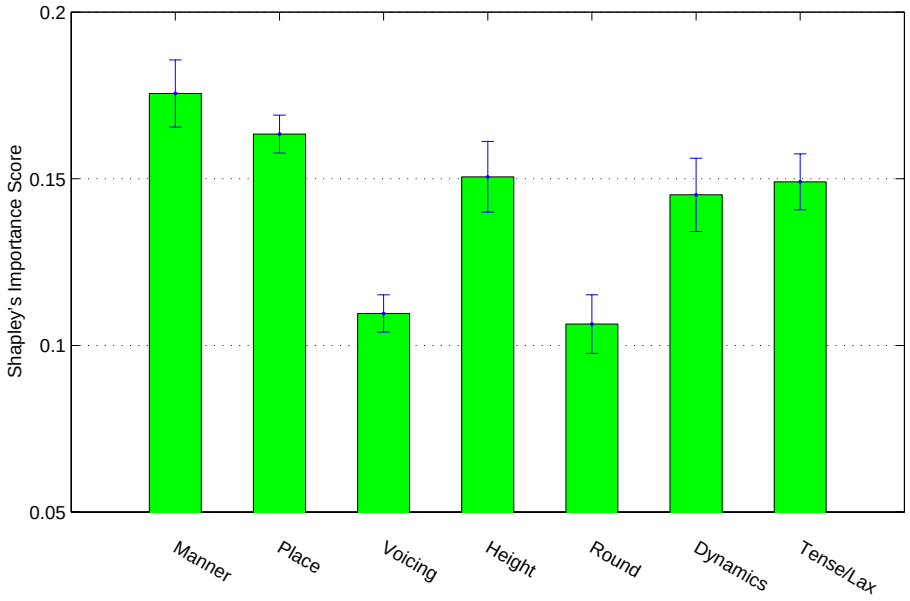

Figure 7.1: Mean Shapley scores computed from the trained fuzzy measures of the baseline condition for different AF dimensions, averaged over 15 random trials. The error-bar associated with each score indicates the range of ± 1 standard deviation. The mean scores sum to 1.0.

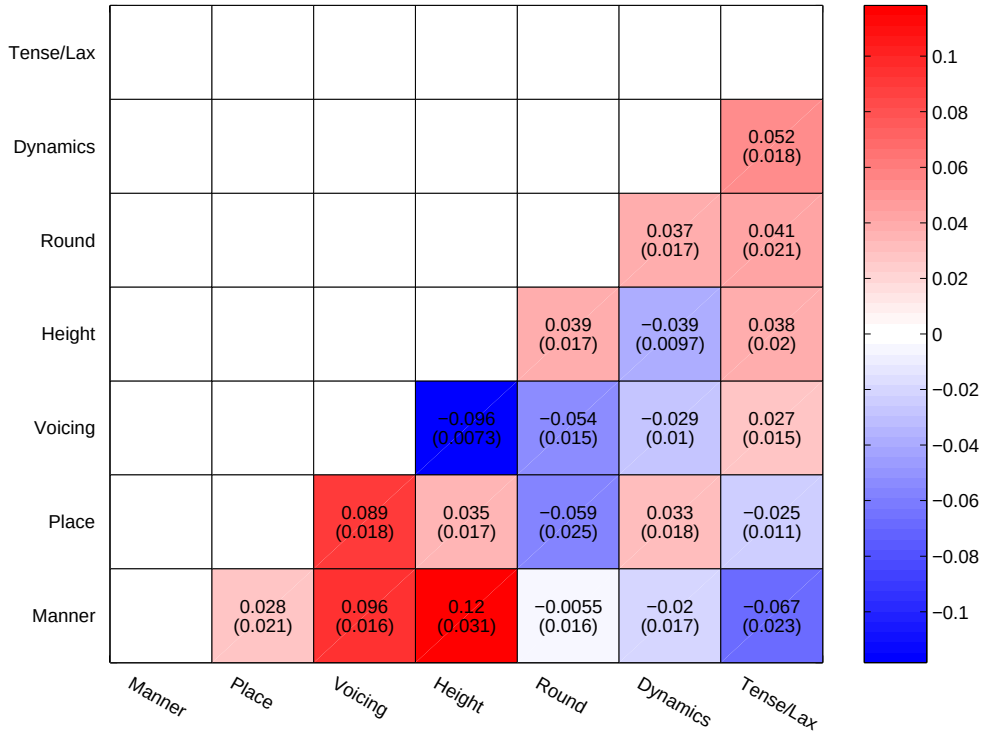

Figure 7.2: Mean two-way interaction indices computed from the trained fuzzy measures of the baseline condition, averaged over 15 random trials. The color-coding represents the magnitude of the interaction where positive and negative interaction indices indicate synergy and redundancy of information contained in the pair of AF dimensions, respectively. The mean value of each interaction index and the standard deviation (in parenthesis) are also shown.

Use Stress-accent			Word-Error Rate%		
Onset	Nucleus	Coda	Baseline	Half-way House	Fabricated
F	F	F	5.98	2.29	1.33
F	T	T	5.85	2.12	1.35
T	F	T	5.95	2.42	1.33
T	T	F	5.87	2.29	1.35
T	T	T	5.59	1.97	1.29

Table 7.3: Comparison of word-error rates with and without incorporating stress-accent information at the onset, nucleus and coda positions in the pronunciation modeling for the three data conditions. An "F" indicates no stress-accent information used at the corresponding syllable position while "T" indicates using stress-accent information.

the five experiments. Each experiment differs from the others in whether the stress-accent estimation is included in the pronunciation-variation statistics at syllable onset, nucleus and coda. An "F" at a syllable position indicates that the pronunciation-variation statistics at that syllable position were computed without conditioning on the stress-accent level. That is, the S in $P(C|T,S)$ (Equation 6.8) only includes syllable position information but not stress-accent levels. This was accomplished by computing $P(C|T,S)$ (cf. Section 6.3.6) using *all* material at each syllable position in the training set, irrespective of stress-accent level. On the other hand, a "T" indicates that the pronunciation variation statistics were conditioned on both syllable position and stress-accent level. Therefore, the first experiment ("F","F","F") does not use stress-accent information at all in modeling the AF variation patterns, while the last experiment ("T","T","T") is simply the regular sort of processing. The other experiments use stress-accent estimation at some syllable positions but not at others.

From Table 7.3 we observe that stress-accent information appears to have greater utility for the baseline and half-way house data conditions than for the fabricated-data condition. This result suggests that stress accent contains information complementary to the initial AF processing, especially when AF classification is sub-optimal. However, the magnitude of reduction in word error when using stress-accent information is not very large, most likely due to the relatively canonical pronunciation and small stress-accent variation in utterances contained in this corpus. A greater improvement from using stress-accent information may be expected if the experiments were performed on more comprehensive spontaneous speech material such as the Switchboard corpus [42].

7.4 Testing Pronunciation Modeling

The previous section described experiments for quantitatively assessing the contribution of stress-accent information to recognition by the systematic control of pronunciation variation modeling. In this section the overall effect of pronunciation variation modeling is analyzed (cf. Table 7.4). The experiments described in this section bear some resemblance

Pronunciation Variation			Word-Error Rate%		
Onset	Nucleus	Coda	Baseline	Half-way House	Fabricated
F	F	F	7.03	2.81	1.76
F	F	T	6.15	2.31	1.50
F	T	F	7.03	2.70	1.56
F	T	T	5.91	2.16	1.33
T	F	F	6.77	2.63	1.82
T	F	T	5.91	2.21	1.61
T	T	F	6.70	2.55	1.63
T	T	T	5.59	1.97	1.29

Table 7.4: Comparison of word-error rates with and without incorporating pronunciation variation statistics at the onset, nucleus and coda positions for the three data conditions. An "F" indicates no pronunciation variation statistics used at the corresponding syllable position while "T" indicates using the pronunciation variation statistics.

to the ones in the previous section. Again, experiments were performed for each data condition and the pronunciation-variation statistics applied or withheld for the various onset, nucleus and coda positions. In this case, if an experiment did not apply pronunciation variation modeling to a specific syllable position (e.g., an "F" in Table 7.4), the $P(C|T,S)$ (cf. Equation 6.8) was simply represented by an identity matrix (i.e. the realized [transcribed] form was assumed to be the same as the canonical form for each AF).

Table 7.4 shows that incorporating pronunciation variation modeling reduces word-error rates by 20% to 30% for the three data conditions. Since this performance improvement is observed even for the fabricated data condition where acoustic modeling is assumed to be near-optimal (using data derived from manually annotated phonetic transcript), it appears that having very accurate acoustic modeling alone is not sufficient, and that some form of pronunciation variation modeling is required to achieve better recognition performance.

More interestingly, the utility of pronunciation variation modeling (as reflected in the amount of word-error reduction) differs across syllable positions. In all three conditions, the onset position has the smallest word-error rate difference (between using, and not using, pronunciation variation information), when the nucleus and coda positions are held constant – both using pronunciation variation statistics. This suggests that onset elements of syllables are the most canonical, conforming to the pronunciation variation patterns of spontaneous speech described in Chapters 4 and 5. In contrast, syllable nuclei and codas exhibit greater variation in pronunciation, and therefore, benefit more from pronunciation variation modeling.

Just as different syllable positions benefit to varying degrees from pronunciation variation modeling, so do different words in the corpus. Table 7.5 presents word-error rates (and relative error reductions) of each of the three data conditions, partitioned by whether or not a word is "canonically" pronounced (according to the manual phonetic transcripts)[2].

[2]There are approximately 17% word tokens pronounced (transcribed) non-canonically.

Word Canonicality	Pronunciation Variation Used	Word-Error Rate%		
		Baseline	Half-way House	Fabricated
Canonical	F	4.94	1.01	0.03
Canonical	T	4.01	0.85	0.03
Error Reduction%		*18.8*	*15.8*	*0*
Non-Canonical	F	17.15	11.51	10.14
Non-Canonical	T	13.27	7.38	7.51
Error Reduction%		*22.6*	*35.9*	*25.9*

Table 7.5: Comparison of the effects of pronunciation variation modeling on word-error rates for the canonically and non-canonically realized words. An "F" indicates no pronunciation variation statistics is used while a "T" indicates the pronunciation variation statistics is used at all syllable positions.

For the non-canonically pronounced words, all three conditions exhibit significant error reduction using pronunciation variation statistics. However, for the canonical word instances, since the fabricated data condition has already achieved almost perfect recognition, no further error reduction is observed. For the baseline and half-way house conditions that make use of automatically derived data, there are significant error reductions even for the canonical word instances (although smaller than that for the non-canonical words). This suggests that either the pronunciation variation statistics is able to partially compensate for inaccuracies in the initial AF classification and segmentation, or the manual transcription contains errors (so that some non-canonical word instances are considered as canonical).

In summary, results described in this section showed that there is a significant advantage in incorporating pronunciation variation modeling in recognition, even when acoustic modeling is near optimal. This is especially true for the non-canonically pronounced words. Since a very large proportion of words were pronounced canonically in the Numbers95 corpus, a greater utility of pronunciation variation modeling may be observed by extending the experiments to a corpus that contains a high proportion of non-canonical pronunciations (e.g. the Switchboard corpus).

7.5 Testing the Contribution of Syllable Position

The results described in the previous sections show that different constituents within the syllable respond differently to pronunciation variation modeling. In this section we test the contributions of the onset, nucleus and coda for word recognition. The current implementation allows us to neutralize information pertaining to syllable position while evaluating within-syllable matching for a specific AF dimension (cf. Section 6.3.6). This was achieved by setting the matching score for a particular syllable position to a constant value regardless of what the inputs and references were. For example, for testing the contribution of the onset, all onset matching scores (or penalty scores) were set to a constant value (e.g. 0.5) so that no useful information could be gained by looking at the onset scores.

Neutralized Syllable Position	Word-Error Rate%		
	Baseline	Half-way House	Fabricated
Onset	15.70	11.27	9.70
Nucleus	20.22	13.28	5.95
Coda	10.13	6.60	3.92
None	5.59	1.97	1.29

Table 7.6: Comparison of the effect on word-error rate by withholding (neutralizing) the contribution from each of the onset, nucleus and coda positions.

Table 7.6 shows word-error rates associated with different syllable position information neutralized. The "None" case is simply the standard system result. It can be observed that neutralizing the "coda" information results in the least increase in errors relative to the standard system, for all three data conditions. This suggests that the coda position contributes the least to lexical access for this particular task. Neutralizing "nucleus" information results in a larger increase in errors relative to neutralizing "onset" information for both the baseline and half-way house conditions (but not for the fabricated condition). This suggests that the baseline and the half-way house conditions rely more heavily on information contained in vocalic nuclei than does the fabricated condition.

7.6 Summary

This chapter has described a number of controlled experiments performed on the Numbers95 corpus using the test-bed implementation of the multi-tier model introduced in the previous chapter. Experiments were partitioned into three separate data conditions: (1) a baseline condition with all automatically derived data, (2) a fabricated condition with almost all information derived from manual phonetic transcripts except stress-accent estimates, and (3) a half-way house condition where some information was automatically derived from the acoustic signal and some information (the initial segmentation and knowledge of whether a segment being vocalic or not) was derived from manual transcription.

Some significant results obtained from the experiments are as follows:

- The overall system performance for the half-way house condition was much closer to the fabricated condition than to the baseline, suggesting accurate segmentation and knowledge of vocalic nuclei is very important for recognition within this framework.
- Shapley scores associated with various AF dimensions and interaction indices of subsets of AF dimensions were computed from the trained fuzzy measures, indicating the relative contribution of each AF dimension to overall recognition, as well as the synergy and redundancy of information contained in subsets of AF dimension scores. This provides a potentially useful technique for feature selection.
- The utility of stress-accent information in word recognition was ascertained by performing experiments either with or without conditioning AF statistics on stress-accent

information. Results showed that word recognition performance improved significantly when stress accent was incorporated, in the baseline and the half-way house conditions. However, stress-accent information did not improve word recognition performance for the fabricated condition, which presumably has near-optimal acoustic modeling.

- The contribution of pronunciation variation modeling was assessed by performing recognition experiments while withholding pronunciation variation statistics associated with some or all of the onset, nucleus and coda positions within each syllable. Significant improvements in word recognition were observed in all three data conditions when pronunciation variation modeling was used, even for the fabricated condition, suggesting the advantage of modeling pronunciation variation explicitly even when acoustic modeling was near optimal. Results also showed that syllable onsets appear to be the most canonical and benefited the least from pronunciation variation modeling, while codas are the least canonical and benefited the most from pronunciation variation modeling. It was also shown that pronunciation variation modeling provides greater gain in word-recognition performance for words that were non-canonically pronounced than words that were canonically realized, but the difference was smaller for the baseline and halfway-house conditions than for the fabricated data condition.

- The test-bed implementation allowed explicit testing of the contribution of the onset, nucleus and coda to word recognition, by neutralizing input information associated with each of the syllable positions separately. Results showed that onsets and nuclei provided greater contributions to lexical access than codas, suggesting a differential treatment of different syllable positions may be appropriate.

The experiments and analysis described in this chapter show that there may be significant advantages of incorporating information from various linguistic tiers, such as articulatory-acoustic features and stress accent, within a syllable-centric model of speech. Pronunciation variation modeling within this framework was effective in improving recognition performance of spontaneous speech. These results also suggest promising directions that could be the focus of future development.

Chapter 8

Conclusions and Future Work

Relying on the conventional phone-based model of speech and statistical pattern recognition techniques, current-generation automatic speech recognition systems are able to perform well on many tasks with certain constraints. However, the performance of state-of-the-art ASR systems still falls far short of expectation on unconstrained, large vocabulary, spontaneous speech tasks. At least part of the problem lies in the potential mismatch between assumptions made by the conventional phonemic-beads-on-a-string model of speech and the reality of spoken language, particularly with respect to pronunciation variation phenomena of spontaneous, natural speech. In Chapter 1 it was suggested that an accurate and efficient alternative model of speech should incorporate information from several linguistic tiers both below and above the phone level; successful recognition may be a result of converging evidence from various sources. An approach was outlined for finding significant elements and structure in speech, and for seeking an alternative model of speech beyond the conventional ones. The remaining chapters described the first few steps along that approach. This concluding chapter first summarizes the main findings of the preceding chapters along with further discussion, then describes some promising future research directions, and ends with some concluding thoughts in the final section.

8.1 Summary and Conclusions

8.1.1 Linguistic Dissection of LVCSR Systems

The first step toward finding a better alternative model of speech was to identify significant factors underlying recognition errors made by state-of-the-art ASR systems on a large vocabulary, spontaneous speech discourse – the Switchboard corpus. The details of the linguistic dissection of LVCSR systems were described in Chapter 2. A statistical analysis was performed on the system outputs at both the word and phonetic-segment levels for both unconstrained recognition and forced-alignment systems, with respect to dozens of linguistic and acoustic parameters. It was found that a correlation exists between word-error rate and phone-error rate across the evaluation sites, suggesting a dependency of word recognition on accurate phone classification and hence on accurate acoustic modeling. It was also found

that there is a significant correlation between word-recognition performance and the average number of pronunciations per word found in the systems' output. Although this comparison did not control for all parameters across the systems, a general dependency of recognition performance could be discerned on the level of sophistication of the pronunciation lexicon. Furthermore, the average number of pronunciations per word found in the systems' output (between 1 and 2.5) was an order of magnitude smaller than that found in the manually annotated transcription, indicating an apparent gap between models and observed data.

Many other linguistic parameters were found to significantly affect word recognition performance. For example, certain types of syllable structure (e.g. vowel-initial forms) exhibit much higher word-error rates than others (e.g. consonant-initial and polysyllabic forms). Prosodic features, such as stress accent and speaking rate, also affect word recognition performance. Systems' tolerance to articulatory feature errors varies as a function of position within the syllable and particular feature dimension of interest. These findings provided valuable information on the significant factors underlying recognition errors that the subsequent development should focus on.

8.1.2 Detailed Analysis of the Elements of Speech

Chapter 3 focused on a more granular representation of the phonetic tier of speech – articulatory-acoustic features. It first reviewed the background of articulatory-acoustic features, surveyed previous research using articulatory-like features and described some of the advantages of using AFs often cited by other authors, particularly the representational flexibility. The rest of the chapter concentrated on the computational aspect of automatic AF processing, providing further evidence for incorporating AFs into speech recognition. A TFM/MLP neural-network-based AF-classification system was described in detail illustrating feasibility of accurate AF extraction from the acoustic signal, as well as demonstrating the utility of extending the AF classification to a system for automatic phonetic labeling. Further experiments were performed on the more comprehensive NTIMIT corpus, and in particular, an "elitist" approach was presented to delineate regions of speech with high confidence in the AF classification and a manner-specific training scheme was described for enhancing place-of-articulation classification, potentially useful in conjunction with accurate manner classification. The cross-linguistic transferability of AF training was assessed quantitatively by testing (American English) NTIMIT-corpus-trained AF-classification networks on a Dutch corpus (VIOS). Experimental results showed that certain AF dimensions (e.g. voicing and manner of articulation) transfer better than others (e.g. place of articulation), suggesting that caution must be exercised when attempting to transfer ASR systems across languages. Further evidence supporting the use of AFs was provided by the robustness of AFs, as demonstrated in experiments involving speech in noisy backgrounds, particularly when the AF-classification system is trained on speech embedded in a variety of noise backgrounds over a wide dynamic range of SNRs ("mixed-training"). Incorporating AF classification as an intermediate stage often yields significant improvements to phonetic classification; system trained under the "mixed-training" scheme not only performed well under noise conditions included in the training but also generalized to superior performance to many novel noise conditions.

Chapter 4 was devoted to the analysis of a suprasegmental unit of speech – the syllable. It described the central role played by the syllable in spoken language and provided evidence to support the syllable being the binding unit of speech, around which information at various linguistic tiers is organized. The stability and importance of the syllable in speech perception was emphasized by several sources of evidence: (1) statistics from spontaneous speech corpora, and (2) acoustics-based syllable detection and segmentation, as well as (3) the significance of syllable duration in speech perception. Through concrete examples of word instances extracted from spontaneous speech material, syllable-level information was shown to be very helpful in describing the observed pronunciation variation patterns. In particular, the analysis showed that the nuclei and codas of syllables are more likely to deviate from canonical pronunciation than the onsets. Moreover, the common types of deviation patterns (substitution, insertion and deletion) differ with respect to position within the syllable. The intimate relationship between articulatory-acoustic features and the syllable was reinforced by the significant gain in AF and phonetic classification accuracies when the syllable position information was incorporated for speech in both clean and noisy backgrounds. The evidence from this chapter supports an explicit incorporation of the syllable as the fundamental binding unit in models of speech recognition. It also illustrated a specific structural framework of speech, through which information from various linguistic tiers can be linked to each other.

Chapter 5 focused on an important prosodic feature of spoken English – stress accent, and especially on the impact of stress-accent level on pronunciation variation in spontaneous speech. In the survey of background information, the perceptual basis of stress accent and the intimate relationship between stress accent and vocalic identity were discussed. The interdependency of vocalic identity and stress-accent levels [63] is a significant departure from the traditional linguistic perspective, but could potentially prove very useful for automatic speech processing [58]. The impact of stress accent on pronunciation variation in spontaneous speech was demonstrated first by using examples of word instances extracted from the Switchboard corpus and then by using overall deviation patterns from canonical pronunciation computed over a subset of the Switchboard material that has been manually labeled at the prosodic stress-accent level. A significant contrast was observed among different stress-accent levels when patterns of pronunciation deviation from the canonical were partitioned by stress-accent level, in conjunction with syllable position. Furthermore, detailed statistics of the pronunciation deviation patterns from the canonical were computed based on the realization of articulatory-acoustic features, with respect to syllable position and stress-accent level. Such results demonstrated that pronunciation variation of spontaneous speech can be largely captured via the systematic relationship among the AFs, syllable structure and stress accent in a parsimonious fashion, and these findings formed the basis of an alternative model of speech. In addition, a neural-network-based system was developed to automatically label stress accent for spontaneous speech and experimental results showed that the automatic system was able to perform at a level comparable to a human transcriber. This development demonstrated the feasibility of automatically extracting prosodic stress-accent information from the acoustics and provided useful information for assessing the contribution of different cues for stress-accent determination. It was found that, in contrast to the traditional linguistic framework, the most salient features for stress accent are related to energy, duration and vocalic identity. Pitch-related features

were found to play only a minor role.

8.1.3 An Alternative Model of Speech

The analysis and experiments presented in Chapters 3-5 provided evidence in support of an alternative model of speech, incorporating articulatory-acoustic features, syllable structure and stress accent. Chapter 6 described a multi-tier model based on this framework. The multi-tier model views speech as organized at the syllable level, with each syllable in an utterance capable of manifesting a distinctive level of stress accent. A syllable consists of a vocalic nucleus and optionally a consonantal onset and coda constituents, which are characterized by features along several quasi-orthogonal AF dimensions. Much pronunciation variation is captured parsimoniously by AF deviation patterns from the canonical within the context of syllable position and stress accent. Recognition using the multi-tier model is viewed as combining evidence from heterogeneous information sources across time and space.

A test-bed implementation was built based on the multi-tier model to perform controlled word-recognition experiments on a limited-vocabulary task. The implementation made a number of simplifying assumptions in order to provide a simple and transparent system to facilitate convenient diagnostic analysis, rather than to build a scalable, high-performance system. Major components of the system are the initial AF classification and manner-based segmentation, stress-accent labeling, word-hypothesis scoring based on syllable alignments, computing individual AF dimension scores and their combination for syllable matching. A fuzzy-measure/fuzzy-integral-based multiple-information-aggregation approach was adopted to combine matching scores from various AF dimensions, which can take into account the relative importance and interaction of various AF dimensions. The differential contributions associated with the features can be interpreted using the automatically learned fuzzy-measure parameters. The statistic patterns of AF-deviations from canonical realization can be obtained from the training data and used in the evaluation of individual AF matching scores at the onset, nucleus and coda positions within the syllable, to facilitate syllable-position and stress-accent-based pronunciation variation modeling.

A number of controlled experiments were performed on the Numbers95 corpus using the test-bed implementation of the multi-tier model. To enable informative diagnostic analysis, three different data conditions were adopted: (1) a baseline condition with entirely automatically derived data, (2) a fabricated data condition with most information derived from manual transcription, and (3) a half-way house condition where the initial segmentation and the knowledge of whether a segment is vocalic or not were derived from the manual transcription (with all other features automatically computed). Analysis of the overall system performance showed that the half-way house condition had a performance very close to that of the fabricated data condition, highlighting the importance of accurate phonetic segmentation and knowledge of vocalic location. This result conformed, from a different perspective, the importance of segmentation and syllable position information as observed in previous research. For example, in their Time Index Model experiments, Konig and Morgan [77] observed that word-error rate could be greatly reduced if accurate segmental boundary information were available. In another study, Wu and colleagues re-

ported significant gains in word recognition performance by integrating syllable boundary information into speech recognition systems [145].

Shapley scores and interaction indices of various AF dimensions were computed from the trained fuzzy measures, providing information pertaining to the relative contributions to word recognition associated with various AF dimensions, as well as the redundancy and synergy among subsets of AF dimensions. The result agreed with the intuition of different utility being provided by various AF dimensions to recognition and may be a potential basis for feature selection. Results from recognition experiments with and without using stress-accent estimation showed that stress-accent information was particularly helpful when the initial AF processing was sub-optimal such as when automatically derived data were used instead of fabricated data. The contribution of pronunciation variation modeling was assessed at different positions within the syllable. It was found that incorporating pronunciation variation statistics significantly reduced recognition error for all three data conditions and utilities of the pronunciation variation modeling were greater at the nucleus and coda positions than at the onset position. This result suggests that it is important to have models of pronunciation variation even when acoustic modeling is near optimal (as in the fabricated data condition). Finally, by neutralizing information from onsets, nuclei or codas separately differential contributions from the three positions were ascertained and it was found that onsets and nuclei are more important than codas for word recognition by machine.

8.2 Future Directions

This thesis has described a few important steps toward the development of speech recognition models and systems that would perform as well as humans do under many realistic conditions, using the approach outlined in Chapter 1. Much remains to be done. This section describes a few promising directions of future research, as well as possible applications of insights and ideas generated from this current work to other approaches.

8.2.1 Incorporation into Conventional Systems

It should be possible to incorporate some of the insights and ideas from the current work into current-generation ASR systems without an overhaul of the conventional model of speech that such systems are based on. One way to accomplish this is through a dynamic pronunciation modeling framework, such as the one developed by Fosler-Lussier in his thesis [36]. Such a method is most suitable in a multi-pass system where an N-best list or a recognition lattice is rescored by a dynamic dictionary containing pronunciation models produced by statistical decision trees based on contextual information from the output of the previous pass [36][37]. Within such a framework, quasi-independent estimation of stress accent and articulatory-acoustic features can be taken into consideration by decision trees (or other modeling techniques) to provide more accurate conditioning for pronunciation variation. Such an approach is likely to provide some gain in recognition performance but may not be able to take full advantage of AF and prosodic information since the models are still constrained by limitations of using the phone as the fundamental unit of speech.

Another perspective that may be taken is that, instead of generating different pronunciation models given the context, the system can re-interpret acoustic modeling results depending on the recognition context, while keeping pronunciation models relatively canonical and simple. This is, in fact, the approach taken by the current work in the multi-tier model and the test-bed implementation where detailed AF classification results are re-interpreted depending on syllable-position and stress-accent context. How this approach may be incorporated into a conventional system would depend on the particular implementation. A combination of the two perspectives may be the most desirable approach. Since a system incorporating additional information such as AF and stress-accent information is likely to have different characteristics from a purely conventional phone-based system, combining the two should, in principle, yield superior (or at the very least not inferior) performance than that of either one alone.

8.2.2 Further Analysis and Experiments

The exploratory analysis and experiments described in Chapters 3-5 concentrated on the roles played by AF, syllable structure and stress accent in modeling spoken English. Although these primary components, interacting within a systematic structure, are able to provide an accurate and parsimonious description of pronunciation variation in spontaneous speech, it would be useful to extend the detailed analysis to other parameters. For example, one may consider speaking rate, intonation, word predictability, hesitation and disfluency, dialectal accent, etc., many of which have been shown to be important factors influencing the specific phonetic realization of speech (e.g. [36][51][119]). It would be important to ascertain how much pronunciation variation due to these various factors has been captured by such factors as syllable position and stress accent. It would also be interesting to perform comparable analysis on different languages other than English, in particular on languages that have very different prosodic accent characteristics (such as Japanese) and on those languages that make use of certain linguistic properties differently from English (such as tonal languages where the tone is used for lexical distinction). It is expected that these other languages would have some different manifestation of detailed pronunciation variation patterns and have different parameters to provide contextual information. However, it is likely that a similar structural organization (with possibly different elements) is shared in common across different languages and a similar framework to the one described in the current work applies.

One of the contributions of the test-bed implementation and controlled experiments is the identification of those aspects of the multi-tier model which would bring the most benefit to overall recognition. For example, through the comparison of recognition performance among the baseline, fabricated and half-way house data conditions, the initial phonetic segmentation and the identification of vocalic nuclei of syllables exhibit great importance in successful recognition and therefore may be worth significantly improving. As the basis for segmentation and subsequent processing, the AF classification certainly requires further enhancement. Among various articulatory-acoustic feature dimensions, the relative contribution to recognition may be used as an indicator for allocating resources to improve classification accuracy. The combination of the "elitist" approach and the manner-

specific training also provides a promising direction of further investigation. Although the front-end signal processing has not been a major focus of this thesis, development of better front-end techniques would certainly be highly beneficial to AF processing. For example, since different AF dimensions possess very different spectro-temporal characteristics, it may be useful to adopt different front-end processing strategies for various AF dimensions.

The test-bed implementation and experiments described in the current work have been limited to a constrained task with a relatively modest vocabulary size. Although the test material possesses many characteristics of spontaneous speech, it does not cover the full range of variability observed in large vocabulary, spontaneous corpora such as the Switchboard corpus. Consequently the experimental results may have not fully shown the advantage of modeling techniques employed and the analysis may not be conclusive. Therefore it would be very useful to extend the implementation and experiments to an unconstrained, large vocabulary spontaneous speech corpus such as Switchboard. Such an extension would be a quite challenging task due to the requirement of scalability as well as the increased difficulty of acoustic modeling and greater reliance on higher-level processing such as language modeling.

8.2.3 An Improved Framework and Implementation

The test-bed implementation described in the current work made many simplifying assumptions such as (1) manner-based segmentation, (2) synchronization of AF dimensions to manner segments, and (3) the heuristic matching score evaluation of AF dimensions for each syllable position. Although these simplifying assumptions were not without merit and did not cause any significant performance degradation on the constrained recognition task, they should be addressed in future system implementations, particularly if more comprehensive and difficult recognition tasks are to be considered.

In both the multi-tier model of speech as well as the test-bed implementation, higher-level processing such as language modeling, semantic and pragmatic modeling were not explicitly considered, in order to simplify system development. However, these higher-level factors are certainly very important in speech recognition and may also exert a significant influence on pronunciation variation. In most practical situations, there will always be a certain degree of confusibility in pronunciation and acoustic models that requires the higher-level processing to reduce or eliminate. Future developments, especially those that aim at good scalability and high performance, would certainly need to incorporate such higher-level processing.

In the long run, a unified framework that combines information from various linguistic levels is desirable; in addition, more elegant and sophisticated mathematical techniques should be adopted where appropriate. A successful model of speech recognition and its implementation must also provide satisfactory solutions to problems of adaptability and robustness.

8.3 Coda

The current study relied heavily on the availability of realistic speech data, particularly that with high-quality manual annotation at various linguistic tiers, at least during the exploratory analysis and diagnostic experiment stages. Such data provide important information pertaining to properties of natural speech that laboratory speech (e.g. speech data that are planned, scripted and recorded in a laboratory environment) may not be able to provide, such as broad range of variability in pronunciation. Statistical analyses of natural speech data with respect to manual annotation by linguistically trained individuals help us gain insights into speech perception by human listeners, as well as provide us with a more accurate reference for evaluating experimental results than existing automatic systems provide. Interestingly, the ability to analyze a significant amount of natural speech data may lead to observations inconsistent with traditional linguistic theories, and very often the statistical analysis of natural speech presents a more realistic characterization and may thus provide a basis for refining current linguistic theory. Although the benefits are abundant, acquiring natural speech data is often an expensive and time-consuming task, especially manual annotation, which requires intensive labor and expertise from highly trained individuals. Consequently, such data exist for few languages. Thus, one of the goals of developing accurate automatic transcription systems is to make such data collection process simpler and cheaper; the benefit is likely to go beyond just ASR applications, but should also advance the state of linguistic and cognitive science studies in general.

A major force behind the progress in ASR research during the past few decades is the advancement of modern, machine-learning techniques, particularly statistical pattern-recognition methods such as hidden Markov models and neural networks. These technologies not only enable elegant recognition system design but also provide means to automatically learn the necessary system parameters from a vast amount of data in a relatively efficient manner. However, the contribution of computational technology is paralleled by linguistic knowledge, and such insights (from linguistics, cognitive and neural science) can also aid in the development of the technology. Such scientific understanding helps direct the focus of computational and engineering developments to more efficiently navigate through a near-infinitely large space of models and techniques toward building superior systems. Thus, the surest path to a solution of the ASR problem may be to maintain a relative balance between basic science and engineering.

The approach taken in the current work is toward a general solution to the problem of speech recognition that can be adopted for many different tasks. However, it is often the case that how a system is developed, trained and assessed depends heavily on the criterion for evaluating system performance. The most commonly used criterion, word-error rate (the sum of the substitution, deletion and insertion rates), may not be optimal for every task. For example, a recognition result with a 25% word-error rate is far from optimal speech transcription but may be sufficient for certain information retrieval tasks [143] where redundancy in the transcription output can be productively exploited. As another example, it has often been observed that increasing the accuracy of phonetic classification or reducing the language-model perplexity does not necessarily lead to a significantly decreased word-error rate in conventional ASR systems[18]. Therefore, caution must be taken when selecting

a suitable evaluation metric, which should be closely linked to the ultimate goal of the application.

Naturally spoken speech is full of variability and uncertainty. Complete reliance on a single source of information, derived from only one single linguistic tier, is likely to fail in capturing the full range of possibilities and thus will result in serious performance degradation when the test environment changes. When an unexpected variation occurs, evidence from a single source may be distorted and misleading but recognition as a result of converging evidence from many (quasi-)independent sources is likely to remain robust since a certain invariance is likely to be found in some of the various representations for a particular condition. Similarly, agreement derived from an ensemble of techniques is a good indication that an accurate and reliable result has been achieved.

Natural speech is a very complex phenomenon and a rich information source. However, not all information contained in the speech signal is relevant for a particular recognition task. It is therefore important to identify structure that helps capture the relevant information in a well-organized and parsimonious fashion. The computational effort may be better concentrated on interesting and reliable cues of speech rather than divided equally across time and space. Finally a computational technique should not attempt to provide an overly precise characterization of elements of speech that are inherently vague and ambiguous. An overly strong commitment to a precise description may lead to unnecessarily restricted hypotheses and ultimately to recognition errors.

Appendix A

Supplementary Information on Linguistic Dissection of LVCSR Systems

This appendix provides additional information on the linguistic dissection of Switchboard-corpus LVCSR systems described in Chapter 2. A mapping procedure between the reference phone set and submission sites' phone set, along with inter-labeler agreement patterns, is presented first. The following section provides a detailed description of the evaluation procedure, including file format conversion, recognition output scoring and analysis data generation.

A.1 Phone Mapping Procedure and Inter-labeler Agreement

Because each of the participating sites used a quasi-custom phone set, it was necessary to convert each submission to a common format. This was done by first devising a mapping from each site's phone set to a common reference phone set (cf. Table A.1 for a description of the reference phone set), which was based on the STP [49] material, but was adapted to match the less granular symbol sets used by the submission sites. The reference phone set was also inversely mapped to the submission site phone sets to ensure that variants of a phone were given due credit in the scoring procedure. For example, [em] (a syllabic nasal) was mapped to [ix] + [m] and the vowel [ix] was mapped in certain instances to both [ih] and [ax], depending on the specifics of the phone set. This two-way phone mapping procedure was used in both years' evaluations.

For the Year-2001 evaluation in particular, we have added another phone mapping procedure to allow for certain phones commonly confused among human transcribers to be scored as "correct" even though they would otherwise be scored as "wrong." We call this specific mapping the transcription-compensated form, in contrast to the uncompensated form where only common phone ambiguities were allowed.

In order to devise the transcription-compensated phone mappings, we analyzed

Phone	Example/Description	Phone	Example/Description
b	'bob'	el	'bottle'
d	'dad'	r	'red'
g	'gag'	w	'wet'
p	'pop'	y	'yet'
t	'tot'	hh	'hay'
k	'kick'	iy	'beat'
dx	'forty'	ih	'bit'
q	glottal stop	eh	'bet'
jh	'judge'	ey	'bait'
ch	'church'	ae	'bat'
s	'sis'	aa	'robot'
sh	'shoe'	aw	'down'
z	'zoo'	ay	'bite'
zh	'measure'	ah	'much'
f	'fief'	ao	'bought'
th	'thief'	oy	'boy'
v	'verb'	ow	'boat'
dh	'they'	uh	'book'
m	'mom'	uw	'boot'
em	'bottom'	er	'bird'
n	'non'	ax	(unaccented) 'the'
nx	'winner'	ix	(unaccented) 'roses'
ng	'sing'	h#	non-speech other than silence
en	'button'	pv	filled pause-vocalic
eng	'Washington'	pn	filled pause-nasal
l	'led'	sil	silence

Table A.1: Description of the reference phone set used in the Switchboard-corpus LVCSR system phonetic evaluations.

Segment	Uncompensated	Transcription Compensated
[d]	[d]	[d] [dx]
[k]	[k]	[k]
[s]	[s]	[s] [z]
[n]	[n]	[n] [nx] [ng] [en]
[r]	[r]	[r] [axr] [er]
[iy]	[iy]	[iy] [ix] [ih]
[ao]	[ao]	[ao] [aa] [ow]
[ax]	[ax]	[ax] [ah] [aa] [ix]
[ix]	[ix] [ih] [ax]	[ix] [ih] [iy] [ax]

Table A.2: Selected forms of segment interchanges allowed in the transcription-compensated and uncompensated scoring.

the inter-labeler agreement patterns among three transcribers on a subset of the evaluation material. Figure A.1 shows the average agreement rate for each phone among the three transcribers. The overall inter-labeler agreement rate is 74 percent but the disagreement patterns differ across segments. For the consonantal segments, stop (plosive) and nasal consonants exhibit a low degree of disagreement, fricatives exhibit slightly higher degree of disagreement and liquids show a moderate degree of disagreement; for the vocalic segments, lax monophthongs exhibit a high degree of disagreement, diphthongs show a relatively low degree of disagreement and tense, low monophthongs show relatively little disagreement.

From these inter-labeler disagreement patterns, we have devised transcription-compensated phone mappings for the Year-2001 evaluation. Table A.2 shows the forms of tolerances allowed in the transcription-compensated scoring.

A.2 Evaluation Procedure

A.2.1 File Format Conversion

In order to score submissions in terms of phone-segments and words correct, as well as perform detailed analyses of the error patterns, it was necessary to convert the submissions into a common format. The following steps were required:

- A reference set of materials at the word, syllable and phone levels was created from the transcript to include:
  - word-to-phone mapping
  - syllable-to-phone mapping
  - word-to-syllable mapping
  - time points for the phones and words in the reference material

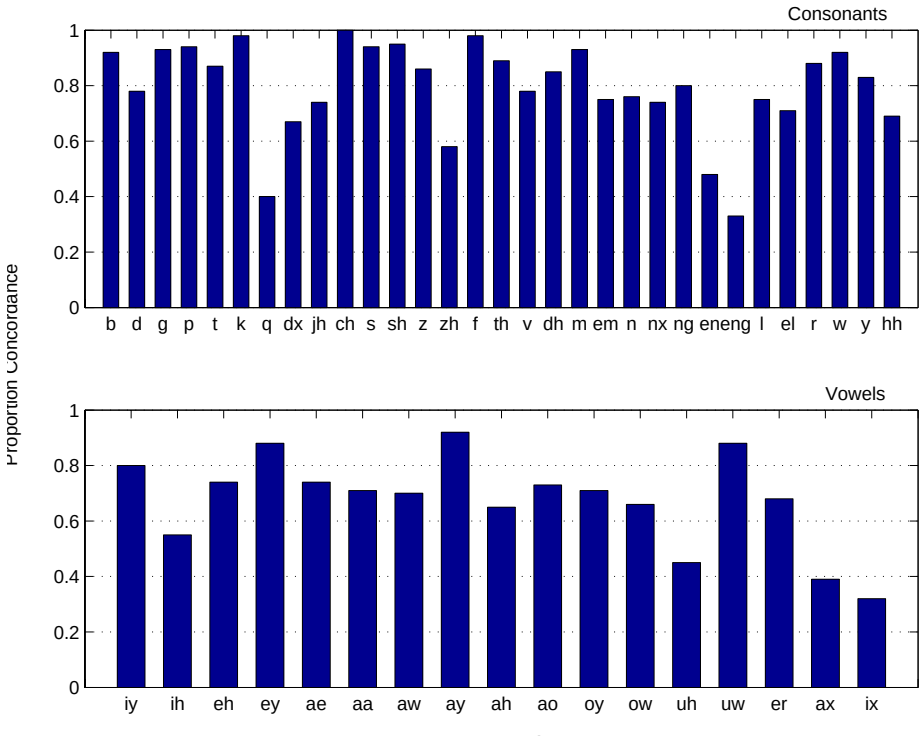

Figure A.1: Average concordance for each phone (partitioned into consonants and vowels) among three transcribers. The overall inter-labeler agreement rate is 74%. For the consonantal segments, stop (plosive) and nasal consonants exhibit a low degree of disagreement, fricatives exhibit slightly higher degree of disagreement and liquids show a moderate degree of disagreement; for the vocalic segments, lax monophthongs exhibit a high degree of disagreement, diphthongs show a relatively low degree of disagreement and tense, low monophthongs show relatively little disagreement.

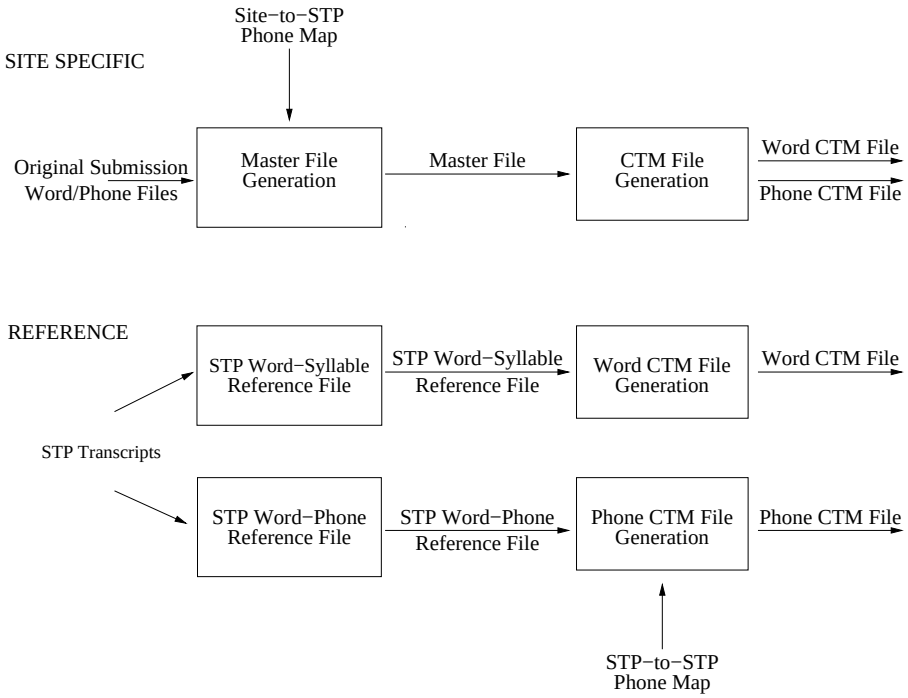

Figure A.2: Flow diagram of the file conversion process for the diagnostic evaluation. For each site, word and phone files (in CTM format) were generated from the original submission and a site-to-STP phone map was applied to phone labels. Reference word and phone files (in CTM format) were generated from the original STP transcripts, and a word-syllable-mapping file and a word-phone-mapping file were also created.

- Word and phone level reference files were created in NIST's Time-Marked Conversation (CTM) format [99].
- Phone mapping procedures as described in the previous section were applied to the sites' submissions (both unconstrained and constrained recognitions) and files (in NIST's CTM format) were generated at both the word and phone levels.

The NIST's CTM format specifies for each token (a word or phone segment) a label, the beginning time point, the duration and optionally a confidence score. In addition, the CTM format allows alternative tokens to be specified for the same temporal segment, this capability was very useful for the transcription-compensated phonetic scoring where certain phonetic segments were allowed to map without penalty to any of several different reference phones. Figure A.2 shows the steps involved in the file-conversion process.

A.2.2 Scoring the Recognition Systems

Once the reference and submission files were converted into the CTM format we were able to temporally align the words and phones in the submission material unambigu-

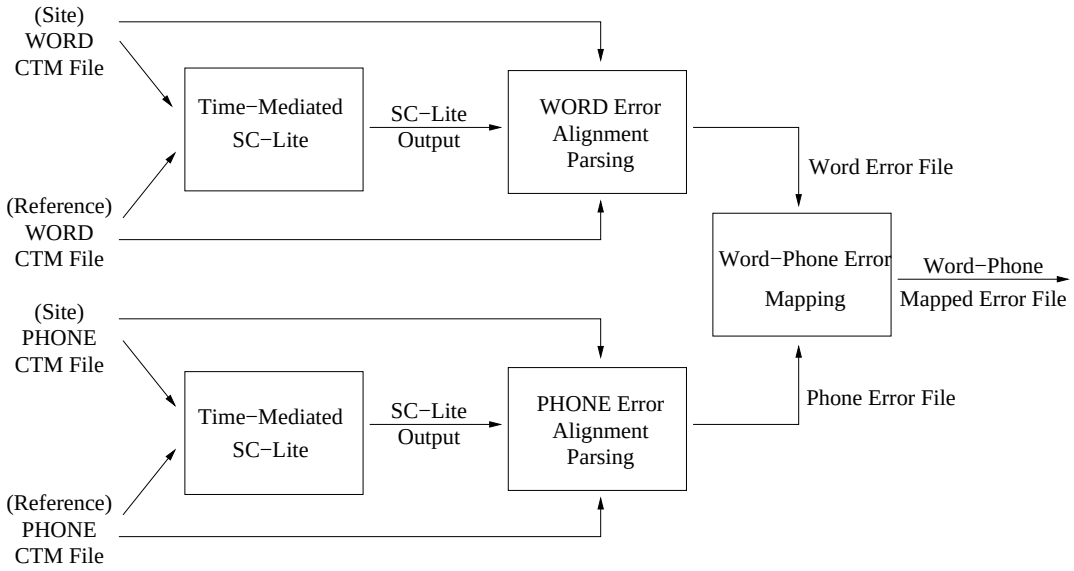

Figure A.3: Flow diagram of the scoring procedure used in the diagnostic evaluation. For each site, time-mediated alignment between submission material and reference material was performed at both the word and phone levels separately. Each word and phone segment was scored in terms of being correct or not; each incorrect segment was also assigned an error type. Word- and phone-error files were aligned and merged into a single file according to the temporal relationship between words and phones.

ously to that in the reference set, and to perform time-mediated scoring using SC-Lite [99], a program developed by NIST to score competitive ASR evaluation submissions. This *strict* time mediation was used in both years' evaluations. However, since the word and phone segmentation of the submission material often deviates from those of the STP-based reference material, for the Year-2001 evaluation we also developed a *lenient* time mediation by de-weighting the time-mismatch penalty in the SC-Lite's alignment algorithm.

SC-Lite scores each word (and phone) segment in terms of being correct or not, as well as designating the error as one of three types – a substitution (i.e., $a \rightarrow b$), an insertion ($a \rightarrow a + b$) or a deletion ($a \rightarrow \phi$). A fourth category, *null*, occurs when the error cannot be clearly associated with one the the other three categories (and usually implies that the error is due to some form of formatting discrepancy). The *lenient* time mediation generates about 20% fewer phone "errors" than the *strict* time mediation. The sample output in Tables A.3 and A.4 illustrates the two time-mediated scoring methods.

Finally, the SC-Lite output files at the word and phone levels were aligned and merged, and the resulting word-phone error mapping was used as the basis for generating the data contained in the summary tables ("big lists") (described in Section A.2.3). The entire scoring procedure is depicted in Figure A.3.

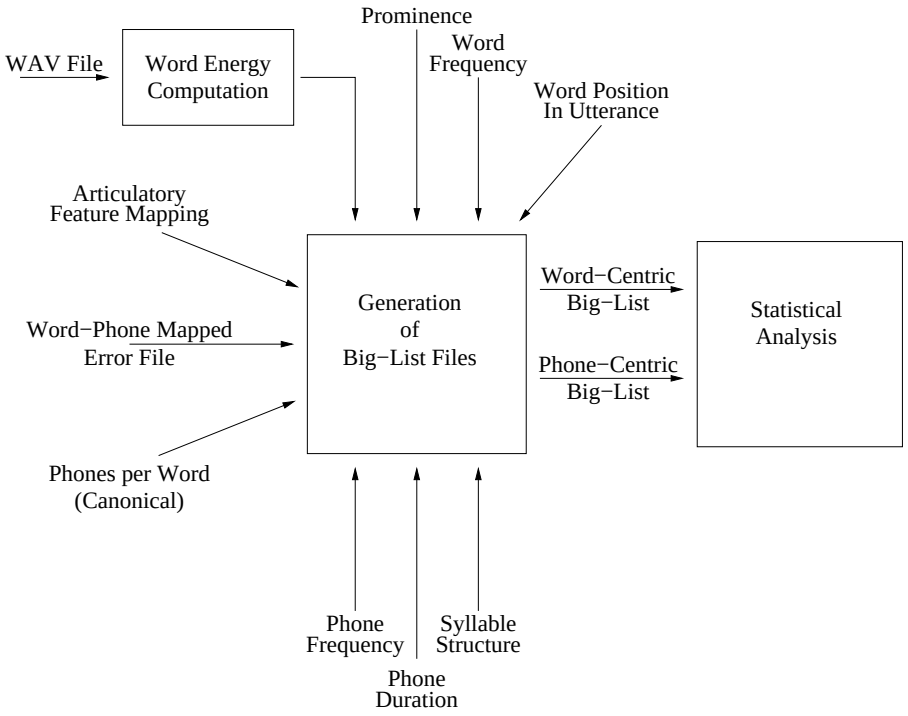

Figure A.4: Flow diagram of the generation of "big lists" used in the diagnostic evaluation. Word-centric and phone-centric "big lists" were separately generated from each word-phone mapped error file (cf. Figure A.3). A number of linguistic parameters pertaining to each word and phone segment were computed and included in the "big lists," which were used to perform statistical analyses.

ID	REF WD	HYP WD	WS	RB	HB	RP	HP	PS	RB	HB
SWB_	I'D	UP	S	2.82	2.82	AY	AX	S	2.82	2.82
40035						D	**	S	2.89	**
-A-						P	P	C	2.97	2.92
0053	PREFER	FOR	S	2.97	3.10	F	F	C	3.12	3.10
						ER	ER	C	3.19	3.17
	THE	THE	C	3.31	3.27	DH	DH	C	3.31	3.27
						AX	AX	C	3.37	3.36
	H#	H#	C	3.42	3.43	H#	H#	C	3.42	3.43
	CITY	CITY	C	3.71	3.71	S	S	C	3.71	3.71
						IH	IH	C	3.79	3.78
						**	DX	I	**	3.84
						DX	**	D	3.82	**
						IY	IY	C	3.85	3.87

Table A.3: A sample, composite output from SC-Lite *strict*-time-mediated scoring at the word and phone levels. ID (SWB_40035-A-0053) pertains to the entire word sequence, REF WD is the the reference word (H# for silence), HYP WD is the recognized word, WS is the word score (C=correct, S=substitution), RB is the beginning time (in seconds) of the reference word or phone, HB is the beginning time of the recognizer output, RP is the reference phone, HP is the recognized phone, and PS is the phone score (I=insertion, D=deletion). The insertion/deletion of the phone, DX, is due to temporal misalignment (cf. Table A.4 for *lenient* time-mediation).

ID	REF WD	HYP WD	WS	RB	HB	RP	HP	PS	RB	HB
...										
SWB_	CITY	CITY	C	3.71	3.71	S	S	C	3.71	3.71
40035						IH	IH	C	3.79	3.78
-A-						DX	DX	C	3.82	3.84
0053						IY	IY	C	3.85	3.87

Table A.4: A sample, composite output from the SC-Lite *lenient*-time-mediated scoring at the word and phone levels for the word "CITY" in the same utterance as shown in Table A.3. Note that the insertion/deletion of the phone DX in the *strict* time-mediation is scored as "correct" in the *lenient* time-mediation.

A.2.3 Data Generation

In order to easily manipulate the scoring results and perform statistical analyses on the data, we generated summary tables ("big lists") from the word-phone mapped files and included dozens of separate parameters pertaining to speaker-specific, linguistic and acoustic properties of the speech material, including energy level, duration, stress-accent pattern, syllable structure, speaking rate and so on (cf. Table A.5 for a complete list of the parameters). The "big lists" were generated in either word-centric or phone-centric format (cf. Figure A.4) to provide for analysis at either word or phone level. A sample subset of a word-centric "big list" file is shown in Table A.6.

Lexical Level Parameters		Phone Level Parameters	
1	Word Error Type - Sub, Del, Ins, Null	22	Phone ID (Reference and Hypothesized)
2	Word Error Type Context (Preceding,Following)	23	Phone Duration (Reference and Hypothesized)
3	Word (Unigram)Frequency (in Switchboard Corpus)	24	Phone Position within the Word
4	Position of the Word in the Utterance	25	Phone Frequency (Switchboard Corpus)
5	Word Duration (Reference and Hypothesized)	26	Phone Energy
6	Word Energy	27	Phone Error Type (Sub, Del, Ins, Null)
7	Time Alignment Between Ref. and Hyp. Word	28	Phone Error Context (Prec., Following Phone)
8	Lexical Compound Status (Part of a Compound or Not)	29	Time Alignment Between Ref. and Hyp. Phone
9	Prosodic Prominence (Max and Avg. Stress Accent)	30	Position Within the Syllable
10	Prosodic Context - Max/Avg. Accent (Prec/Following)	31	Manner of Articulation
11	Occurrence of Non-Speech (Before, After)	32	Place of Articulation
12	Number of Syllables in Word (Canonical and Actual)	33	Front-Back (Vocalic)
13	Syllable Structure (CVC,CV, etc.-Canonical & Actual)	34	Voicing
14	Number of Phones in Word (Canonical and Actual)	35	Lip Rounding
15	Number of Phones Incorrect in the Word	36	Cumulative Phonetic Feature Distance
16	Type and Number of Phone Errors in Word	Utterance Level Parameters	
17	Phonetic Feature Distance Between Hyp./Ref. Word	37	Utterance ID
Speaker Characteristics		38	Number of Words in Utterance
18	Dialect Region	39	Utterance Duration
19	Gender	40	Utterance Energy
20	Recognition Difficulty (Very Easy to Very Hard)	41	Utterance Difficulty (Very Easy to Very Hard)
21	Speaking Rate - Syls/Second and MRATE	42	Speaking Rate - Syllables per Second
		43	Speaking Rate - Acoustic Measure (MRATE)

Table A.5: A list of the speaker, utterance, linguistic (prosodic, lexical, phonetic) and acoustic characteristics computed for the diagnostic component of the Switchboard evaluation, the output of which was compiled into summary tables ("big lists") for each submission.

ERR	REFWORD	HYPWORD	UTID	WORDPOS	FREQ	ENERGY	MRATE	SYLRATE	ETC.
C	A	A	40035-A-0032	0.53	-1.67230	0.98	3.750	4.190	...
C	WOMAN	WOMAN	40035-A-0032	0.58	-3.95740	0.98	3.750	4.190	...
C	NAMED	NAMED	40035-A-0032	0.63	-4.72990	0.90	3.750	4.190	...
C	TERRY	TERRY	40035-A-0032	0.68	-5.47590	0.94	3.750	4.190	...
S	GILLIAN	GET	40035-A-0032	0.74	-6.02000	0.97	3.750	4.190	...
I	***	ONE	40035-A-0032	0.79	-6.02000	0.95	3.750	4.190	...
C	LAST	LAST	40035-A-0032	0.84	-3.17830	0.88	3.750	4.190	...
C	SPRING	SPRING	40035-A-0032	0.89	-4.08800	0.88	3.750	4.190	...
C	H#	H#	40035-A-0032	0.95	-6.02000	0.85	3.750	4.190	...
C	AND	AND	40035-A-0033	0.04	-1.50240	1.10	3.440	5.730	...
C	WE	WE	40035-A-0033	0.08	-2.10820	0.96	3.440	5.730	...
N	H#	***	40035-A-0033	0.12	-6.02000	0.79	3.440	5.730	...
C	PUT	PUT	40035-A-0033	0.17	-3.04610	1.05	3.440	5.730	...
S	ADS	THAT	40035-A-0033	0.21	-4.88960	0.99	3.440	5.730	...
C	IN	IN	40035-A-0033	0.25	-1.92950	0.88	3.440	5.730	...
C	THE	THE	40035-A-0033	0.29	-1.55280	0.69	3.440	5.730	...
C	CITY	CITY	40035-A-0033	0.33	-3.62110	0.92	3.440	5.730	...

Table A.6: Sample of a word-centric "big list" file. "ERR" is word error type: "C" is correct, "S" is substitution, "I" is insertion, "N" is *null* error (see text for an explaination). "REFWORD" and "HYPWORD" refer to reference word label and hypothesis word label, respectively. "UTID" is the Switchboard utterance ID. "WORDPOS" is the position of the word within the utterance normalized to between 0 and 1. "FREQ" is the unigram frequency of the reference word (in log probability). "ENERGY" is the normalized energy of the word (over the entire utterance). "MRATE" is the MRate, an acoustic measure of speaking rate of the utterance [92]. "SYLRATE" is a linguistic measure of speaking rate (syllables per second) for the utterance.

Appendix B

Pronunciations of "But"

In Chapter 4 various phonetic realizations of the word "that," extracted from the Year-2001 phonetic evaluation material, were compared to the canonical pronunciation, [dh ae t](cf. Tables 4.1 and 4.2). The deviations from canonical were further partitioned by stress-accent level in Chapter 5 (cf. Tables 5.1-5.6). The word "that" is unusual with respect to the onset segment (i.e., [dh] in the canonical pronunciation), which tends to have a greater number of deviations from canonical than other onset segments in general. To get a perspective using a different word, this Appendix shows sample pronunciations of the word "but" extracted from the same set of material. Tables B.1 and B.2 compare various pronunciations of the word "but" to the canonical pronunciation, [b ah t], as well as display a summary of deviation-from-canonical statistics (partitioned by syllable position). Tables B.3-B.8 further partition the pronunciation variants of the word "but" by stress accent and display the associated deviation-from-canonical statistics. Note that the onset segment of the word "but" is generally pronounced far more canonically than the onset segment of the word "that" (which is representative of words in the Switchboard corpus).

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
b ah t	35	-	-	-	ah t	1	D	-	-
b ah	24	-	-	D	ax dx	1	D	S	S
b ax t	14	-	S	-	ax t	1	D	S	-
b ah dx	8	-	-	S	b ih d	1	-	S	S
b ax	8	-	S	D	b iy	1	-	S	D
b ax dx	7	-	S	S	b ow	1	-	S	D
ah dx	2	D	-	S	m ah t	1	S	-	-
ax	2	D	S	D	v ah	1	S	-	D
b ax q	2	-	S	S	v ah dx	1	S	-	S
ah	1	D	-	D	v ah t	1	S	-	-

Table B.1: Pronunciation variants of the word "but" found in the Year-2001 phonetic evaluation material. For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([b ah t]) with respect to syllable onset, nucleus and coda. "S" is substitution, "D" is deletion, "I" is insertion and "-" is no deviation.

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	89.4	3.5	7.1	0	113
Nucleus	66.4	33.6	0	0	113
Coda	46.9	19.5	33.6	0	113
Overall	67.6	18.9	13.6	0	339

Table B.2: Summary of phonetic deviation (from canonical) in terms of percentage of total segments (last column) for each syllable position (and overall), for the word "but" (cf. Table B.1) from the Year-2001 phonetic evaluation material.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
b ax t	12	-	S	-	ax	2	D	S	D
b ah	8	-	-	D	b ax q	2	-	S	S
b ax	7	-	S	D	ah t	1	D	-	-
b ah t	5	-	-	-	b ah dx	1	-	-	S
b ax dx	4	-	S	S	b iy	1	-	S	D

Table B.3: Unaccented instances of the word "but" from the Year-2001 phonetic evaluation material. For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([b ah t]) with respect to syllable onset, nucleus and coda. "S" is substitution, "D" is deletion, "I" is insertion and "-" is no deviation.

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	93.0	0	7.0	0	43
Nucleus	34.9	65.1	0	0	43
Coda	41.9	16.3	41.9	0	43
Total	56.6	27.1	16.3	0	129

Table B.4: Summary of phonetic deviation (from canonical) in terms of percentage of total segments (last column) for each syllable position (and overall), for the unaccented instances of the word "but" (cf. Table B.3) from the Year-2001 phonetic evaluation material.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
b ah t	27	-	-	-	ax t	1	D	S	-
b ah	15	-	-	D	b ax	1	-	S	D
b ah dx	6	-	-	S	b ih d	1	-	S	S
b ax dx	3	-	S	S	b ow	1	-	S	D
ah dx	2	D	-	S	m ah t	1	S	-	-
b ax t	2	-	S	-	v ah	1	S	-	D
ah	1	D	-	D	v ah dx	1	S	-	S
ax dx	1	D	S	S					

Table B.5: Lightly accented instances of the word "but" from the Year-2001 phonetic evaluation material. For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([b ah t]) with respect to syllable onset, nucleus and coda. "S" is substitution, "D" is deletion, "I" is insertion and "-" is no deviation.

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	87.5	4.7	7.8	0	64
Nucleus	84.4	15.6	0	0	64
Coda	48.4	21.9	29.7	0	64
Total	73.4	14.1	12.5	0	192

Table B.6: Summary of the phonetic deviation (from canonical), in terms of percentage of total segments (last column) in each syllable position (and overall), for the lightly accented instances of the word "but" (cf. Table B.5) from the Year-2001 phonetic evaluation material.

Pron.	#	Deviation Type			Pron.	#	Deviation Type		
		Onset	Nuc.	Coda			Onset	Nuc.	Coda
b ah t	3	-	-	-	b ah dx	1	-	-	S
b ah	1	-	-	D	v ah t	1	S	-	-

Table B.7: Fully accented instances of the word "but" from the Year-2001 phonetic evaluation material. For each pronunciation the number of occurrences (#) is given, as well as the pattern of deviation from the canonical pronunciation ([b ah t]) with respect to syllable onset, nucleus and coda. "S" is substitution, "D" is deletion, "I" is insertion and "-" is no deviation.

Syllable Position	Deviations from Canonical				
	Canonical%	Substitution%	Deletion%	Insertion%	Total
Onset	100.0	0	0	0	6
Nucleus	100.0	0	0	0	6
Coda	50.0	16.7	33.3	0	6
Total	83.3	5.6	11.1	0	18

Table B.8: Summary of the phonetic deviation (from canonical), in terms of percentage of total segments (last column) in each syllable position (and overall), for the fully accented instances of the word "but" (cf. Table B.7) from the Year-2001 phonetic evaluation material.

Appendix C

Learning Fuzzy Measures

This appendix describes a supervised, gradient-based algorithm for learning the fuzzy measures used in combining AF-dimension matching scores into a syllable-level score (cf. Section 6.3.5), similar in spirit to an algorithm of fuzzy measure learning introduced by Grabisch and Nicholas [47][44]. The following sections first provide a formulation of the learning problem, then describe the algorithm in steps, followed by derivation of the parameter update equations under two different error criteria.

C.1 Formulation of the Problem

For the purpose of learning the fuzzy measures, let us re-cast the combining of AF-dimension matching scores into a single syllable score as a classification problem. Suppose there are M reference syllable classes, N AF dimensions under consideration and D data points (input syllables) in the training set. Following the notation from Chapter 6, let $X = \{x_1, \cdots, x_N\}$ denote the set of N AF dimensions (each as a separate information source). For the d^{th} input syllable data point, let $H_d^m = \{h_d^m(x_i), i = 1, \cdots, N\}$ be the set of the matching scores provided by the AF dimensions for the m^{th} reference syllable, and $y_d^m = f^m(H_d^m)$ the combined score for reference syllable class m, where $f^m()$ is the overall combining function for reference syllable class m (with its parameters). A well-known result in pattern recognition [29] is that the minimum expected classification error is obtained if one always selects the winning class according to the maximum of the posterior probabilities of the classes given the input. In [108], Richard and Lippmann showed that a discriminant function trained with one-of-M target outputs (i.e., one for the correct class and zeros for the others) approximates the posterior probabilities of the classes using either a minimum-squared-error (MSE) or a minimum-cross-entropy (MCE) error criterion. Thus, in the current implementation, the target outputs for the combining function adopt the one-of-M scheme such that for each input data point d, the target output for class m is $t_d^m = 1$ if the input "belongs" to class m and $t_d^m = 0$ otherwise. The total error for the MCE criterion is

$$E^{MCE} = \sum_{d=1}^{D} E_d^{MCE} = \sum_{d=1}^{D} \sum_{m=1}^{M} t_d^m \log \frac{t_d^m}{y_d^m} \tag{C.1}$$

and the total error for the MSE criterion is

$$E^{MSE} = \sum_{d=1}^{D} E_d^{MSE} = \sum_{d=1}^{D} \sum_{m=1}^{M} (y_d^m - t_d^m)^2. \tag{C.2}$$

In the formulation so far, the combining function $f^m()$ is not constrained to any specific form. In the subsequent description, let $f^m()$ take the following form:

$$f^m(H_d^m) = \frac{\exp(\beta C^m(H_d^m))}{\sum_{j=1}^{M} \exp(\beta C^j(H_d^j))} \tag{C.3}$$

where $C^j(H_d^j) = \int_C H_d^j \circ g^j$ is the (Choquet) fuzzy integral (cf. Definition 5 in Section 6.3.5) of the AF dimension matching scores for reference syllable j with respect to the fuzzy measure g^j. The β is a scaling factor and the exponential softmax form in (C.3) serves two purposes: (1) to provide a normalization for the fuzzy integral outputs so that they are proper probability terms, (2) to adjust the sharpness of the estimated posterior probability distribution where a large β represents a narrow distribution concentrating on the maximum class output and a small β is associated with a broad distribution. For the current implementation system performance is not very sensitive to the choice of the scaling factor, β, over a large numerical range. β was set to 10 using trial-and-error methods for the experiments described in Chapters 6 and 7. Note that the softmax transformation is only used for learning the fuzzy measures and it is actually $C^j(H^j)$ terms that are used during the recognition. Thus, the learning problem is essentially to find the optimal fuzzy measures, g, minimizing the error criteria E^{MCE} or E^{MSE} given the input AF dimension matching scores H and the desired target outputs t (in one-of-M representation).

C.2 Algorithm Description

Before presenting the detailed steps of the algorithm, a description to a lattice representation of a fuzzy measure is in order, following the original introduction by Grabisch and Nicholas [47]. For a fuzzy measure g on an N-dimension set $X = x_1, \cdots, x_N$, there are a total of 2^N fuzzy measure parameters and they can be arranged in a lattice with the usual ordering of real numbers. The main purpose of the lattice is to show the monotonicity of the fuzzy measure parameters and the particular values involved in a fuzzy integral evaluation. Figure C.1 shows a lattice representation of a fuzzy measure, with $N = 4$, i.e., $X = x_1, x_2, x_3, x_4$ (adopted from [47]).

The lattice contains $N+1$ layers (referred to as layer $0, \cdots, N$ from top to bottom) of nodes. Each node represents the fuzzy measure of a particular subset of X (for simplicity, g_{23} denotes $g(\{x_2, x_3\})$ and similarly for others) and it is assumed $g(\phi) = 0$ and $g(X) = 1$. Two nodes in adjacent layers are connected only if there is a set-inclusion relationship between the two subsets of X whose measures they represent. A node in layer l thus has l connected nodes in layer $l-1$ and $N-l$ connected nodes in layer $l+1$. By monotonicity, for any pair of nodes that are directly connected, the fuzzy measure of the node in the upper layer is less than or equal to the fuzzy measure of the node in the lower layer. For one fuzzy

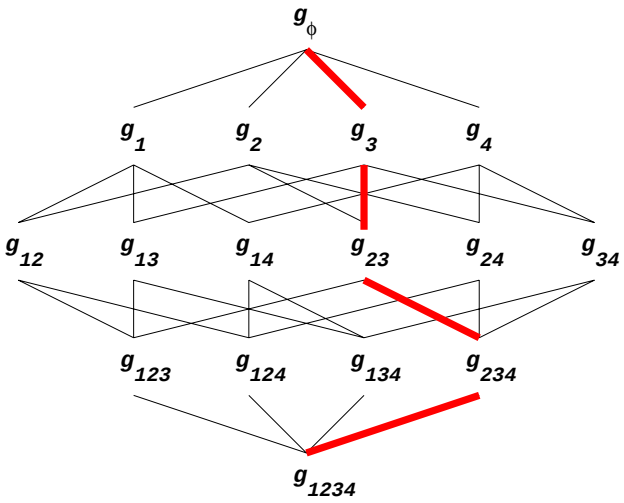

Figure C.1: An example of a lattice representation of a fuzzy measure, with $N = 4$. The path $< g_\phi, g_3, g_{23}, g_{234}, g_{1234} >$ is highlighted. (Adapted from [47].)

integral evaluation of $H = h(x_1), \cdots, h(x_N)$ with respect to g, only one path from $g(\phi)$ to $g(X)$ is involved, depending on the order of H. For example, if $h(x_1) \leq h(x_4) \leq h(x_2) \leq h(x_3)$, the path involved in the fuzzy integral evaluation contains nodes g_ϕ, g_3, g_{23}, g_{234} and g_{1234} (cf. the highlighted path in Figure C.1).

With the lattice representation of the fuzzy measure, we are now ready to describe the detailed steps of the learning algorithm. The algorithm is similar in spirit to that described in [44], with some differences in the detailed update equations due to the different normalization and error criteria adopted. The major steps of the algorithm are:

- Step 0 – initialization: each of the fuzzy measures g^m for $m = 1, \cdots, M$ is initialized at the so-called equilibrium state [44], i.e., $g^m(\{x_i\}) = 1/N$ for all $i = 1, \cdots, N$ and the g^m is additive (i.e., $g^m(A \cup B) = g^m(A) + g^m(B)$ $\forall A \subseteq X, B \subseteq X$ *and* $A \cap B = \phi$). With respect to this initial fuzzy measure, the (Choquet) fuzzy integral reduces to simply computing the arithmetic mean – a least committing averaging operator in the absence of any information.

- Step 1 – computing the fuzzy integral: for an input-output learning data pair ($H = \{H^m,\ m = 1, \cdots, M\}$, $t = \{t^m,\ m = 1, \cdots, M\}$), evaluate the fuzzy integrals and the softmax normalization to obtain $y^m = f^m(H^m)$ for $m = 1, \cdots, M$.

- Step 2 – updating fuzzy measures: for each of the $m = 1, \cdots, M$, do the following:

  - Step 2.1: compute the output error $\epsilon^m = y^m - t^m$.
  - Step 2.2: let $g^m_{(0)}, g^m_{(1)}, \cdots, g^m_{(N)}$ denote the nodes on the path involved in the fuzzy integral evaluation of y^m in the order from g_ϕ to $g_{1,\cdots,N}$ where the parenthesis "()" around a subscript indicates its being ordered. (Of course, this order is determined by the input matching scores such that $h^m(x_{(N)}) \leq h^m(x_{(N-1)}) \leq$

$\cdots \leq h^m(x_{(1)})$. For example, for the highlighted path in Figure C.1 $g^m_{(0)} = g_\phi = 0$, $g^m_{(1)} = g_3$, $g^m_{(2)} = g_{23}$, $g^m_{(3)} = g_{234}$ and $g^m_{(4)} = g_{1234} = 1$.)
For each $i = 1, \cdots, N-1$, compute $\delta_{(i)} = h^m(x_{(N+1-i)}) - h^m(x_{(N-i)})$ and then update each $g^m_{(i)}$ by

$$g^{m^{new}}_{(i)} = g^{m^{old}}_{(i)} - \alpha\beta\epsilon^m\delta_{(i)} \tag{C.4}$$

for the MCE error criterion and by

$$g^{m^{new}}_{(i)} = g^{m^{old}}_{(i)} - 2\alpha\beta y^m[\epsilon^m - \sum_{q=1}^{M}\epsilon^q y^q]\delta_{(i)} \tag{C.5}$$

for the MSE error criterion and in both equations α is a learning rate parameter, which can be decreased over the course of learning to obtain better convergence. Note that for the MSE error criterion Step 2.1 has to be completed for all m first.

- Step 2.3: verify the monotonicity relations to ensure that each g^m is still a proper fuzzy measure. For each $g^m_{(i)}$ updated in the previous step, its value is compared to its neighboring (connected) nodes in the layers above or below. If a violation of monotonicity occurs between $g^m_{(i)}$ and some node, $g^m_{(i)}$ is set to the value of that node, which results in the least amount of correction required. The verification should be carried out in the order from $g^m_{(1)}$ to $g^m_{(N-1)}$ if $\epsilon^m > 0$, and in the reverse order if $\epsilon^m < 0$.

For each training epoch, Steps 1 and 2 are repeated for all training data points. Several training epochs can be carried out and a separate cross-validation data set should be used to determine the total number of epochs to avoid over-training. In [44], an extra step was described to smooth those fuzzy measures that have not been updated previously due to the scarcity of training data. However, because of the large amount of data in our experiments, this additional step was not necessary.

C.3 Derivation of Parameter Update Equations

This section shows the derivation of the update equations (C.4) and (C.5). Material pertaining to the MCE criterion is presented first, followed by analogous material pertaining to the MSE criterion where steps common to both criteria are not repeated. To make the derivations easier to follow, a description of the symbols used is listed here:

Symbol	Description
E	total error on the training data, equivalent to E^{MCE} for minimum-cross-entropy criterion and E^{MSE} for minimum-sum-of-squared-error criterion
E^{MCE}	total MCE error on the training data
E^{MSE}	total MSE error on the training data

E_d	error on the d^{th} data point
g^k	fuzzy measure for the syllable class k
$g^k_{(i)}$	i^{th} ordered fuzzy measure parameter for syllable class, k, on a particular path determined by the input, H^k
H^k	the vector of N AF-dimension matching scores for syllable class k
$h^k(x_{(i)})$	the i^{th} ordered value in H^k
C^k	the fuzzy integral function with respect to fuzzy measure g^k for syllable class k
y^k	the output value for class k (the result of the softmax normalization of C^k as in Equation (C.3))
β	the scaling factor used in the softmax normalization in Equation (C.3)
t^k	the desired (target) output for syllable class k

C.3.1 Cross-Entropy Error

Following the problem formulation and notations from the previous sections, the goal of the learning problem is to find the optimal fuzzy measure g^{k^*} for each syllable class k such that

$$g^{k^*} = argmin_{g^k} E \tag{C.6}$$

where

$$E = E^{MCE} = \sum_{d=1}^{D} \sum_{m=1}^{M} t_d^m \log \frac{t_d^m}{y_d^m}$$

for the MCE criterion given by Equation (C.1). Thus the iterative, gradient-descent learning algorithm requires the computation of the gradient $\partial E / \partial g^k_{(i)}$ for each i and k. Since the algorithm is online (i.e., parameters are updated after each data point is presented), we will only consider the gradient computed for each data point (i.e., $\partial E_d / \partial g^k_{(i)}$ for some d) in the subsequent derivations. However, for convenience of notation, the subscript d is omitted from all symbols except E_d without confusion.

For each $g^k_{(i)}$, the partial derivative of E_d can be computed as

$$\frac{\partial E_d}{\partial g^k(i)} = \frac{\partial E_d}{\partial C^k(H^k)} \cdot \frac{\partial C^k(H^k)}{\partial g^k_{(i)}}. \tag{C.7}$$

From the definition of fuzzy integral (cf. Equation (6.4)), we have

$$C^k(H^k) = \sum_{i=1}^{N} [h^k(x_{(N+1-i)}) - h^k(x_{(N-i)})] g^k_{(i)} \quad with \quad h^k(x_{(0)}) = 0. \tag{C.8}$$

The second term on the right-hand side of Equation (C.7) is thus

$$\frac{\partial C^k(H^k)}{\partial g^k_{(i)}} = h^k(x_{(N+1-i)}) - h^k(x_{(N-i)}). \tag{C.9}$$

The first term on the right-hand side of Equation (C.7) is evaluated as:

$$\frac{\partial E_d}{\partial C^k(H^k)} = \sum_{j=1}^{M} \frac{\partial E_d}{\partial y^j} \cdot \frac{\partial y^j}{\partial C^k(H^k)} \tag{C.10}$$

where, from Equation (C.3),

$$y^j = f^j(H^j) = \frac{\exp(\beta C^j(H^j))}{\sum_{q=1}^{M} \exp(\beta C^q(H^q))}.$$

Therefore, we get

$$\frac{\partial y^j}{\partial C^k(H^k)} = \begin{cases} \frac{\beta \exp(\beta C^k)}{\sum_{q=1}^{M} \exp(\beta C^q)} + \frac{-\beta \exp(\beta C^k) \exp(\beta C^k)}{(\sum_{q=1}^{M} \exp(\beta C^q))^2} = \beta y^k(1-y^k), & for j = k \\ \exp(\beta C^j) \cdot \frac{-\beta \exp(\beta C^k)}{(\sum_{q=1}^{M} \exp(\beta C^q))^2} = -\beta y^j y^k, & for j \neq k \end{cases} \tag{C.11}$$

From the definition of the MCE criterion, we get:

$$\begin{aligned} \frac{\partial E_d}{\partial y^j} &= \frac{\partial \sum_{q=1}^{M} t^q \log \frac{t^q}{y^q}}{\partial y^j} \\ &= t^j \cdot \frac{y^j}{t^j} \cdot \frac{-t^j}{(y^j)^2} \\ &= -\frac{t^j}{y^j} \end{aligned} \tag{C.12}$$

Using Equations (C.11) and (C.12), Equation (C.10) becomes

$$\begin{aligned} \frac{\partial E_d}{\partial C^k(H^k)} &= -\frac{t^k}{y^k} \cdot \beta y^k(1-y^k) + \sum_{j \neq k} (-\frac{t^j}{y^j})(-\beta y^j y^k) \\ &= -t^k \cdot \beta + t^k \cdot \beta y^k + \beta \sum_{j \neq k} t^j y^k \\ &= -t^k \cdot \beta + \beta y^k \underbrace{(t^k + \sum_{j \neq k} t^j)}_{=1} \\ &= \beta(y^k - t^k) \end{aligned} \tag{C.13}$$

where the last step uses the fact that the desired outputs are in one-of-M form. Putting together Equations (C.7), (C.9) and (C.13), we get the gradient term

$$\frac{\partial E_d}{\partial g^k(i)} = \beta(y^k - t^k)(h^k(x_{(N+1-i)}) - h^k(x_{(N-i)})) \tag{C.14}$$

and hence the update equation (C.4).

C.3.2 Sum-of-Squares Error

When the MSE criterion is used, the goal of the learning problem is to find the optimal fuzzy measure g^{k^*} for each syllable class k such that

$$g^{k^*} = argmin_{g^k} E \tag{C.15}$$

where

$$E = E^{MSE} = \sum_{d=1}^{D} E_d^{MSE} = \sum_{d=1}^{D} \sum_{m=1}^{M} (y_d^m - t_d^m)^2.$$

Again since the algorithm is online, as for the MCE criterion, we will only consider the gradient computed for each data point (i.e., $\partial E_d / \partial g_{(i)}^k$ for some d) in the subsequent derivations. Now for the MSE criterion, Equations (C.7) through (C.11) remain identical to that of the MCE criterion. However, from the definition of the MSE criterion, instead of Equation (C.12) we get:

$$\begin{aligned} \frac{\partial E_d}{\partial y^j} &= \frac{\partial \sum_{q=1}^{M} (y^q - t^q)^2}{\partial y^j} \\ &= 2(y^j - t^j) \end{aligned} \tag{C.16}$$

Now with Equations (C.11) and (C.16), Equation (C.10) for the MSE criterion becomes

$$\begin{aligned} \frac{\partial E_d}{\partial C^k(H^k)} &= 2(y^k - t^k) \cdot \beta y^k (1 - y^k) - \sum_{j \neq k} 2(y^j - t^j) \cdot \beta y^k y^j \\ &= 2\beta y^k [(y^k - t^k)(1 - y^k) - \sum_{j \neq k} (y^j - t^j) y^j] \qquad \text{(C.17)} \\ &= 2\beta y^k [(y^k - t^k) - \sum_{j=1}^{M} (y^j - t^j) y^j]. \qquad \text{(C.18)} \end{aligned}$$

Finally, putting together Equations (C.7), (C.9) and (C.18), we get the gradient term for the MSE criterion

$$\frac{\partial E_d}{\partial g^k(i)} = 2\beta y^k [(y^k - t^k) - \sum_{j=1}^{M} (y^j - t^j) y^j] \cdot [h^k(x_{(N+1-i)}) - h^k(x_{(N-i)})] \tag{C.19}$$

and hence the update equation (C.5).

C.3.3 Parameter Sharing

The descriptions so far assume that each syllable class $m = 1, \cdots, M$ has its separate fuzzy measure g^m in the most general case. However, this assumption is not necessary and different syllable classes can certainly share the same fuzzy measure. For example, in the extreme case, only one fuzzy measure is shared among all of the M syllable classes. The reduced number of parameters leads to less risk of over-training and can be

more easily interpreted. Interestingly, in case of such sharing of the fuzzy measure, the algorithm and the parameter update equations described above are still valid, as shown in the simple proof for the one-fuzzy-measure-for-all case (which can also be extended to other cases of fuzzy-measure sharing).

Suppose $g^1 = g^2 = \cdots = g^M = r$. Let us denote the error criterion (either MCE or MSE) by $E(g^1, \cdots, g^M)$ to make its dependency on the fuzzy measures explicit. The partial derivative of E with respect to r is then

$$\begin{aligned} \frac{\partial E(g^1, \cdots, g^M)}{\partial r} &= \frac{\partial E(g^1, \cdots, g^M)}{\partial g^1} \cdot \frac{\partial g^1}{\partial r} + \cdots \frac{\partial E(g^1, \cdots, g^M)}{\partial g^M} \cdot \frac{\partial g^M}{\partial r} \\ &= \sum_{i=1}^{M} \frac{\partial E(g^1, \cdots, g^M)}{\partial g^i} \end{aligned} \tag{C.20}$$

since $\frac{\partial g^i}{\partial r} = 1$ for all $i = 1, \cdots, M$. This shows that the gradient of the error criterion with respect to the shared fuzzy measure is the same as the sum of the gradients with respect to the fuzzy measures considered individually. Thus, the algorithm described above also achieves the appropriate learning effect for the shared fuzzy measure.

Bibliography

[1] T. Arai and S. Greenberg. The temporal properties of spoken Japanese are similar to those of English. In *Proceedings of the 5th Eurospeech Conference on Speech Communication and Technology (Eurospeech-97)*, pages 1011–1014, Rhodes, Greece, September 1997.

[2] T. Arai and S. Greenberg. Speech intelligibility in the presence of cross-channel spectral asynchrony. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, pages 933–936, Seattle, WA, May 1998.

[3] T. Arai, H. Hermansky, M. Pavel, and C. Avendano. Intelligibility of speech with filtered time trajectories of spectral envelope. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-96)*, pages 2489–2492, Atlanta, GA, May 1996.

[4] The Aurora Evaluation of Distributed Speech Recognition Systems. http://www.isip.msstate.edu/projects/aurora/.

[5] M. Beckman. *Stress and Non-Stress Accent.* Foris Publications, Dordrecht, Holland, 1986.

[6] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1996.

[7] W. Bolton. *A Living Language: the History and Structure of English.* Random House, New York, 1981.

[8] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, Boston, MA, 1993.

[9] E. Carlson and D. Miller. Aspects of voice quality: Display, measurement and therapy. *International Journal of Language and Communication Disorders*, (33):suppl., 1998.

[10] J. Carson-Berndsen and M. Walsh. Defining constraints for multilinear speech processing. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 2281–2284, Aalborg, Denmark, September 2001.

[11] Dept. of Computer Science Center for Spoken Language Understanding and Oregon Graduate Institute Engineering. OGI Stories corpus, Release 1.0, 1995.

[12] Dept. of Computer Science Center for Spoken Language Understanding and Oregon Graduate Institute Engineering. Numbers corpus, Release 1.0. (Numbers95), 1995.

[13] S. Chang, S. Greenberg, and M. Wester. An elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 1725–1728, Aalborg, Denmark, September 2001.

[14] S. Chang, L. Shastri, and S. Greenberg. Automatic phonetic transcription of spontaneous speech (American English). In *Proceedings of the International Conference on Spoken Language Processing*, pages 330–333, Beijing, China, October 2000.

[15] S. Chang, L. Shastri, and S. Greenberg. Robust phonetic feature extraction under a wide range of noise backgrounds and signal-to-noise ratios. In *Proceedings of the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark, September 2001.

[16] S. Chang, L. Shastri, and S. Greenberg. Robust phonetic feature extraction under a wide range of noise backgrounds and signal-to-noise ratios. *submitted to Speech Communication Special Issue on Recognition and Organization of Real-World Sound*, 2002.

[17] S. Chang, M. Wester, and S. Greenberg. An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *submitted to Computer Speech and Language*, 2002.

[18] S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, February 1998.

[19] S. Cho and J. Kim. Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Transaction on Systems, Man, Cybernetics*, (2):280–384, 1995.

[20] J. Clark and C. Yallup. *Introduction to Phonology and Phonetics.* Blackwell, Oxford, UK, 1990.

[21] L. Comerford, D. Frank, P. Gopalakrishnan, R. Gopinath, and J. Sedivy. The IBM personal speech assistant. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-01)*, Salt Lake City, UT, May 2001.

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

[23] L. Deng and K. Erler. Structural design of hidden markov model speech recognizer using multivalued phonetic features: Comparison with segmental speech units. *Journal of Acoustic Society of America*, 92(6):3058–3066, June 1992.

[24] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein. Large vocabulary word recognition using context-dependent allophonic Hidden Markov Models. *Computer Speech and Language*, 4:345–357, 1990.

[25] L. Deng, G. Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. *Speech Communication*, 22(2):93–112, August 1997.

[26] L. Deng and D. Sun. Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-94)*, pages 45–48, Adelaide, Australia, April 1994.

[27] L. Deng and D. Sun. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of Acoustic Society of America*, 95(5):2702–2719, May 1994.

[28] L. Deng, K. Wang, A. Acero, H.W. Hon, J. Droppo, C. Boulis, Y.Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X.D. Huang. Distributed speech processing in MiPad's multimodal user interface. *IEEE Transactions on Speech and Audio Processing, to appear*, 2002.

[29] R. Duda and P. Hart. *Pattern classification and scene analysis.* John Wiley and Sons, New York, 1973.

[30] E. Dupoux. The time course of prelexical processing: The syllabic hypothesis revisited. In G. Altmann and R. Shillcock, editors, *Cognitive Models of Speech Processing – the Second Sperlonga Meeting*, pages 81–113. Lawrence Erlbaum Associates, Hove, UK, 1993.

[31] D. Ellis. SYLLIFY. In-house software at ICSI. Tcl/TK interface for Fisher's TSYLB2 program.

[32] W. Feller. *Introduction to the Probability Theory and Its applications*, volume 2. John Wiley and Sons, New York, 3rd edition, 1968.

[33] R. Drullman J. Festen and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of Acoustic Society of America*, 95:2670–2680, 1994.

[34] R. Drullman J. Festen and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of Acoustic Society of America*, 95(4):1053–1064, 1994.

[35] T. Fontaine and L. Shastri. A hybrid system for handprinted word recognition. In *Proceedings of the 9th Conference on Artificial Intelligence for Applications (CAIA'93)*, pages 227–234, Los Alamitos, CA, March 1993. IEEE Computer Society Press.

[36] E. Fosler-Lussier. *Dynamic Pronunciation Models for Automatic Speech Recognition.* PhD dissertation, Department of EECS, University of California, Berkeley, 1999.

[37] E. Fosler-Lussier, S. Greenberg, and N. Morgan. Incorporating contextual phonetics into automatic speech recognition. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, pages 611–614, San Francisco, CA, August 1999.

[38] J. Frankel, K. Richmond, S. King, and P. Taylor. An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proceedings of the International Conference on Spoken Language Processing*, pages 145–148, Beijing, China, October 2000.

[39] J. Fritsch. *Hierarchical Connectionist Acoustic Modeling for Domain-Adaptive Large Vocabulary Speech Recognition.* PhD dissertation, University of Karlsruhe, October 1999.

[40] A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchoff, M. Ordowski, and B. Wheatley. Syllable – A promising recognition unit for LVCSR. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 207–214, Santa Barbara, California, December 1997.

[41] G. Gimson. *Introduction to the Pronunciation of English.* Edward Arnold, London, UK, 3rd edition, 1980.

[42] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-92)*, pages 517–520, San Francisco, CA, March 1992.

[43] M. Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, (69):279–298, 1994.

[44] M. Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *Proceedings of the 4th IEEE International Conference on Fuzzy Systems and the 2nd International Fuzzy Engineering Symposium*, pages 145–150, Yokohama, Japan, 1995.

[45] M. Grabisch. k-order additive fuzzy measures. In *Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 1345–1350, Granada, Spain, 1996.

[46] M. Grabisch. Fuzzy integral for classification and feature extraction. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 415–434. Physica-Verlag, New York, 2000.

[47] M. Grabisch and J. Nicolas. Classification by fuzzy integral: Performance and tests. *Fuzzy Sets and Systems*, (65):225–271, 1994.

[48] M. Grabisch and M. Roubens. Application of the Choquet integral in multicriteria decision making. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 348–373. Physica-Verlag, New York, 2000.

[49] S. Greenberg. The Switchboard transcription project research report #24. In *Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*, Baltimore, MD, 1996. Center for Language and Speech Processing, Johns Hopkins University.

[50] S. Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-a-Mousson, France, April 1997.

[51] S. Greenberg. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–176, 1999.

[52] S. Greenberg. From here to utility – Melding phonetic insight with speech technology. In W. Barry and W. Domelen, editors, *to appear in Integrating Phonetic Knowledge with Speech Technology*. Kluwer Press, Boston, 2002.

[53] S. Greenberg. Understanding spoken language using statistical and computational methods. Presentation at Workshop on Patterns of Speech Sounds in Unscripted Communication – Production, Perception, Phonology (http://www.icsi.berkeley.edu/~ steveng), October 8-11, 2000.

[54] S. Greenberg and T. Arai. The relation between speech intelligibility and the complex modulation spectrum. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 473–476, Aalborg, Denmark, September 2001.

[55] S. Greenberg, H.M. Carvey, and L. Hitchcock. The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *Proceedings of ISCA Workshop on Prosody in Speech Processing (Speech Prosody 2002)*, Aix-en-Provence, April 2002.

[56] S. Greenberg, H.M. Carvey, L. Hitchcock, and S. Chang. Beyond the phoneme: A juncture-accent model of spoken language. In *Proceedings of Human Language Technology Conference (to be published)*, San Diego, March 2002.

[57] S. Greenberg and S. Chang. Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pages 195–202, Paris, France, September 2000.

[58] S. Greenberg, S. Chang, and L. Hitchcock. The relation between stress accent and vocalic identity in spontaneous American English discourse. In *Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 51–56, Red Bank, NJ, Oct. 2001.

[59] S. Greenberg, S. Chang, and J. Hollenback. An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems. In *Proceedings of the NIST Speech Transcription Workshop*, College Park, MD, May 16-19 2000.

[60] S. Greenberg and B. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, pages 1647–1650, Munich, April 1997.

[61] H. Hermansky. Perceptual linear predictive (PLP) analysis for speech. *Journal of Acoustic Society of America*, (4):1738–1752, 1990.

[62] H. Hermansky and S. Sharma. Temporal patterns (TRAPS) in ASR of noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-99)*, pages 289–292, Phoenix, Arizona, May 1999.

[63] L. Hitchcock and S. Greenberg. Vowel height is intimately associated with stress accent in spontaneous American English discourse. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 79–82, Aalborg, Denmark, September 2001.

[64] J. Hogden, I. Zlokarnik, A. Lofqvist, V. Gracco, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. *Journal of Acoustic Society of America*, (3):1819–1834, 1996.

[65] C. Jankowski, A. Kalyanswamy, S. Basson, , and J. Spitz. NTIMIT: A phonetically balanced, continuous speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-90)*, pages 109–112, Albuquerque, NM, April 1990.

[66] F. Jelinek. *Statistical Methods for speech recognition.* MIT Press, Cambridge, MA, 1997.

[67] R. Jones, S. Downey, and J. Mason. Continuous speech recognition using syllables. In *Proceedings of the 5th Eurospeech Conference on Speech Communication and Technology (Eurospeech-97)*, pages 1171–1174, Rhodes, Greece, September 1997.

[68] D. Jurafsky, W. Ward, J. Zhang, K. Herold, X. Yu, and S. Zhang. What kind of pronunciation variation is hard for triphones to model? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-01)*, pages 577–580, Salt Lake City, UT, May 2001.

[69] M. Kean. *The theory of markedness in generative grammar.* PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1975.

[70] P. Keating. Coronal places of articulation. In C. Paradis and J. Prunet, editors, *Phonetics and Phonology: the special status of coronals internal and external evidence*, pages 29–48. Academic Press, San Diego, California, 1991.

[71] J. Keller, P. Gader, and A. Hocaoglu. Fuzzy integrals in image processing and recognition. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 435–466. Physica-Verlag, New York, 2000.

[72] J. Kessens, M. Wester, , and H. Strik. Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation. *Speech Communication*, (2):193–207, 1999.

[73] S. King and P. Taylor. Detection of phonological features in continuous speech using neural networks. *Computer, Speech and Language*, (4):333–345, 2000.

[74] B. Kingsbury. *Perceptually Inspired Signal-Processing Strategies for Robust Speech Recognition in Reverberant Environments.* PhD dissertation, Department of EECS, University of California, Berkeley, 1998.

[75] B. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25:117–132, 1998.

[76] K. Kirchhoff. *Robust Speech Recognition Using Articulatory Information.* PhD dissertation, University of Bielefeld, June 1999.

[77] Y. Konig and N. Morgan. Modeling dynamics in connectionist speech recognition - the time index model. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-94)*, pages 1523–1526, Yokohama, Japan, September 1994.

[78] P. Ladefoged. *A Course in Phonetics.* Harcourt Brace, Fort Worth, TX, 3rd edition, 1993.

[79] L. Lamel, J. Garafolo, J. Fiscus, W. Fisher, and D. Pallett. *TIMIT: The DARPA acoustic-phonetic speech corpus.* National Technical Information Service Publication Publication PB91-505065INC, 1990.

[80] C.H. Lee, B.H. Juang, F.K. Soong, and L. Rabiner. Word recognition using whole word and subword models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-89)*, pages 683–686, Glasgow, UK, 1989.

[81] L. Lee. Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine*, (97):63–101, 1997.

[82] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation. *Computer Speech and Language*, 9:171–185, 1995.

[83] I. Lehiste. Suprasegmental features of speech. In N. Lass, editor, *Principles of Experimental Phonetics*, pages 226–244. Mosby, St. Louis, MO, 1996.

[84] P. Lieberman and S. Blumstein. *Speech physiology, speech perception, and acoustic phonetics.* Cambridge University Press, Cambridge, UK, 1988.

[85] M. Lindau. The story of /r/. In V. Fromkin, editor, *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, pages 157–168. Academic Press, Orlando, FL, 1985.

[86] R. Lippmann. Speech recognition by machines and humans. *Speech Communication*, (22):1–15, 1997.

[87] D. McAllaster, L. Gillick, F. Scattone, and M. Newman. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*, pages 1847–1850, Sydney, Australia, November 1998.

[88] N. Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition.* PhD dissertation, Department of EECS, University of California, Berkeley, 1998.

[89] N. Mirghafori, E. Fosler, and N. Morgan. Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes. In *Proceedings of the 4th Eurospeech Conference on Speech Communication and Technology (Eurospeech-95)*, pages 491–494, Madrid, Spain, September 1995.

[90] N. Morgan and H. Bourlard. An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, pages 24–42, May 1995.

[91] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proceedings of the 5th Eurospeech Conference on Speech Communication and Technology (Eurospeech-97)*, pages 2079–2083, Rhodes, Greece, September 1997.

[92] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, pages 729–732, Seattle, WA, May 1998.

[93] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (III): interaction index. In *Proceedings of the 9th Fuzzy System Symposium*, pages 693–696, Sapporo, Japan, 1993.

[94] T. Murofushi and M. Sugeno. An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure. *Fuzzy Sets and Systems*, (29):202–227, 1989.

[95] T. Murofushi and M. Sugeno. The Choquet integral in multiattribute decision making. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 333–347. Physica-Verlag, New York, 2000.

[96] T. Murofushi and M. Sugeno. Fuzzy measures and fuzzy integrals. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 3–41. Physica-Verlag, New York, 2000.

[97] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu. Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition. In *Proceedings of the International Workshop on Automatic Speech Recognition and Understanding*, pages 197–200, Keystone, CO, December 1999.

[98] National Institute of Standard and Technology (NIST). Benchmark tests. http://www.nist.gov/speech/tests.

[99] National Institute of Standard and Technology (NIST). SC-Lite - speech recognition scoring program. http://www.nist.gov/speech/tools/index.htm.

[100] M. Ostendorf. Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of the International Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO, December 1999.

[101] S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 15(4):263–322, 2000.

[102] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of Acoustic Society of America*, (2):688–700, 1992.

[103] C. Paradis and J. Prunet. Asymmetry and visibility in consonant articulation. In C. Paradis and J. Prunet, editors, *Phonetics and Phonology: the special status of coronals internal and external evidence*, pages 1–28. Academic Press, San Diego, California, 1991.

[104] J. Peters, L. Han, and S. ramanna. The Choquet integral in a rough software cost decision system. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals:Theory and Applications*, Studies in fuzziness and soft computing, pages 392–414. Physica-Verlag, New York, 2000.

[105] T. Pham and H. Yan. Combination of handwritten-numeral classifiers with fuzzy integral. In C. Leondes, editor, *Fuzzy Theory Systems: Techniques and Applications*, pages 1111–1127. Academic Press, San Diego, CA, 1999.

[106] V. Prasad. *Segmentation and recognition of continuous speech.* PhD dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai, 2002.

[107] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall Signal Processing Series, Englewood Cliffs, NJ, 1993.

[108] M. Richard and R. Lippmann. Neural network classifiers estimate Bayesian a posteori probabilities. *Neural Computation*, (3):461–483, 1991.

[109] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulatory markov models for speech recognition. In *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pages 133–139, Paris, France, September 2000.

[110] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modeling from hand-labeled phonetic corpora. In *Proceedings of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Rolduc, The Netherlands, 1998.

[111] S. Roweis and A. Alwan. Towards articulatory speech recognition: Learning smooth maps to recover articulator information. In *Proceedings of the 5th Eurospeech Conference on Speech Communication and Technology (Eurospeech-97)*, pages 1227–1230, Rhodes, Greece, September 1997.

[112] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP research group., editors, *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*. MIT Press, Cambridge, MA, 1986.

[113] L. Saul, M. Rahim, and J. Allen. A statistical model for robust integration of narrowband cues in speech. *Computer Speech and Language*, (15):175–194, 2001.

[114] H. Schramm and P. Beyerlein. Towards discriminative lexicon optimization. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 1457–1460, Aalborg, Denmark, September 2001.

[115] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[116] L. Shapley. A value for n-person games. In H. Kuhn and A. Tucker, editors, *Contributions to the Theory of Games*, number 28 in Annals of Mathematics Studies, pages 307–317. Princeton University Press, 1953.

[117] L. Shastri, S. Chang, and S. Greenberg. Syllable detection and segmentation using temporal flow neural networks. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, volume 3, pages 1721–1724, San Francisco, CA, August 1999.

[118] M. Shire. Syllable onset detection from acoustics. Master's thesis, Department of EECS, University of California, Berkeley, 1997.

[119] E. Shriberg. Phonetic consequences of speech disfluency. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, pages 619–622, San Francisco, August 1999.

[120] M. Siegler and R. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-95)*, volume 1, pages 612–615, Detroit, MI, May 1995.

[121] R. Silipo and S. Greenberg. Automatic detection of prosodic stress in american english discourse. Technical Report TR-00-001, International Computer Science Institute, Berkeley, CA, March 2000.

[122] R. Sillipo and S. Greenberg. Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of the NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[123] R. Singh, B. Raj, and R. Stern. Automatic generation of sub-word units for speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, 10(2):89–99, 2002.

[124] V. Steinbiss, H. Ney, R. Haeb-Umbach, B. Tran, U. Essen, R. Kneser, M. Oerder, H. Meier, Aubert, C. Dugast, and D. Geller. The Philips research system for large-vocabulary continuous-speech recognition. In *Proceedings of the 3rd Eurospeech Conference on Speech Communication and Technology (Eurospeech-93)*, pages 2125–2128, Berlin, Germany, September 1993.

[125] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris. Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-00)*, Beijing, China, October 2000.

[126] K. Stevens. *Acoustic Phonetics*. MIT Press, 1998.

[127] H. Strik, A. Russell, H. van den Heuvel, C. Cucchiarini, and L. Boves. A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology*, (2):119–129, 1997.

[128] M. Sugeno. *Theory of fuzzy integrals and its applications*. PhD dissertation, Tokyo Institute of Technology, 1974.

[129] J. Sun and L. Deng. An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal of Acoustic Society of America*, 111(2):1086–1101, February 2002.

[130] D. van Kuijk and L. Boves. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, (27):95–111, 1999.

[131] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.

[132] J. Verhasselt and J.-P. Martens. A fast and reliable rate of speech detector. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-96)*, pages 2258–2261, Philadelphia, PA, October 1996.

[133] W. Vieregge and T. Broeders. Intra- and interspeaker variation of /r/ in Dutch. In *Proceedings of the 3rd Eurospeech Conference on Speech Communication and Technology (Eurospeech-93)*, pages 267–270, Berlin, Germany, September 1993.

[134] Z. Wang and G. Klir. *Fuzzy Measure Theory*. Plenum Press, New York, 1992.

[135] R. L. Watrous. Phoneme discrimination using connectionist networks. *Journal of Acoustic Society of America*, (87):1753–1772, 1990.

[136] R. L. Watrous. *GRADSIM: A connectionist network simulator using gradient optimization techniques*. Siemens Corporate Research, Inc., Princeton, NJ, 1993.

[137] R. L. Watrous and L. Shastri. Learning phonetic discrimination using connectionist networks: An experiment in speech recognition. Technical Report MS-CIS-86-78, University of Pennsylvania, 1986.

[138] R. L. Watrous and L. Shastri. Learning phonetic features using connectionist networks: An experiment in speech recognition. In *Proceedings of the IEEE First Annual International Conference on Neural Networks*, pages IV:381–388, San Diego, June 1987.

[139] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass. Effect of speaking style on LVCSR performance. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-96)*, pages Addendum:16–19, Philadelphia, PA, October 1996.

[140] J. Westbury. *X-Ray Microbeam Speech Production Database User's Handbook*. Waisman Center on Mental Retardation and Human Development, University of Wisconsin, Madison, WI., 1994.

[141] M. Wester. *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD dissertation, Department of Language and Speech, University of Nijmegen, Nijmegen, Netherlands, 2002.

[142] M. Wester, S. Greenberg, and S. Chang. A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-01)*, pages 1729–1732, Aalborg, Denmark, September 2001.

[143] M. Witbrock and A. Hauptmann. Speech recognition and information retrieval. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, Chantilly, VA, February 1997.

[144] S.L. Wu. *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*. PhD dissertation, Department of EECS, University of California, Berkeley, 1998.

[145] S.L. Wu, M.L. Shire, S. Greenberg, and N. Morgan. Integrating syllable boundary information into speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, Munich, Germany, April 1997.

[146] S. Young. Statistical modelling in continuous speech recognition (CSR). In *Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, August 2001.

[147] J. Zacks and T. Thomas. A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech and Language*, (8):189–209, 1994.

[148] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, (1):3–28, 1978.

[149] J. Zheng, H. Franco, and A. Stolcke. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. *Speech Communication, to appear*, 2002.

[150] D. Zipser. Subgrouping reduces complexity and speeds up learning in recurrent networks. In D. Touretzky, editor, *Advances in Neural Information Processing systems II*, pages 638–641. Morgan Kaufmann, San Mateo, CA, 1990.

[151] G. Zweig. *Speech Recognition with Dynamic Bayesian Networks.* PhD dissertation, Department of EECS, University of California, Berkeley, 1998.